Estimation of GMRFs by Recursive Cavity Modeling

by

Jason K. Johnson

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2003

Author . . . .
Department of Electrical Engineering and Computer Science
March 4, 2003

Certified by . . . .
Alan S. Willsky
Professor of Electrical Engineering
Thesis Supervisor

Accepted by . . . .
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Estimation of GMRFs by Recursive Cavity Modeling

by

Jason K. Johnson

Submitted to the Department of Electrical Engineering and Computer Science
on March 4, 2003, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

This thesis develops the novel method of recursive cavity modeling as a tractable approach to approximate inference in large Gauss-Markov random fields. The main idea is to recursively dissect the field, constructing a cavity model for each subfield at each level of dissection. The cavity model provides a compact yet (nearly) faithful model for the surface of one subfield sufficient for inferring other parts of the field. This basic idea is developed into a two-pass inference/modeling procedure which recursively builds cavity models by an “upward” pass and then builds complementary blanket models by a “downward” pass. Marginal models are then constructed at the finest level of dissection. Information-theoretic principles are employed for model thinning so as to develop compact yet faithful cavity and blanket models thereby providing tractable yet near-optimal inference. In this regard, recursive cavity modeling blends recursive inference and iterative modeling methodologies. While the main focus is on Gaussian processes, general principles are emphasized throughout suggesting the applicability of the basic framework for more general families of Markov random fields. The main objective of the method is to provide efficient, scalable, near-optimal inference for many problems of practical interest. Experiments performed thus far, with simulated Gauss-Markov random fields defined on two-dimensional grids, indicate good reliability and scalability of the method.

The recursive cavity modeling method intersects with a variety of inference techniques arising in the graphical modeling literature including multiscale modeling, junction trees, projection filtering, Markov-blanket filtering, expectation propagation and other methods relying on reduction of embedded models. These connections are explored and important distinctions and extensions are noted. The author believes this thesis represents a significant generalization of existing methods, extending the class of Markov random fields for which reliable, scalable inference is available. But much work remains to better characterize and investigate this claim. Recommendations for furthering this agenda are outlined.

Thesis Supervisor: Alan S. Willsky
Title: Professor of Electrical Engineering

Acknowledgments

To begin with, I would like to thank all of my friends and colleagues at Alphatech for providing the inspiration which led to my decision to attend graduate school in electrical engineering. Among these I especially thank Bob Washburn, Mark Luettgen, Bill Irving, Alan Chao, Mahendra Mallick, Gerry Morse and Ron Chaney. It was a pleasure to work with you all and I have learned something from you each. In particular, I thank Bob for this very positive, dynamic work experience and also for his notes *Markov Trees* and *Gauss-Markov Trees*, the foundation upon which my understanding of Markov random fields and probabilistic inference has developed.

I owe a debt of gratitude to my thesis advisor, Alan Willsky. I thank him for the opportunity to have performed this research and for the tremendous care and effort with which he has read multiple drafts of this thesis. I first became acquainted with Alan's reputation for excellence while working with several of his earlier students (all exceptional researchers) and am privileged to benefit from the same care and skill with which Alan mentors all of his students.

An important aspect of this experience has been interacting with Alan's talented group of students in an intellectually stimulating environment fueled, in large part, by Alan's own foresight, energy and enthusiasm. I have especially benefited from weekly discussions with past and present "graphical model groupies" Mike Schneider, Eric Sudderth, Dewey Tucker and Martin Wainwright who have all influenced my understanding of various topics pertaining to this thesis. I thank Dmitry Malioutov and Ayres Fan for their collaboration on a term project which laid the groundwork for some model thinning ideas developed further in this thesis. I also thank the infamous Taylore Kelly for her friendship and for her thesis pep-talks.

Finally, I thank my parents, Van and Brenda, for always encouraging me to choose my own path in life and for their continuing support and advice down through the years.

Contents

Chapter 1

Introduction

A *Markov random field* (MRF) is a collection of random variables characterized by a set of interactions (dependencies) among subsets of these random variables [44, 45, 66, 18, 107, 37, 97, 67, 25, 137]. Typically, these random variables (state variables) are naturally indexed by spatial locations (sites) where the interactions among random variables involve sets of nearby sites. While originally developed as phenomenological models to aid in the understanding of physical processes, MRFs now play an increasingly important role in the numeric solution, by digital computer, of large-scale problems of statistical inference. In this context, MRFs are often described as *graphical models* indicating the importance of graph theory in the description, representation and inference of these statistical models [105, 88, 77]. In any case, MRFs have been exploited for modeling, simulation and estimation in a wide variety of scientific and engineering disciplines including: physics [20, 62, 101]; geophysics and remote sensing [52]; communication and coding [59]; image processing [134, 60, 19, 87]; medical imaging [90]; speech and natural language processing [108, 106]; as well as artificial intelligence and machine learning [76, 100]. *Gauss-Markov random fields* (GMRFs) are jointly Gaussian[1] MRFs and are prevalent for modeling random fields with continuous-valued states [118, 43, 124, 117].

This thesis introduces *recursive cavity modeling* (RCM) as a novel framework for approximate inference of large, intractable MRFs and develops this framework for GMRFs in particular. This is a recursive inference approach which adopts information theoretic modeling principles – such as information projection, iterative scaling and model selection by generalization of the Akaike information criterion – to develop tractable models of the interaction between subfields of an otherwise intractable MRF.

This introductory chapter poses the fundamental inference problem in the context of GMRFs (introducing some useful notation and terminology), motivates the need for tractable methods of approximate inference for large GMRFs, provides a high-level preview of the RCM method to be developed, and summarizes the content of later chapters.

[1] Multivariate Gaussian distributions are reviewed in Chapter 2.

1.1 Problem Statement

An appealing representation of a GMRF as a graphical model is provided by the *information parameterization* of the Gaussian density. This representation traditionally arises in the information filter implementation of the Kalman filter and related smoothing algorithms (Kalman [78], Kalman and Bucy [79], Rauch, Tung and Striebel [110], Fraser [58]). Consider a Gaussian random vector $\mathrm{x} \sim \mathcal{N}(\hat{x}, P)$, with moment parameters $(\hat{x}, P)$ given by the mean vector $\hat{x} = E\{\mathrm{x}\}$ and the covariance matrix $P = E\{(\mathrm{x}-\hat{x})(\mathrm{x}-\hat{x})'\}$.[2] The information parameterization is defined as $(h, J) = (P^{-1}\hat{x}, P^{-1})$. For non-singular systems, where P and J are invertible, moment parameters and information parameters are in one-to-one correspondence. Moment parameters are recovered from information parameters by $(\hat{x}, P) = (J^{-1}h, J^{-1})$. Adopting notation as in Sudderth [125], we write $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$ to indicate that the random vector x is distributed according to the Gaussian distribution with information parameters (h, J). The vector h will be referred to as the *influence vector* and the matrix J as the *interaction matrix.*

Consider a GMRF having real-valued vector states $\{x_\gamma \in R^{n_\gamma} | \gamma \in \Gamma\}$, where x_γ is the state of site γ, n_γ is the state-dimension, and Γ is the set of all sites. We will let x_γ denote a state variable (a random variable) and let x_γ denote a specific state value (a realization of that random variable). For GMRFs, the joint state $\mathrm{x} = (\mathrm{x}_\gamma, \forall\gamma \in \Gamma)$ is a Gaussian random vector of dimension $n = \sum_\gamma n_\gamma$, which has moment and information parameters as above. We respectively denote by $\hat{x}_\gamma$ and h_γ the corresponding n_γ-dimensional subvectors of $\hat{x}$ and h. For a pair of sites $\gamma, \lambda \in \Gamma$, we denote by $P_{\gamma,\lambda}$ and $J_{\gamma,\lambda}$ the corresponding $n_\gamma \times n_\lambda$ submatrices of P and J. We employ the abbreviations $J_\gamma = J_{\gamma,\gamma}$ and $P_\gamma = P_{\gamma,\gamma}$. The state variable $\mathrm{x}_\gamma \sim \mathcal{N}(\hat{x}_\gamma, P_\gamma)$ is also a Gaussian random vector with marginal moment parameters $(\hat{x}_\gamma, P_\gamma)$ and with marginal information parameters defined as $(\hat{h}_\gamma, \hat{J}_\gamma) = (P_\gamma^{-1}\hat{x}_\gamma, P_\gamma^{-1})$. An important inference problem is the calculation of marginal distributions $p(x_\gamma) = \int p(x) dx_{\Gamma\setminus\gamma}$, the integral of $p(x)$ over all other state variables $x_{\Gamma\setminus\gamma}$ besides x_γ, given the information model $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$. This may be posed as either computation of marginal moment parameters $\mathrm{x}_\gamma \sim \mathcal{N}(\hat{x}_\gamma, P_\gamma)$ or, equivalently, as computation of marginal information parameters $\mathrm{x}_\gamma \sim \mathcal{N}^{-1}(\hat{h}_\gamma, \hat{J}_\gamma)$.

One compelling reason for considering the information parameterization (h, J) is that this often provides a compact graphical model of a GMRF so that the interaction matrix J is sparse. This occurs because the fill-pattern of the interaction matrix reflects the Markov structure of the field (reviewed in Section 2.1). Specifically, the interactions $J_{\gamma,\lambda}$ between sites γ and λ are zero when (only when) the states x_γ and x_λ are conditionally independent given the joint state of all other sites. The Markov structure of the field is then summarized by an undirected graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$, with vertices Γ and edges $\mathcal{E}_\Gamma$, where vertices represent sites of the field and edges represent interactions between sites. Only those pairs of sites such that $J_{\gamma,\lambda} \neq 0$ are linked by an edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$. We will further review graphical models and the Markov

[2] We let $E\{\cdot\}$ denote the expectation operator such that, for a continuous-valued random variable x, $E\{f(\mathrm{x})\} = \int p(x)f(x)dx$ where $p(x)$ is the probability density function (pdf) of x.

property in Section 2.1.

Another reason for considering the information parameterization is that it is well-suited for incorporating local observations of the form $y_\gamma = C_\gamma x_\gamma + v_\gamma$, a linear observation of the state x_γ corrupted by additive Gaussian measurement noise $v_\gamma \sim \mathcal{N}(0, R_\gamma)$. Provided the measurement noise is independent from site to site, these observations are readily absorbed into the information parameters by updating local influences $h_\gamma \leftarrow h_\gamma + C'_\gamma R_\gamma^{-1} y_\gamma$ and interactions $J_{\gamma,\gamma} \leftarrow J_{\gamma,\gamma} + C'_\gamma R_\gamma^{-1} C_\gamma$. Note that such local observations do not change the interaction structure of the field but only update local influence and interaction parameters. Given this updated information model, the inference problem then is to infer the *conditional* marginal distribution $\mathrm{x}_\gamma | y \sim \mathcal{N}(\hat{x}_\gamma(y), \hat{P})$, where $y = (y_\gamma, \forall \gamma \in \Gamma)$ are the observations, $\hat{x}_\gamma(y) = E\{\mathrm{x}_\gamma | y\}$ is the conditional mean and $\hat{P}_\gamma = E\{(\mathrm{x}_\gamma - \hat{x}_\gamma(y))(\mathrm{x}_\gamma - \hat{x}_\gamma(y))' | y\}$ is the conditional covariance.[3] This is essentially the same calculation as that discussed previously, only the parameter values have changed. Hence, without any loss of generality, we may omit explicit reference to the observations and focus on the fundamental problem of inferring the marginal moments $\{(\hat{x}_\gamma, P_\gamma) | \forall \gamma \in \Gamma\}$ given the information model $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$. Several prototypical inference problems are graphically depicted in Figure 1-1.

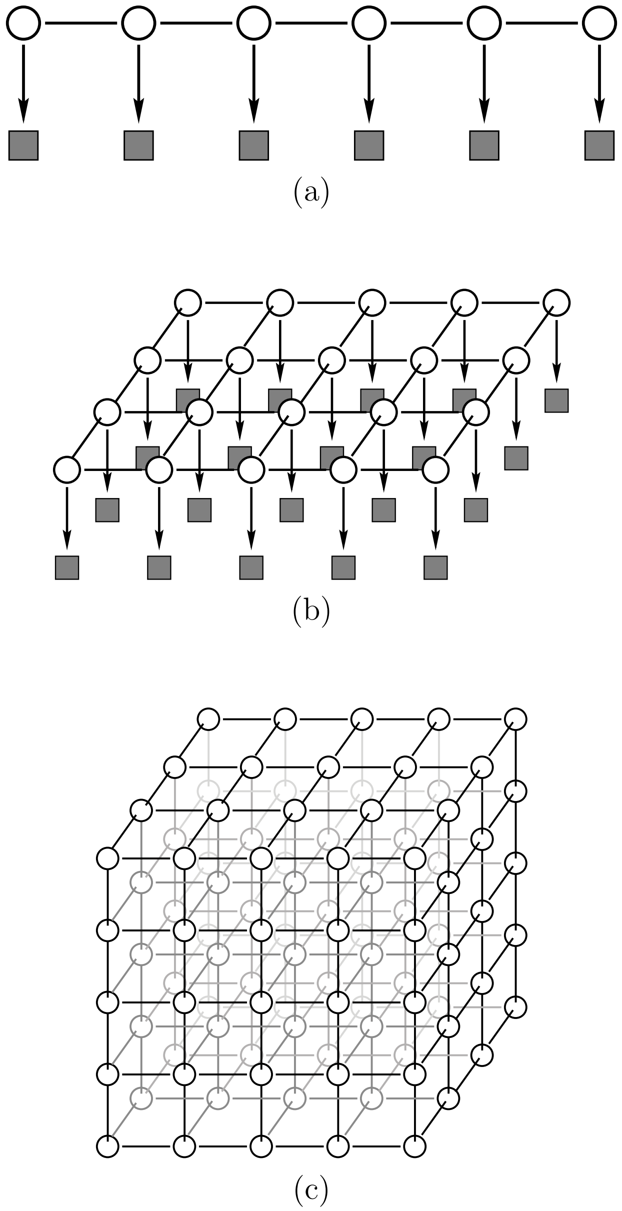
1.2 Motivation

In principle, computation of the marginals is straightforward. The full set of Gaussian moment parameters is recovered from the information parameters by $(\hat{x}, P) = (J^{-1}h, J^{-1})$. The marginal densities are then given by selecting the appropriate subsets of marginal moments $(\hat{x}_\gamma, P_\gamma)$ for each site γ. This "brute-force" inference for an n-dimensional GMRF requires $\mathcal{O}(n^3)$ computation with $\mathcal{O}(n^2)$ memory storage.

A more economical approach exploits any sparse Markov structure of the field by employing recursive inference methodologies such as in the multiscale modeling literature[4] or by the junction tree approach (Dawid [39, 40], Lauritzen and Spiegelhalter [89], Shenoy and Shafer [122]) reviewed in Section 2.3.2. Yet, even these approaches suffer when the so-called tree-width w, the largest state-space dimension within a tree-structured representation of the random field (see discussion and illustrations in Section 2.3), is large, as these methods require $\mathcal{O}(w^3)$ computations. For instance, in image processing applications tree-widths of order $n^{1/2}$ are common, requiring $\mathcal{O}(n^{3/2})$ computation which does not scale linearly with the dimension of the field. For many image processing applications such methods will prove impractical. For very large problems, such as in oceanographic or meteorological modeling with $n \gg 10^6$ (Fieguth et al [52]), such exact methods are not feasible. The scalability of recursive inference in higher dimensions is even less favorable. For instance, for GMRFs defined on 3-D lattices, we may expect tree widths of order $n^{2/3}$ so that analogous recursive

[3] We let $E\{\cdot|y\}$ denote the conditional expectation operator with respect to the conditional probability density $p(x|y) = p(x,y)/p(y)$, i.e. $E\{f(\mathrm{x})|y\} = \int f(x)p(x|y)dx$.

[4] A recent article by Willsky [132] gives a perspective on this extensive body of work. Selected aspects of the multiscale modeling approach are reviewed in Section 2.3.3.



(a)

(b)

(c)

Figure 1-1: Graphical depiction of several prototypical inference problems involving MRFs. Nodes (circles) indicate sites/states of the field, edges (lines) indicate interactions between states (e.g., for GMRFs, where $J_{\gamma,\lambda} \neq 0$), and boxes indicate noisy observations of states. (a) depicts the hidden Markov model (HMM) scenario arising in signal and speech processing, (b) depicts a 2-D MRF such as arises in image processing applications, and (c) depicts a 3-D MRF (observations suppressed) such as might occur in the modeling of, for instance: solids in physics; volumes of the ocean or atmosphere in the earth sciences; or tissues in medical imaging (e.g; tomography, MRI, ultrasound).

methods then require $\mathcal{O}(n^2)$ computation. Again, the recursive approach is more favorable than the $\mathcal{O}(n^3)$ brute-force calculation (not exploiting Markov structure), but even the recursive approach quickly becomes intractable for larger fields and this effect is more pronounced in higher spatial dimensions. Hence, the motivation for developing scalable, near-optimal inference methods for GMRFs. This is the primary goal of the RCM approach. Furthermore, the basic RCM framework, developed here for GMRFs, should prove applicable and appropriate for other classes of MRFs where exact methods may be even more complex than for GMRFs. For instance, for binary-state MRFs the complexity of exact recursive inference is $\mathcal{O}(2^w)$.

1.3 Preview of RCM

In recent years, several methods for tractable inference of MRFs have arisen in the graphical modeling literature which marry ideas of recursive inference with modeling notions such as approximation by *information projection* (minimizing Kullback-Leibler divergence over a family of less complex graphical models, as we discuss further in Chapter 2). These include *projection filtering* methods for nonlinear filtering in Markov chains (Kulhavý [86]; Brigo [27]); *Markov-blanket filtering* methods for dynamic Bayesian networks (Boyen and Koller [22]; Murphy and Weiss [99]); the *expectation propagation* framework (Minka [96]; Heskes and Zoeter [70]); and the *junction trees edge-removal* method of Kjærulff [82, 83]. A fundamental theme common to all of these methods is the idea of performing information projection of embedded models to less complex families of models thereby easing the burden of subsequent inference by (otherwise) exact methods.

The recursive cavity modeling (RCM) approach continues and extends this trend, choosing a recursive divide-and-conquer approach to inference, inspired by the multiscale modeling and junction tree methods, while adopting information-theoretic modeling principles to select compact yet faithful approximations for the "Markov blanket models" arising in this context. RCM exploits the Markov structure of the field to reduce computation by forming a tree-structured decomposition of the field. This is given by recursively dissecting the field into disjoint subfields (see Figure 1-2). The inference is then structured according to this hierarchical dissection of the field employing a two-pass inference procedure involving an "upward" pass followed by a "downward" pass. In this regard, RCM closely resembles inference procedures developed for tree-structured multiscale models (Luettgen et al [92]). However, to contain the complexity of the inference computations, the RCM approach introduces model thinning operations so as to develop compact models of the surfaces between subfields which are then used to infer other subfields. As in earlier approaches of Taylor [126] and Daniel [36], this is posed as "thinning" of a graphical model by pruning edges. RCM, however, formalizes an information theoretic modeling perspective for model thinning. That is we adopt the information representation of the GMRF (as an exponential family) and then perform edge-pruning by information projection to selected (nearby) lower-order families of graphical models. In the context of the upward pass, this leads to the construction of "cavity models" providing thinned models for the

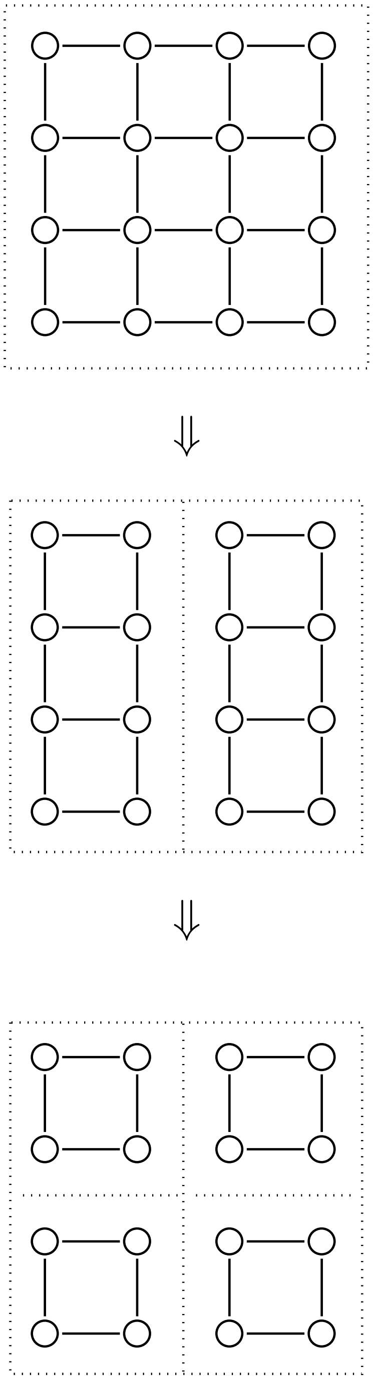


Figure 1-2: Illustration of nested dissection of 4×4 square-grid graphical model. The first cut partitions the sites Γ into two disjoint subsets $\Gamma_1 \cup \Gamma_2 = \Gamma$. This may be viewed as "cutting" the edges J_{Γ_1,Γ_2} producing two partial models $(h_{\Gamma_1}, J_{\Gamma_1})$ and $(h_{\Gamma_2}, J_{\Gamma_2})$. Recursing this cutting procedure generates a tree-structured nested dissection of the graphical model. Subsequent RCM inference procedures are recursively structured according to this dissection procedure operating on the dissected parts of the graphical model.

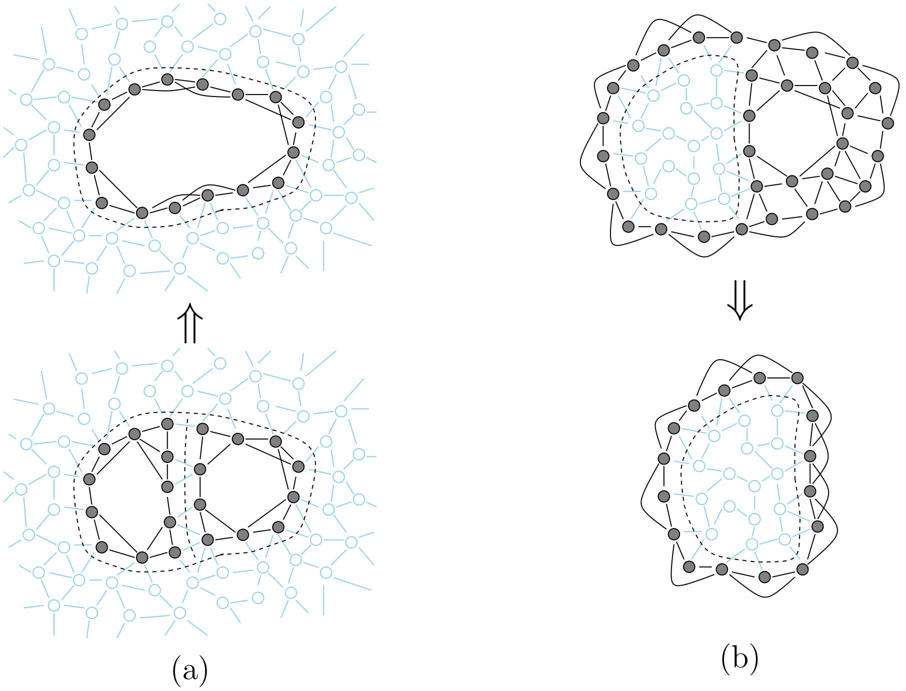


Figure 1-3: Illustration of RCM inference approach: (a) An “upward” cavity modeling procedure constructs cavity models from sub-cavity models. This entails *joining* two cavity models along their common boundary by reinstating those elements of J severed during nested dissection; *marginalizing* over states in the interior of the joined cavity model; and *pruning* weak edges (interactions) by information projection. (b) A “downward” blanket modeling procedure constructs blanket models from adjacent cavity and super-blanket models. This is also implemented by a combination of joining, marginilization, and edge removal.

surfaces of subfields. A corresponding “downward” pass then builds complimentary “blanket models” for each subfield from the cavity models of adjacent subfields. Both inference/modeling procedures are illustrated in Figure 1-3. The blanket models produced at the finest level of dissection then allow for inference of the site marginals. Note that cavity models are constructed recursively from subcavity models and hence the name of our method, recursive cavity modeling.

The model thinning procedure plays a central role in this approach to inference. The technique developed here departs from earlier information projection methods for approximate inference in several regards. Most notable, we do not impose a specific type of model structure, such as “factored” or “tree-structured”, but rather adaptively select model structure according to a model selection criterion inspired by Akaike and Bayesian information criteria (reviewed in Section 2.2.5). This criterion balances the competing objectives of model fidelity (measured by Kullback-Leibler divergence) and model complexity (measured by model order, the number of independent model parameters) where a parameter setting, specifying the acceptable level of information loss (KL-divergence) per removed model parameter, controls how strongly model compactness is favored relative to model fidelity. This provides a principled basis for inductively selecting "weak" interactions to prune from the model. A greedy thinning procedure is developed to decrease this metric by incremental information projections to selected lower-order families of models. We find that this naturally leads to the development of "thin" (low tree-width) cavity and blanket models such that the requisite information projections may be implemented in a tractable manner employing recursive inference and iterative parameter fitting subroutines. A novel adaptation of the iterative scaling method (reviewed in Section 2.2.4) is developed to implement the parameter fitting subroutine.

1.4 Thesis Organization

For the convenience of the reader, a guide indicating the content of subsequent chapters is provided below.

Chapter 2 – Background. This chapter provides a unified review of the relevant literature so as to provide the necessary context for developing and understanding the recursive cavity modeling approach. The main objectives, in addition to acknowledging important influences, are to (i) present the picture of the GMRF as a graphical model, (ii) emphasize connections between information theory and modeling, and (iii) provide an overview of some recursive inference techniques.

Chapter 3 – Model Thinning. Modeling notions play a central role in the RCM inference approach. For this reason, Chapter 3 is devoted to the preliminary task of developing the fundamental modeling ideas and methods which are later employed for model thinning in the context of RCM. The main features of this model thinning procedure are (i) a model selection metric inspired by the Akaike and Bayesian information criteria; (ii) imposing model structure by information projections implemented by a novel adaptation of the iterative scaling procedure; and (iii) a greedy, inductive model thinning procedure which incrementally thins the graphical model by information projections.

Chapter 4 – Recursive Cavity Modeling. This chapter details and discusses the basic RCM inference procedure and also presents two iterative extensions of this basic procedure. The model thinning approach of the previous chapter is employed as a subroutine. Pseudo-code and illustrations are provided. Simulations are performed to investigate the scalability and the reliability of the method with some very favorable results.

Chapter 5 – Conclusion. The main ideas and methods of the RCM approach are summarized. Since the approach shares many ideas in common with a variety of significant methods, an outline of the original contributions made by this thesis is provided. Distinctions between RCM and existing methods are emphasized indicating where, in the authors view, RCM represents a significant improvement and/or generalization over existing methods. Finally, this thesis really only begins to explore a novel framework for tractable inference of MRFs. Hence, in closing, recommendations for further research and development aimed at refining and extending this framework are outlined.

Chapter 2

Background

This chapter provides a broad overview of several topics pertaining to the description, identification and inference of MRFs from the perspective of *graphical modeling.* Section 2.1 introduces graphical models, focusing on the formulation appropriate for MRFs, and discusses the information parameterization of a GMRF from this perspective. In Section 2.2, we step back momentarily, reviewing general principles for statistical model selection and parameter estimation deriving from information theory. Such information-theoretic principles have come to play an increasingly important role in the graphical modeling literature, not only for the selection of model order, structure and parameters; but also for the design and analysis of approximate inference procedures. Section 2.3 then reviews recursive approaches to inference and discusses several approximate inference methods combining recursive inference with information theoretic modeling principles.

2.1 Graphical Models

This section introduces the description of Markov random fields as *graphical models.* Graphical models are often regarded as a marriage of graph theory and probability theory, as the language of graph theory plays a central role in the description of both the statistical structure of these models and of the data structures and algorithms employed to represent and process such models by digital computer. Basic graph theoretic terminology and notation is developed in the first subsection. Subsequent subsections then clarify the role graph theory plays both for describing the statistical structure of Markov random fields and in providing parameterized representations of the probability distribution of a MRF. Specificly, the information parameterization of GMRFs is discussed in the last subsection.

2.1.1 Graphs and Hypergraphs

This subsection provides some basic definitions and notation adopted from graph theory. The definitions for undirected graphs prove useful for specifying the so-called Markov structure of the random field (Section 2.1.2). The more general definition of a hypergraph proves useful for describing the structure of the underlying interactions – local potential functions employed to construct exponential families of Gibbs random fields (Section 2.1.3). We first define hypergraphs and then define graphs as a subclass of hypergraphs.

Hypergraphs. Let Γ be a finite set $\{\gamma_1, \gamma_2, \dots, \gamma_{|\Gamma|}\}$. A *hypergraph* (Berge [15], Lauritzen [88], Yeung et al [137]) based on Γ is a pair $\mathbf{H}_\Gamma = (\Gamma, \mathcal{H}_\Gamma)$ where $\mathcal{H}_\Gamma = (H_i, i \in I)$ is a collection of nonempty subsets $H_i \subset \Gamma$. The elements $\gamma \in \Gamma$ are called the *vertices* and the subsets $H \in \mathcal{H}_\Gamma$ are called *hyperedges.* We will consider only *simple* hypergraphs where $H_i \neq H_j$ for all $i \neq j$. Hence, the collection of hyperedges $\mathcal{H}_\Gamma$ forms a set. Several examples of hypergraphs are illustrated in Figure 2-1.

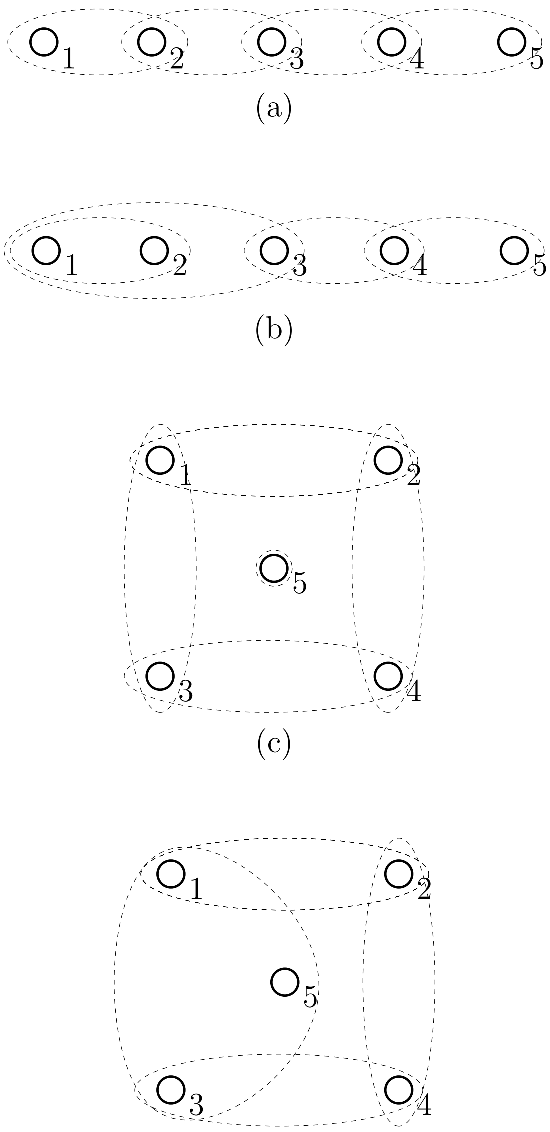
The following definitions are with respect to a given hypergraph $\mathbf{H}_\Gamma$ and pertain only to the vertices and hyperedges of $\mathbf{H}_\Gamma$. A vertex γ and hyperedge H are said to be *incident* in $\mathbf{H}_\Gamma$ when the hyperedge contains the vertex $\gamma \in H$. Two vertices (hyperedges) are *adjacent* in $\mathbf{H}_\Gamma$ when they are incident to a common hyperedge (vertex). A *chain* in $\mathbf{H}_\Gamma$ of length l is an alternating sequence of vertices and hyperedges $(\gamma_0, H_0, \gamma_1, H_1, \dots, H_{l-1}, \gamma_l)$ where the vertices $(\gamma_k, k = 1, \dots, l)$ and the hyperedges $(H_k, k = 0, \dots, l-1)$ are distinct and where $\gamma_k, \gamma_{k+1} \in H_k$ for $k = 0, \dots, l-1$. A *cycle* is a chain with length $l > 1$ and with $\gamma_0 = \gamma_l$. If there is a chain beginning at vertex $\gamma_0 = \gamma$ and ending at vertex $\gamma_l = \lambda$ then these vertices are said to be *connected.* This defines an equivalence relation $\gamma \equiv \lambda$, the equivalence classes of which are the *connected components* of $\mathbf{H}_\Gamma$. A hypergraph which has only one connected component is said to be *connected.* A hypergraph which does not contain any cycles is *acyclic.* Finally, we define the *subhypergraph* of $\mathbf{H}_\Gamma$ induced by $\Lambda \subset \Gamma$ as $\mathbf{H}_\Lambda = (\Lambda, \mathcal{H}_\Lambda)$ where $\mathcal{H}_\Lambda = \{H \in \mathcal{H}_\Gamma | H \subset \Lambda\}$.

Graphs. For our purposes it is convenient to consider only *undirected* graphs.[1] A *graph* is a hypergraph where all hyperedges are doublet sets as in Figure 2-1(a). Such doublet hyperedges are called *edges.* We adopt the usual convention of depicting graphs as in Figure 2-2(c) drawing lines between adjacent vertices to indicate edges. Hence, a graph based on Γ is a pair $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ where $\mathcal{E}_\Gamma$ is a collection of unordered pairs of distinct vertices. Again, we consider only simple graphs such that the collection of edges forms a set. The above definitions for hypergraphs then apply to graphs without modification. A chain in a graph is also called a *path.* Paths may be specified by a sequence of adjacent vertices $(\gamma_0, \dots, \gamma_l)$. We also state some additional definitions which we apply only to graphs.

A subset $\Lambda \subset \Gamma$ is *complete* in $\mathbf{G}_\Gamma$ if every pair of vertices $\gamma, \lambda \in \Lambda$ are adjacent in $\mathbf{G}_\Gamma$. These are referred to as the *cliques* of the graph. A clique is *maximal* if it is not a subset of some other clique. We denote by $\mathcal{C}(\mathbf{G}_\Gamma)$ the class of all cliques in $\mathcal{E}_\Gamma$ and by $\mathcal{C}^*(\mathbf{G}_\Gamma)$ the class of all maximal cliques.

We will say that a graph $\mathbf{G}_\Gamma$ is the *adjacency graph* of a hypergraph $\mathbf{H}_\Gamma$, and denote this by $\mathbf{G}_\Gamma = \text{adj } \mathbf{H}_\Gamma$, when the edges of the graph $\mathcal{E}_\Gamma$ consist of all two-

[1] See Berge [15] for the definition of general *directed* graphs (specified by ordered pairs of vertices called *arcs*).



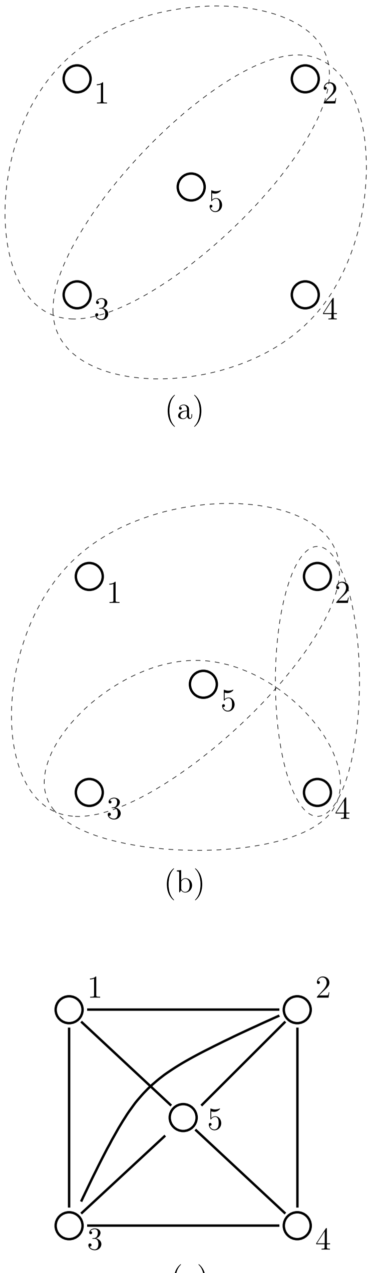
(a)

(b)

(c)

(d)

Figure 2-1: Diagrams of several hypergraphs based on $\Gamma = \{1, 2, 3, 4, 5\}$ with hyperedges $\mathcal{H}_\Gamma$ respectively given by: (a) $\{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 5\}\}$, (b) $\{\{1, 2\}, \{3, 4\}, \{4, 5\}, \{1, 2, 3\}\}$, (c) $\{\{5\}, \{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$, (d) $\{\{1, 2\}, \{2, 4\}, \{3, 4\}, \{1, 3, 5\}\}$. Vertices are indicated by circular nodes and hyperedges by dashed lines enclosing just the members of that hyperedge. It is sometimes useful to require that isolated vertices, such as node 5 in (c), are enclosed by a singleton hyperedge. Only hypergraph (a) qualifies as a graph. Hypergraphs (a),(b) and (d) are connected, (c) has two connected components. Hypergraphs (a) and (b) are acyclic.



(a)

(b)

(c)

Figure 2-2: Illustration of graphical equivalence. All three diagrams (a)-(c) depict graphically equivalent hypergraphs where (c) is the adjacency graph of (a) and (b) (and of itself). Hypergraph (a) is the maximal clique hypergraph of graph (c).

element subsets $\{\gamma, \lambda\} \subset \Gamma$ such that vertices γ and λ are adjacent in $\mathbf{H}_\Gamma$.

$$\mathcal{E}_\Gamma = \{\{\gamma, \lambda\} \subset \Gamma \mid \exists H \in \mathcal{H}_\Gamma : \{\gamma, \lambda\} \subset H\} \tag{2.1}$$

Note that we always have $\mathbf{G}_\Gamma = \text{adj } \mathbf{G}_\Gamma$. Further, we will say that two hypergraphs $\mathbf{H}_\Gamma$ and $\mathbf{H}'_\Gamma$ are *graphically equivalent* if adj $\mathbf{H}_\Gamma =$ adj $\mathbf{H}'_\Gamma$. Note that this defines an equivalence relation over the set of hypergraphs based on Γ. Moreover, the equivalence classes of graphically equivalent hypergraphs are in one-to-one correspondence with the set of all graphs based on Γ. Each graph $\mathbf{G}_\Gamma$ represents a distinct class of graphically equivalent hypergraphs determined by $\mathbf{G}_\Gamma =$ adj $\mathbf{H}_\Gamma$. Also, we define the *clique hypergraph* of graph $\mathbf{G}_\Gamma$ as cliq $\mathbf{G}_\Gamma = (\Gamma, \mathcal{C}(\mathbf{G}_\Gamma))$ such that the hyperedges of the hypergraph are the cliques of the graph. The *maximal clique hypergraph* is defined similarly as cliq* $\mathbf{G}_\Gamma = (\Gamma, \mathcal{C}^*(\mathbf{G}_\Gamma)$. Graphical equivalence and several of the above definitions are illustrated in Figure 2-2.

Finally, the following definitions are used to discuss the adjacency structure of a graph $\mathbf{G}_\Gamma$. The *boundary* $\partial\Lambda$ of subset $\Lambda \subset \Gamma$ is the set of vertices not in Λ which are adjacent to some vertex in Λ. Also, the *closure* of Λ is defined $\bar{\Lambda} = \Lambda \cup \partial\Lambda$. In this notation we allow γ to also denote the singleton set $\{\gamma\}$ (e.g., $\partial\gamma = \partial\{\gamma\}$ and $\bar{\gamma} = \{\gamma\} \cup \partial\gamma$). These definitions may be extended to general hypergraphs, but it is clear that the adjacency structure of a hypergraph $\mathbf{H}_\Gamma$ is completely captured by the adjacency graph $\mathbf{G}_\Gamma =$ adj $\mathbf{H}_\Gamma$. Yet, while $\mathbf{G}_\Gamma$ is determined by these adjacency relations, the hypergraph $\mathbf{H}_\Gamma$ is not.

2.1.2 Markov Random Fields

A *random variable* x is defined by a set $\mathcal{X}$ and a probability distribution $p(x)$ on $\mathcal{X}$.[2] Given a function $f : \mathcal{X} \to \mathcal{R}$, we define the expectation of f (with respect to p) by

$$E_p\{f(\mathrm{x})\} = \int_\mathcal{X} p(x) f(x) dx \tag{2.2}$$

where the integral is to be understood as a sum in the case that $\mathcal{X}$ is discrete (this convention is adopted henceforth). Let (x, y) be a pair of random variables defined by a probability distribution $p(x, y)$ on a product set $\mathcal{X} \times \mathcal{Y}$. The *marginal distribution* of x is $p(x) = \int_\mathcal{Y} p(x, y) dy$ (similarly for y). We say that x and y are *independent* if $p(x, y) = p(x)p(y)$ for all x, y. The *conditional distribution* of x given y is defined as $p(x|y) = p(x, y)/p(y)$ for each y such that $p(y) > 0$. It holds that x and y are independent if and only if $p(x|y) = p(x)$ for all y where $p(y) > 0$ and for all x. Intuitively, observing y does not affect our knowledge of x. For a triplet of random variables $(\mathrm{x}, \mathrm{y}, \mathrm{z})$ we say that x and y are *conditionally independent* given z if $p(x, y|z) = p(x|z)p(y|z)$ for all z where $p(z) > 0$ and for all x, y. Equivalently,

[2] We will say that p is a *probability distribution* on $\mathcal{X}$ if either: (i) $\mathcal{X}$ is discrete (has countably many elements) and p is a probability mass function (pmf), a non-negative function defined on $\mathcal{X}$ normalized such that $\sum_{x \in \mathcal{X}} p(x) = 1$; or (ii) $\mathcal{X}$ consists of a continuum of elements (e.g. $\mathcal{X} = \mathcal{R}$) and p is a probability density function (pdf), a non-negative function defined on $\mathcal{X}$ normalized such that $\int_\mathcal{X} p(x) dx = 1$.

$p(x|z,y) = p(x|z)$ for all x, y, z where $p(z) > 0$ and $p(y|z) > 0$. Intuitively, having already observed z, then observing y does not affect our knowledge of x.

A *random field* is a collection of random variables $\mathrm{x}_\Gamma = (\mathrm{x}_\gamma, \gamma \in \Gamma)$ defined by a probability distribution $p(x_\Gamma)$ on a product set $\mathcal{X}_\Gamma = \prod_{\gamma \in \Gamma} \mathcal{X}_\gamma$. We will refer to an element $\gamma \in \Gamma$ as a *site* of the field and to a random variable x_γ as the *state* of site γ. Each $\Lambda \subset \Gamma$ defines a *subfield* $\mathrm{x}_\Lambda = (\mathrm{x}_\gamma, \gamma \in \Lambda)$ and associated *state-space* $\mathcal{X}_\Lambda = \prod_{\gamma \in \Lambda} \mathcal{X}_\gamma$. We let $\backslash\Lambda$ denote the set complement of Λ in Γ, i.e. $\backslash\Lambda = \{\gamma \in \Gamma | \gamma \notin \Lambda\}$.

A *Markov random field* is a pair $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ satisfying the following property:

Definition 1 (Local Markov) *A random field* x_Γ *is* locally Markov *with respect to graph* $\mathbf{G}_\Gamma$ *if, for every site* $\gamma \in \Gamma$*, the states* x_γ *and* $\mathrm{x}_{\backslash\bar{\gamma}}$ *are conditionally independent given* $\mathrm{x}_{\partial\gamma}$*.*

A.A. Markov [94] introduced this idea into statistics to consider generalization of the law of large numbers for dependent processes. It is sometimes useful to exploit a stronger version of the Markov property. We say that, for a triplet $(\Lambda_1, \Lambda_2, S)$ of subsets of Γ and a graph $\mathbf{G}_\Gamma$, that S *separates* Λ_1 and Λ_2 in graph $\mathbf{G}_\Gamma$ if every path in $\mathbf{G}_\Gamma$ containing vertices of both Λ_1 and Λ_2 must also contain some vertex of S.

Definition 2 (Global Markov) *A random field* x_Γ *is* globally Markov *with respect to graph* $\mathbf{G}_\Gamma$ *if, for every* $(\Lambda_1, \Lambda_2, S)$ *as above, the states* x_{Λ_1} *and* x_{Λ_2} *are conditionally independent given the state of the separator* x_S*.*

Both Markov properties are illustrated in Figure 2-3. These are closely related, as shown by the following proposition.

Proposition 1 (Local/Global Near-Equivalence) *If a random field* x_Γ *is globally Markov with respect to* $\mathbf{G}_\Gamma$ *then it is locally Markov. Conversely, if a (locally) Markov random field* $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ *has non-vanishing probability distribution,* $p(x_\Gamma) > 0$ *for all* $x_\Gamma \in \mathcal{X}_\Gamma$*, then it is globally Markov.*

Proof. That global Markov implies local Markov is immediate, just relate the two definitions identifying $(\Lambda_1, \Lambda_2, S)$ with $(\gamma, \backslash\bar{\gamma}, \partial\gamma)$. The converse proof is not so simple, resembling that of the Hammersley-Clifford theorem below. We instead appeal to that theorem (see remarks following Proposition 2). □

The preceding discussion demonstrates the important role graphs play for describing the Markov structure (conditional independencies) of a random field. Now we approach the issue of how to *represent* the probability distribution of such a Markov random field. The central result in this area is the Hammersley-Clifford theorem relating factorization of the probability distribution to Markov structure. We will utilize the language of hypergraphs to state this relation precisely.

Definition 3 (Hypergraph Factorization) *A probability distribution* $p(x)$ *of a random field* $(x_\gamma, \gamma \in \Gamma)$ *is said to* factor *with respect to a hypergraph* $\mathbf{H}_\Gamma$ *based on* Γ *if*

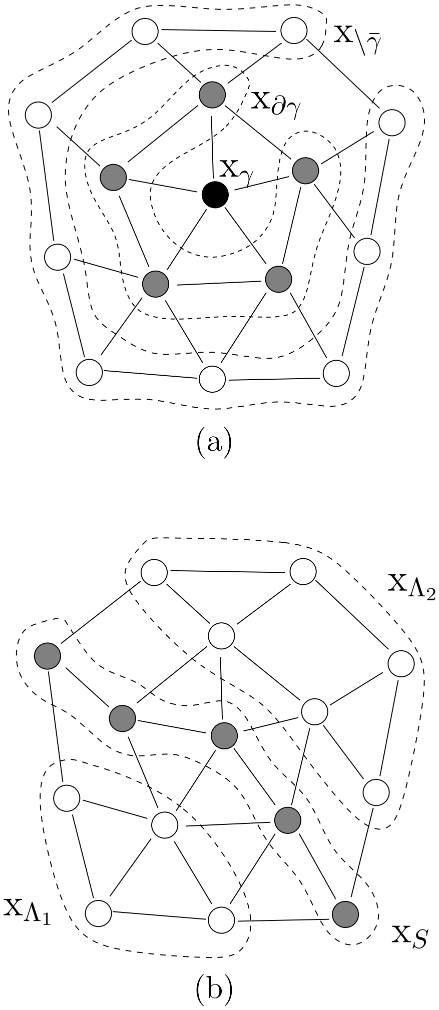


Figure 2-3: Illustration of local and global Markov properties. (a) The local Markov property insures that the state x_γ of site γ (shown in black) is conditionally independent of the state $\mathrm{x}_{\setminus\bar{\gamma}}$ of all sites not adjacent to γ (white) given the state $\mathrm{x}_{\partial\gamma}$ of all sites adjacent to γ (grey). This means that $p(x_\gamma|x_{\setminus\gamma}) = p(x_\gamma|x_{\partial\gamma})$. (b) The stronger global Markov property insures that, conditioned on the state of the separator x_S (grey), the states x_{Λ_1} and x_{Λ_2} of the separated subfields are conditionally independent. This means that $p(x_{\Lambda_1}, x_{\Lambda_2}|x_S) = p(x_{\Lambda_1}|x_S)p(x_{\Lambda_2}|x_S)$.

there exists a collection of non-negative functions ($\psi_\Lambda : \mathcal{X}_\Lambda \to \mathcal{R}^+, \Lambda \in \mathcal{H}_\Gamma$) *and a finite constant* $Z(\psi) < \infty$ *such that*

$$p(x_\Gamma) = \frac{1}{Z(\psi)} \prod_{\Lambda \in \mathcal{H}_\Gamma} \psi_\Lambda(x_\Lambda) \tag{2.3}$$

for all $x_\Gamma \in \mathcal{X}_\Gamma$*, in which case the normalizing constant* $Z(\psi)$ *is given by*

$$Z(\psi) = \int_{\mathcal{X}} \prod_{\Lambda \in \mathcal{H}_\Gamma} \psi_\Lambda(x_\Lambda) dx_\Gamma \tag{2.4}$$

The functions $\psi_\Lambda(x_\Lambda)$ *are sometimes called* compatibility functions *(Wainwright [129]).*

The following proposition states, in terms of hypergraphs, the theorem of Hammersley and Clifford [68]. Several proofs follow Grimmet [66] employing the Möbius inversion lemma (Mitter [97], Guyon [67], Lauritzen [88], Brémaud [25]).

Proposition 2 (Hammersley-Clifford) *A random field* x_Γ *which factors with respect to hypergraph* $\mathbf{H}_\Gamma$ *is globally Markov with respect to graph* $\mathbf{G}_\Gamma = \mathrm{adj}\ \mathbf{H}_\Gamma$*. Conversely, a (locally) Markov random field* $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ *which has positive probability distribution* $p(x_\Gamma) > 0$ *for all* $x_\Gamma \in \mathcal{X}_\Gamma$*, factors with respect to hypergraph* $\mathbf{H}_\Gamma = \mathrm{cliq}^*\ \mathbf{G}_\Gamma$ *(and, hence, is globally Markov).*

Proof. That hypergraph factorization implies the global Markov property is straight forward. Without any loss of generality, we consider just the case that $(\Lambda_1, \Lambda_2, S)$ partitions Γ. Let $\pi_i(x_{\Lambda_i}, x_S)$ $(i = 1, 2)$ denote the product of all factors $\psi_\Lambda(x_\Lambda)$ which depend upon x_{Λ_i}. Let $\pi_S(x_S)$ denote the product of all remaining factors (which depend *only* upon x_S but not x_{Λ_1} or x_{Λ_2}). Then, $p(x_\Gamma) \propto \pi(x_{\Lambda_1}, x_{\Lambda_2}, x_S) \equiv \pi_1(x_{\Lambda_1}, x_S)\pi_2(x_{\Lambda_2}, x_S)\pi_S(x_S)$. The conditional distribution $p(x_{\Lambda_1}|x_S)$ is then given by

$$\begin{aligned} p(x_{\Lambda_1}|x_S) &= \frac{\int \pi(x_{\Lambda_1}, \tilde{x}_{\Lambda_2}, x_S) d\tilde{x}_{\Lambda_2}}{\int\int \pi(\tilde{x}_{\Lambda_1}, \tilde{x}_{\Lambda_2}, x_S) d\tilde{x}_{\Lambda_1} d\tilde{x}_{\Lambda_2}} \\ &= \frac{\pi_1(x_{\Lambda_1}, x_S)}{\int \pi_1(\tilde{x}_{\Lambda_1}, x_S) d\tilde{x}_{\Lambda_1}} \end{aligned} \tag{2.5}$$

and similarly for $p(x_{\Lambda_2}|x_S)$. The conditional distribution $p(x_{\Lambda_1}, x_{\Lambda_2}|x_S)$ then factors as shown below:

$$\begin{aligned} p(x_{\Lambda_1}, x_{\Lambda_2}|x_S) &= \frac{\pi(x_{\Lambda_1}, x_{\Lambda_2}, x_S)}{\int\int \pi(\tilde{x}_{\Lambda_1}, \tilde{x}_{\Lambda_2}, x_S) d\tilde{x}_{\Lambda_1} d\tilde{x}_{\Lambda_2}} \\ &= \left(\frac{\pi_1(x_{\Lambda_1}, x_S)}{\int \pi_1(\tilde{x}_{\Lambda_1}, x_S) d\tilde{x}_{\Lambda_1}}\right)\left(\frac{\pi_2(x_{\Lambda_2}, x_S)}{\int \pi_2(\tilde{x}_{\Lambda_2}, x_S) d\tilde{x}_{\Lambda_2}}\right) \\ &= p(x_{\Lambda_1}|x_S) p(x_{\Lambda_2}|x_S) \end{aligned} \tag{2.6}$$

The converse is shown by construction employing the Möbius inversion lemma. We examine this further when we discuss the Gibbs canonical potential in the next subsection. □

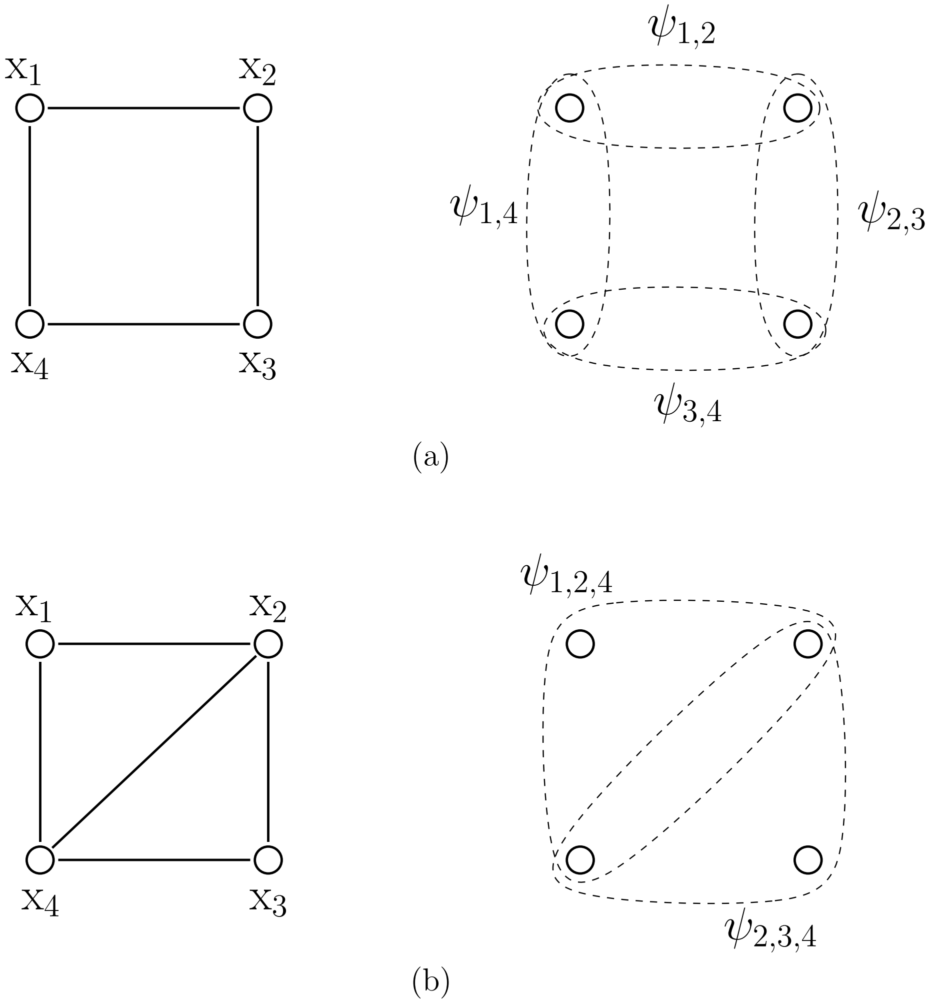


Figure 2-4: Illustration of Hammersley-Clifford theorem for two MRFs each with four sites $\Gamma = \{1, 2, 3, 4\}$. In each example, the MRF $(x_\Gamma, \mathbf{G}_\Gamma)$ is shown to the left and the corresponding hypergraph factorization $(\mathbf{H}_\Gamma, \psi)$ is shown to the right. (a) The Markov structure implies that the probability distribution may be factored as a product of pairwise compatibility functions $p(x) \propto \psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3)\psi_{3,4}(x_3, x_4)\psi_{4,1}(x_4, x_1)$. (b) The Markov structure of the field implies $p(x)$ may be factored as a product of triplet-wise compatibility functions $p(x) \propto \psi_{1,2,4}(x_1, x_2, x_4)\psi_{2,3,4}(x_2, x_3, x_4)$.

This theorem is illustrated in Figure 2-4. The converse is conservative in that $p(x_\Gamma)$ might actually factor with respect to some hypergraph $\mathbf{H}_\Gamma$ which is graphically equivalent to $\mathbf{G}_\Gamma$ but has lower-order interactions than in the maximal clique hypergraph cliq* $\mathbf{G}_\Gamma$. We will see this is the case for GMRFs. Note also that this confirms the converse of Proposition 1 since a random field which is locally Markov with respect to $\mathbf{G}_\Gamma$ factors with respect to $\mathbf{H}_\Gamma = \text{cliq } \mathbf{G}_\Gamma$ (by the above converse) and hence is globally Markov with respect to $\mathbf{G}_\Gamma = \text{adj } \mathbf{H}_\Gamma$ (by the first part of the proposition).

2.1.3 Gibbs Random Fields

In this subsection we discuss representation of graphical models as Gibbs random fields. Here, interaction potential functions take the place of the compatibility functions discussed previously. The connection between these formalisms is shown and a procedure for constructing potential specifications of Markov random fields is given. In later sections we connect this discussion to exponential families by considering parameterized families of Gibbs distributions and discuss representation of Gauss-Markov random fields from this point of view.

Gibbs distributions are named in honor of american physicist J. Willard Gibbs who, building upon the earlier work of Boltzmann [20], developed the formulation of statistical mechanics based on these probability distributions [62]. This formulation arises by considering the probability of states of a "partition" of a physical system where the entire system has equiprobable distribution over joint states (as in Boltzmann's formulation) and where the system is large relative to the partition (the so called "bulk limit"). This perspective is closely related to the maximum-entropy property of exponential families (subject to certain moment constraints) which we discuss later (Section 2.2.1). For an introduction to statistical mechanics, the reader is referred to the text of Bowley and Sánchez [21].

Definition 4 (Gibbs Distribution) *We say that* x_Γ *is a* Gibbs random field *if it has probability distribution* $p^\phi(x_\Gamma)$ *specified by a family* $\mathcal{H}_\Gamma^\phi$ *of subsets of* Γ *and a collection of functions* $\phi = (\phi_\Lambda : \mathcal{X}_\Lambda \to \mathcal{R}, \Lambda \subset \mathcal{H}_\Gamma^\phi)$ *as*

$$p^\phi(x_\Gamma) = \frac{1}{Z(\phi)} \exp\left\{ \sum_{\Lambda \in \mathcal{H}_\Gamma^\phi} \phi_\Lambda(x_\Lambda) \right\} \tag{2.7}$$

with $Z(\phi)$ *a finite normalizing constant. This* $p^\phi(x_\Gamma)$ *is* Gibbs distribution *with* potential specification ϕ. *The functions* ϕ_Λ *are* interaction potentials. *The normalizing* $Z(\phi)$ *is the* partition function *depending upon any parameters of the potentials (which, for the moment, we do not specify). We call* $\mathbf{H}_\Gamma^\phi = (\Gamma, \mathcal{H}_\Gamma^\phi)$ *the* interaction hypergraph *of the potential specification* ϕ.

Observe that the Gibbs distribution then factors with respect to the interaction hypergraph. Conversely, any positive probability distribution which factors with respect to hypergraph $\mathbf{H}_\Gamma$ may be expressed as a Gibbs distribution with potential specification ϕ such that $\mathbf{H}_\Gamma^\phi = \mathbf{H}_\Gamma$. Hence, identifying compatibility functions $\psi_\Lambda(x_\Lambda)$ with exponential factors $\exp\{\phi(x_\Lambda)\}$, we may restate the Hammersley-Clifford theorem as below.

Corollary 1 (Markov/Gibbs Near-Equivalence) *If* x_Γ *is a Gibbs random field having potential specification* ϕ*, then* x_Γ *is globally Markov with respect* $\mathbf{G}_\Gamma^\phi = \mathrm{adj}\ \mathbf{H}_\Gamma^\phi$*. Conversely, any (locally) Markov random field* $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ *with positive probability distribution,* $p(x_\Gamma) > 0$ *for all* $x_\Gamma \in \mathcal{X}_\Gamma$*, may be expressed as a Gibbs random field for some potential specification* ϕ *with* $\mathbf{H}_\Gamma^\phi = \mathrm{cliq}\ \mathbf{G}_\Gamma$*.*

The interpretation of potential functions is clarified by the following proposition which shows how the set of "local" potentials, intersecting some subfield $\Lambda \subset \Gamma$, determines the conditional distribution of that subfield given the state outside that subfield (or, equivalently, the state on just the boundary of that subfield).

Proposition 3 (Conditional Distribution) *For a Gibbs random field* x_Γ *with potential specification* ϕ*, the conditional distribution of subfield* x_Λ *conditioned on* $\mathrm{x}_{\partial\Lambda}$ *is given by*

$$p(x_\Lambda | x_{\partial\Lambda}) = \frac{1}{Z(x_{\partial\Lambda})} \exp\left\{ \sum_{H \in \mathcal{H}_\Gamma^\phi : \Lambda \cap H \neq \varnothing} \phi_H(x_H) \right\} \tag{2.8}$$

with normalizing function $Z(x_{\partial\Lambda})$ *given by the integral of the exponent over* $\mathcal{X}_\Lambda$*. Moreover, the conditional distribution* $p(x_\Lambda | x_{\setminus\Lambda})$ *only depends upon* $x_{\setminus\Lambda}$ *through* $x_{\partial\Lambda}$ *and is given by* $p(x_\Lambda | x_{\setminus\Lambda}) = p(x_\Lambda | x_{\partial\Lambda})$*.*

Proof. This is shown by computation of $p(x_\Lambda | x_{\setminus\Lambda})$ as $p(x_\Lambda, x_{\setminus\Lambda})/p(x_{\setminus\Lambda})$ and cancellation of common factors in the numerator and denominator. Let $\mathcal{H}_\Lambda \subset \mathcal{H}_\Gamma^\phi$ denote those hyperedges which intersect Λ and let $\mathcal{H}_{\setminus\Lambda} \subset \mathcal{H}_\Gamma^\phi$ denote those hyperedges which do not. Then,

$$\begin{aligned}
p(x_\Lambda | x_{\setminus\Lambda}) &= \frac{p(x_\Lambda, x_{\setminus\Lambda})}{\int p(\tilde{x}_\Lambda, x_{\setminus\Lambda})\, d\tilde{x}_\Lambda} \\
&= \frac{\exp\{\sum_H \phi_H(x_\Lambda, x_{\setminus\Lambda})\}}{\int \exp\{\sum_H \phi_H(\tilde{x}_\Lambda, x_{\setminus\Lambda})\} d\tilde{x}_\Lambda} \\
&= \frac{\exp\{\sum_{H \in \mathcal{H}_\Lambda} \phi_H(x_\Lambda, x_{\partial\Lambda})\} \exp\{\sum_{H \in \mathcal{H}_{\setminus\Lambda}} \phi_H(x_{\setminus\Lambda})\}}{\left(\int \exp\{\sum_{H \in \mathcal{H}_\Lambda} \phi_H(\tilde{x}_\Lambda, x_{\partial\Lambda})\} d\tilde{x}_\Lambda\right) \Big(\exp\{\sum_{H \in \mathcal{H}_{\setminus\Lambda}} \phi_H(x_{\setminus\Lambda})\Big)} \\
&= \frac{\exp\{\sum_{H \in \mathcal{H}_\Lambda} \phi_H(x_\Lambda, x_{\partial\Lambda})\}}{\int \exp\{\sum_{H \in \mathcal{H}_\Lambda} \phi_H(\tilde{x}_\Lambda, x_{\partial\Lambda})\} d\tilde{x}_\Lambda} \qquad (2.9)
\end{aligned}$$

By the Markov property, $p(x_\Lambda | x_{\setminus\Lambda}) = p(x_\Lambda | x_{\partial\Lambda})$ which completes the proof. □

In general, specification of the probability distribution $p(x_\Gamma)$ of a Markov random field $(x_\Gamma, \mathbf{G}_\Gamma)$ does not uniquely determine the Gibbsian potential specification

$(\mathbf{H}_\Gamma^\phi, \phi)$. We may generate many distinct Gibbs potential specifications $\phi = (\phi_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$ depending on how we split the interactions of the field among the potential functions. These distinct potential specifications correspond to different possible factorizations of the probability distribution. To remove this degeneracy, it is convenient to consider the following special choice of potential specification.

Canonical Potential Specification. We now give an explicit procedure for generating a potential specification $(\mathbf{H}_\Gamma^\phi, \phi)$ from an arbitrary positive probability distribution $p(x_\Gamma) > 0$. For a Markov random field $(x_\Gamma, \mathbf{G}_\Gamma)$, this procedure generates a sparse potential specification such that adj $\mathbf{H}_\Gamma^\phi = \mathbf{G}_\Gamma$. Furthermore, this choice of potential specification satisfies a certain normalization property with respect to a specified state $x_\Gamma^* \in \mathcal{X}_\Gamma$ and is the unique potential specification having this property. Hence, requiring this normalization forces uniqueness into the representation. We call this the *canonical potential specification.*

The canonical potentials are constructed relative to a specified state $x_\Gamma^* \in \mathcal{X}_\Gamma$ which we call the *ground state*. Relative to the chosen ground state, define the collection of functions $U = (U_\Lambda(x_\Lambda), \forall \Lambda \subset \Gamma)$ by

$$U_\Lambda(x_\Lambda) = \log \frac{p(x_\Lambda, x_{\setminus\Lambda}^*)}{p(x_\Gamma^*)} \tag{2.10}$$

where $p(x_\Lambda, x_{\setminus\Lambda}^*)$ is the value of the probability distribution $p(x_\Gamma)$ evaluated for the state $x_\Gamma = (x_\Lambda, x_{\setminus\Lambda}^*) \in \mathcal{X}_\Gamma$. These functions $U_\Lambda(x_\Lambda)$ measure the increase in log-likelihood due to a local departure x_Λ from the global ground state x_Γ^*. Next, define a second collection of functions $V = (V_\Lambda(x_\Lambda), \forall \Lambda \subset \Gamma)$ by

$$V_\Lambda(x_\Lambda) = \sum_{\Lambda' \subset \Lambda} (-1)^{|\Lambda \setminus \Lambda'|} U_{\Lambda'}(x_{\Lambda'}) \tag{2.11}$$

These functions V_Λ may be viewed as "irreducible" versions of the functions U_Λ where lower-order dependencies are recursively removed. By the *Möbius inversion lemma* (Möbius [98], Bazant [12, 13]), we may recover the collection U from the collection V by

$$U_\Lambda(x_\Lambda) = \sum_{\Lambda' \subset \Lambda} V_{\Lambda'}(x_{\Lambda'}). \tag{2.12}$$

Noting that $p(x_\Gamma) = p(x_\Gamma^*) \exp\{U_\Gamma(x_\Gamma)\}$, we have that

$$p(x_\Gamma) = \frac{1}{Z} \exp\left\{ \sum_{\Lambda \subset \Gamma} V_\Lambda(x_\Lambda) \right\} \tag{2.13}$$

with $Z^{-1} = p(x_\Gamma^*)$ such that V gives a Gibbs potential specification for $p(x_\Gamma)$ relative to the (apparently) complete hypergraph (where every subset of Γ is a hyperedge). Yet, the utility of this construction lies in that, for Markov random fields, this representation is actually sparse such that many of the potential functions in V are zero.

Proposition 4 (Sparsity of Canonical Potential) *For a (locally) Markov random field $(x_\Gamma, \mathbf{G}_\Gamma)$ with non-vanishing probability distribution, $p(x_\Gamma) > 0$ for all $x_\Gamma \in \mathcal{X}_\Gamma$, the collection $V = (V_\Lambda(x_\Lambda), \Lambda \subset \Gamma)$ is sparse such that, for any Λ which is* not *a clique in $\mathbf{G}_\Gamma$ the potential $V_\Lambda(x_\Lambda)$ is identically zero for all x_Λ.*

Proof. See Brémaud [25] or Lauritzen [88]. □

Hence, we define the *canonical potential specification* ϕ from V by collecting all non-zero potentials in V. This defines an interaction hypergraph $\mathbf{H}^\phi_\Gamma$ with hyperedges obtained by collecting all subsets $\Lambda \subset \Gamma$ where $V_\Lambda(x_\Lambda)$ is non-zero for some $x_\Lambda \in \mathcal{X}_\Lambda$.

$$\mathcal{H}^\phi_\Gamma = \{\Lambda \subset \Gamma | \exists x_\Lambda \in \mathcal{X}_\Lambda \ : \ V_\Lambda(x_\Lambda) \neq 0\} \tag{2.14}$$

By Proposition 4, this hypergraph has the property that adj $\mathbf{H}^\phi_\Gamma = \mathbf{G}_\Gamma$. Hence, this procedure generates a sparse potential specification exposing any Markov structure of the random field. Absorbing lower-order interactions into higher-order interactions gives an equivalent specification with respect to the maximal clique hypergraph cliq* $\mathbf{G}_\Gamma$ thus proving the converse of the Hammersley-Clifford theorem.

Another remarkable feature of the canonical potential specification is that, for the chosen ground state x^*_Γ, it gives the *unique* potential specification satisfying the following *normalization property.*

Proposition 5 (Normalization of Canonical Potential) *Among all potential specifications for $p(x_\Gamma)$ of the form $V = (V_\Lambda, \forall \Lambda \subset \Gamma)$, the canonical potential specification is the only specification having the property that $V_\Lambda(x_\Lambda) = 0$ whenever $(x_\Lambda)_\gamma = x^*_\gamma$ for some $\gamma \in \Lambda$.*

Proof. See Brémaud [25]. □

That is, each potential function $V_\Lambda(x_\Lambda)$ vanishes whenever the state of *any* site in Λ is set to its ground state value. The following is a consequence of this normalization property. We define the *partial potential specification* ϕ^Λ for each subfield $\Lambda \subset \Gamma$ as the collection of all interaction potentials defined within that subfield. That is,

$$\phi^\Lambda = (\phi_H, H \in \mathcal{H}^\phi_\Lambda) \tag{2.15}$$

where $\mathcal{H}^\phi_\Lambda = \{H \in \mathcal{H}^\phi_\Gamma | H \subset \Lambda\}$ are the hyperedges of the subhypergraph $\mathbf{H}^\phi_\Gamma$ induced by Λ. This *partial model* $(\mathbf{H}^\phi_\Lambda, \phi^\Lambda)$, the embedded graphical model based on the induced subhypergraph $\mathbf{H}^\phi_\Lambda$ and associated interaction potentials ϕ^Λ, specifies the conditional distribution of subfield x_Λ assuming ground-state boundary conditions.

$$p(x_\Lambda | x^*_{\partial\Lambda}) \propto \exp\{\phi^\Lambda(x_\Lambda)\} \equiv \exp\left\{\sum_{\Lambda' \subset \Lambda} \phi_{\Lambda'}(x_{\Lambda'})\right\} \tag{2.16}$$

This follows from Proposition 3 and 5 where those interactions with sites not contained in Λ vanish due to the normalization property. Conditioned on these boundary conditions, the Markov structure of the subfield is then specified by $\mathbf{G}^\phi_\Lambda =$ adj $\mathbf{H}^\phi_\Lambda$.

This interpretation plays a role in the RCM approach which we discuss further in Chapter 4.

We later construct the information parameterization for GMRFs from this perspective (Section 2.2.3). But first, we consider certain classes of parameterized Gibbs distributions which form exponential families.

2.1.4 Exponential Families

Exponential families and certain geometric characterizations of these parameterized families of probability distributions have been studied extensively by Chentsov [29], Efron [47], Barndorff-Nielsen [7], Amari [3] and many others.

An exponential family consists of probability distributions, for a variable (or collection of variables) $x \in \mathcal{X}$, having the form

$$f(x;\theta) = b(x)\exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{2.17}$$

The family is based on a positive distribution function $b(x) > 0$, a collection of statistics $t(x) \in \mathcal{R}^d$ and associated parameters $\theta \in R^d$ which scale the sensitivity to each statistic. The normalization constant $\varphi(\theta)$ is given by

$$\varphi(\theta) = \log \int b(x)\exp\{\theta \cdot t(x)\}\, dx \tag{2.18}$$

and, viewed as a function of θ, is called the *cumulant function.* The set of all *admissible* parameters Θ is defined by the effective domain of the cumulant function where $\varphi(\theta) < \infty$.

$$\Theta = \{\theta \in \mathcal{R}^d | \varphi(\theta) < \infty\} \tag{2.19}$$

The *exponential family* $\mathcal{F}$, for specified $(\mathcal{X}, b(\cdot), t(\cdot))$, is defined as the set of all such normalizable probability distributions.

$$\mathcal{F} = \{f(\cdot;\theta) | \theta \in \Theta\} \tag{2.20}$$

The family is said to be *regular* if the domain Θ contains an open subset of $\mathcal{R}^d$. This representation of the family is said to be *minimal* if the functions $((t_0(x) = 1, t(x)), \forall x \in \mathcal{X})$ are linearly independent. Then, admissible parameters Θ are in one-to-one correspondence with probability distributions $\mathcal{F}$, and the number of statistics (parameters) d is called the *dimension* of the family. To show that no two distinct choices or parameters $\theta_1 \neq \theta_2$ gives the same probability distribution, write the log-likelihood ratio of two such probability distributions as

$$\log \frac{f(x;\theta_1)}{f(x;\theta_2)} = \vec{a} \cdot (1, t(x)) \tag{2.21}$$

with $\vec{a} = (\varphi(\theta_2) - \varphi(\theta_1), \theta_1 - \theta_2) \neq 0$. If the two probability distributions are identical then we must have that $\vec{a} \cdot (1, t(x)) = 0$ for all $x \in \mathcal{X}$. This contradicts the presumed linear independence of the functions $(1, t(\cdot))$. Hence, assuming minimality, the admissible parameters Θ gives a finite-dimensional coordinate system for the exponential family of probability distributions $\mathcal{F}$. The parameters θ are sometimes called *exponential coordinates*. We now discuss some important properties of minimal representations of regular exponential families.

Moments and Fisher Information. First, we discuss the *moment-generating property* of the cumulant function. As is easily verified, differentiation of $\varphi(\theta)$ gives the expected value of the statistics $t(x)$.

$$\frac{\partial \varphi(\theta)}{\partial \theta_i} = E_\theta\{t_i(\mathrm{x})\} \tag{2.22}$$

These expectations $\eta = E_\theta\{t(x)\}$ are called the *moments* of the distribution. We call $\eta(\Theta) \equiv \{\eta \in \mathcal{R}^d | \exists \theta \in \Theta : E_\theta\{t(x)\} = \eta\}$ the *achievable moments*. Taking second-order partial derivatives gives the covariance of the statistics:

$$\frac{\partial^2 \varphi(\theta)}{\partial \theta_i \partial \theta_j} = E_\theta\{(t_i(\mathrm{x}) - \eta_i)(t_j(\mathrm{x}) - \eta_j)\} \tag{2.23}$$

Let $G(\theta) = (g_{i,j}(\theta))$ denote the Hessian matrix obtained by collecting these second-order partial derivatives. This is the curvature of $\varphi(\cdot)$ at the point θ. An equivalent expression for $G(\theta)$ is

$$g_{i,j}(\theta) = E_\theta \left\{ \frac{\partial \log p(\mathrm{x};\theta)}{\partial \theta_i} \frac{\partial \log p(\mathrm{x};\theta)}{\partial \theta_j} \right\}, \tag{2.24}$$

This is the *Fisher information* associated with the parameterized family $f(x;\theta)$ which plays a fundamental role in the theory of parameter estimation.[3] Note also, since first-order differentiation gives the moments, second-order differentiation gives the sensitivity of moments to parameters:

$$g_{i,j}(\theta) = \frac{\partial \eta_i}{\partial \theta_j}, \tag{2.25}$$

Hence, $G(\theta)$ may also be interpreted as the Jacobian matrix $\partial \eta / \partial \theta$ of the mapping from parameters θ to moments η.

Since the curvature matrix is given by a positive semidefinite covariance matrix (positive definite assuming minimal representation), this shows that the cumulant function is a (strictly) convex function of θ over its effective domain Θ. That is, for all $\theta_1, \theta_2 \in \Theta$ and $\lambda \in [0,1]$ we have

$$\varphi(\lambda\theta_1 + (1-\lambda)\theta_2) \leq \lambda\varphi(\theta_1) + (1-\lambda)\varphi(\theta_2)$$

[3] By the Cramer-Rao inequality, the (appropriately scaled) inverse Fisher information gives a lower bound estimate for the uncertainty of parameter estimates based on a collection of independent samples of $\mathrm{x} \sim f(x;\theta)$. Roughly speaking, the Fisher information characterizes the information gained about θ per observation of x.

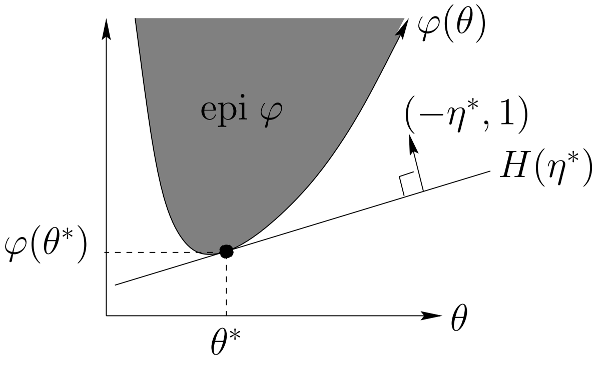


Figure 2-5: Illustration of geometric argument showing one-to-one correspondence between parameters θ and moments η. Each $\eta^* \in \eta(\Theta)$ corresponds to a unique supporting hyperplane $H(\eta^*)$ of the epigraph $E = \text{epi } \varphi$ with normal vector $(-\eta^*, 1)$. These supporting hyperplanes are tangent to the epigraph satisfying the relation $\eta^* = \nabla\varphi(\theta)$ with respect to any intersection point $(\theta, \varphi(\theta)) \in H(\eta^*) \cap E$. Assuming minimal representation, E is strictly convex such that each tangential supporting hyperplane intersects E at just one point $(\theta^*, \varphi(\theta^*)) = H(\eta^*) \cap E$. Hence, each $\eta^* \in \eta(\Theta)$ is generated by a unique choice of parameters $\theta^* \in \Theta$.

For minimal representations, the convexity is strict such that, for $\theta_1 \neq \theta_2$, equality occurs only at the endpoints $\lambda \in \{0, 1\}$. Finally, we remark that the cumulant function is *essentially smooth*. This means that $\varphi(\theta)$ is a continuous, smooth function of θ over its effective domain Θ and is *steep* in that it diverges to infinity near the boundary of Θ. This property, together with convexity, makes the cumulant function very well-suited for convex analysis.

A crucial point is that, for minimal representations, the mapping from admissible parameters Θ to achievable moments $\eta(\Theta)$ is in fact one-to-one. This is shown by the following geometric argument illustrated in Figure 2-5. Consider the *epigraph* of the function $\varphi(\theta)$, defined as epi $\varphi = \{(\theta, h) | h \geq \varphi(\theta)\} \subset \mathcal{R}^{d+1}$. We say that a d-dimensional hyperplane H is a *supporting hyperplane* of $E = \text{epi } \varphi$ if it intersects E at some point while containing E in its upper closed half-space. Due to the essential smoothness and convexity of $\varphi(\theta)$, the set of supporting hyperplanes $\mathcal{H}$ is generated by $\{H(\eta), \eta \in \eta(\Theta)\}$ where $H(\eta)$ is the unique supporting hyperplane of E normal to $(-\eta, 1)$. Furthermore, the set of supporting hyperplanes is precisely the set of d-dimensional tangent hyperplanes $\{T(\theta), \theta \in \Theta\}$ where $T(\theta)$ is tangent to E at the point $(\theta, \varphi(\theta))$ and, hence, normal to the vector $(-\nabla\varphi(\theta), 1)$. By the moment-generating property, we have $H(\eta) = T(\theta)$ where $\eta = \nabla\varphi(\theta)$. Assuming strict convexity (minimal representation), each tangent hyperplane $T(\theta)$ intersects E at just one point so that the mapping from admissible parameters Θ to supporting hyperplanes $\mathcal{H} = \{T(\theta)\}$ is one-to-one. Likewise, for the mapping from parameters Θ to achievable moments $\eta(\Theta)$. Hence, the moments η are sometimes called *moment coordinates*. The set of achievable moments $\eta(\Theta)$ provides an alternative finite-dimensional coordinate system for the exponential family $\mathcal{F}$. We indicate this parameterization of the exponential family by $f^*(x;\eta) \equiv f(x;\theta(\eta))$. Also, we let $G^*(\eta) = (g^*_{i,j}(\eta))$ denote the Fisher information,

$$g^*_{i,j}(\eta) = E_\eta \left\{ \frac{\partial \log f^*(\mathrm{x};\eta)}{\partial \eta_i} \frac{\partial \log f^*(\mathrm{x};\eta)}{\partial \eta_j} \right\}, \tag{2.26}$$

with respect to this moment parameterization $f^*(x;\eta)$.

Convex Duality. Convex duality plays a central role in the geometric analysis of exponential families. In Fenchel duality (Fenchel [49, 50], Rockafellar [114, 115], Bertsekas et al [17]), the *convex conjugate function* of $\varphi(\theta)$ is defined as:

$$\varphi^*(\beta) = \sup_{\theta \in \Theta} \{\beta \cdot \theta - \varphi(\theta)\} \tag{2.27}$$

For, essentially smooth and convex $\varphi(\theta)$, Fenchel duality reduces to Legendre duality (Rockafellar [114], Bauschke and Borwein [11], Bauschke et al [10]). Consider minimizing $f(\theta) = \varphi(\theta) - \beta \cdot \theta$ (a "tilted" version of the cumulant function) with respect to $\theta \in \Theta$. This is a strictly convex, essentially smooth function such that if there exists any local minimum $\theta^* \in \Theta$ then this is also the (unique) global minimum. Setting the gradient to zero, $\nabla f(\theta) = \eta(\theta) - \beta = 0$, gives the necessary and sufficient condition $E_{\theta^*}\{t(x)\} = \beta$ for θ^* to be the global minimum. Hence, for $\beta \in \eta(\Theta)$, there exists a unique local minimum (this being the global minimum) θ^* and β then gives the corresponding moments $\eta^* = E_{\theta^*}\{t(\mathrm{x})\}$. We may write the convex conjugate function for $\eta \in \eta(\Theta)$ as

$$\varphi^*(\eta) = \eta \cdot \theta - \varphi(\theta) \tag{2.28}$$

where θ and η are dually-coupled by the condition $E_\theta\{t(x)\} = \eta$. This is known as the *Legendre transform* (Rockafellar [114]). It may also be shown that φ^* is an essentially smooth, strictly convex function with effective domain $\eta(\Theta)$. When $\varphi(\theta)$ is strictly convex, so is $\varphi^*(\eta)$. This pair of functions (φ, φ^*) satisfy a variety of useful duality relations which we summarize.

First, the previously discussed bijective coordinate transformation between exponential coordinates θ and moment coordinates η may now be characterized by the following dual differential relations:

$$\frac{\partial \varphi(\theta)}{\partial \theta_i} = \eta_i \tag{2.29}$$

$$\frac{\partial \varphi^*(\eta)}{\partial \eta_i} = \theta_i \tag{2.30}$$

This shows that the cumulant function $\varphi(\theta)$ and the convex conjugate function $\varphi^*(\eta)$ are dually related by a "slope transform" as illustrated in Figure 2-6. The first relation is just the moment-generating property of the cumulant function. The second relation shows that exponential coordinates are recovered from moment coordinates by evaluating the gradient of $\varphi^*(\eta)$ at those moment coordinates.

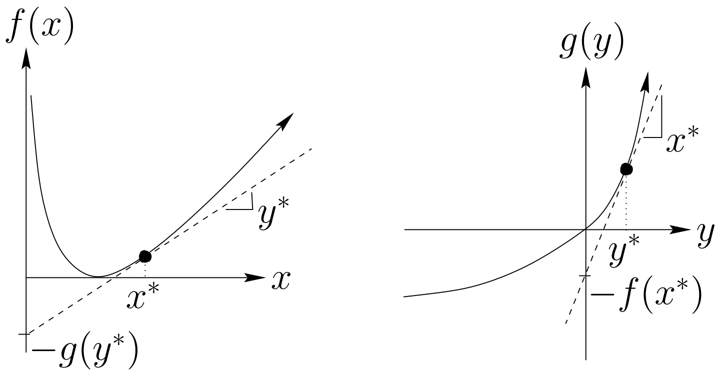


Figure 2-6: Depiction of Legendre transform between a dual pair of convex-conjugate functions. This example depicts the functions $f(x) = (x - 1) - \log x$ (for $x > 0$) and $g(y) = -\log(1 - y)$ (for $y < 1$) but is not drawn to scale. Coordinates (x^*, y^*) are dually-coupled, such that $x^* = \arg\min_x\{xy^* - f(x)\}$ ($y^* = \arg\min_y\{x^*y - g(y)\}$), if and only if $x^* = g'(y^*)$ ($y^* = f'(x^*)$). Then, $f(x^*) + g(y^*) = x^*y^*$ such that the zero-intercepts of tangents gives minus the value of the conjugate function.

Second, the curvature of the two functions $\varphi(\theta)$ and $\varphi^*(\eta)$ are related by the dual relations

$$\frac{\partial^2\varphi(\theta)}{\partial\theta_i\partial\theta_j} = g_{i,j}(\theta) \tag{2.31}$$

$$\frac{\partial^2\varphi^*(\eta)}{\partial\eta_i\partial_j\eta_j} = g^*_{i,j}(\eta) \tag{2.32}$$

where $G^*(\eta) = G^{-1}(\theta)$ for dually-coupled exponential/moment coordinates (θ, η). That is, the curvature of $\varphi^*(\eta)$ is the Fisher information in the moment parameterization of the family which is inversely related to the curvature (Fisher information) within the exponential parameterization. Also, we have that $G^*(\eta) = \partial\theta/\partial\eta$, the Jacobian matrix of the inverse transform taking moment coordinates back to exponential coordinates. For minimal representations, this shows that the curvature $G^*(\eta)$ is also positive definite for all $\eta \in \eta(\Theta)$. Hence, $\varphi^*(\eta)$ is shown to be strictly convex over its domain $\eta(\Theta)$.

Due to the convexity of $\varphi^*(\eta)$, we may write $\varphi(\theta) = \theta\cdot\eta - \varphi^*(\eta)$, viewing this as the Legendre transform of $\varphi^*(\eta)$ which recovers the original function $\varphi(\theta)$. Equivalently,

$$\varphi(\theta) = \sup_{\beta\in\eta(\Theta)} \{\beta \cdot \theta - \varphi^*(\beta)\} \tag{2.33}$$

showing the duality between $\varphi(\theta)$ and $\varphi^*(\eta)$. We later see, in Section 2.2.1, that $\varphi^*(\eta)$ may be interpreted as a measure of the *information content* of the probability distribution $f^*(\cdot;\eta)$ relative to the base distribution $b(x)$.

Exponential Family Graphical Models. As in Della Pietra et al [106], we approach graphical modeling by considering parameterized families of Gibbs distributions which form exponential families. This is accomplished by restricting the interaction potentials of the Gibbs distributions to have the form $\phi_\Lambda(x_\Lambda; \theta_\Lambda) = \theta_\Lambda \cdot t_\Lambda(x_\Lambda)$ based on a set of local state statistics $t_\Lambda(x_\Lambda)$ and associated exponential parameters θ_Λ which scale the sensitivity to each statistic. Collecting these local statistics for all interaction potentials gives a global vector statistic $t(x) = (t_\Lambda(x_\Lambda), \Lambda \in \mathcal{H}^\phi_\Gamma)$. Likewise, we collect parameters $\theta = (\theta_\Lambda, \Lambda \in \mathcal{H}^\phi_\Gamma)$ such that $\theta \cdot t(x) = \sum_\Lambda \theta_\Lambda \cdot t_\Lambda(x_\Lambda)$. The (parameterized) Gibbs distribution then has the form of an exponential model $p(x) \propto \exp\{\theta \cdot t(x)\}$ (with base distribution $b(x) = 1$ for all x).

We also define the collections of statistics $t^\Lambda(x_\Lambda) = (t_{\Lambda'}(x_{\Lambda'}), \Lambda' \subset \Lambda)$, exponential parameters $\theta^\Lambda = (\theta_{\Lambda'}, \Lambda' \subset \Lambda)$ and moment parameters $\eta^\Lambda = (\eta_{\Lambda'}, \Lambda' \subset \Lambda)$. These are obtained by collecting all statistics (parameters) defined on Λ or any subset of Λ. Note that $(\theta^\Lambda, t^\Lambda)$ specifies the partial potential specification $\phi^\Lambda(x_\Lambda) = \exp\{\theta^\Lambda \cdot t^\Lambda(x_\Lambda)\}$. Also, we say that the family $\mathcal{F}$ is *marginalizable* when, for all $p \in \mathcal{F}$ and all cliques of the adjacency graph $\Lambda \in \mathcal{C}(\mathbf{G}^\phi_\Gamma)$, the marginal distribution $p(x_\Lambda) = \int p(x_\Gamma) dx_{\setminus\Lambda}$ has the form of an exponential distribution based on $t^\Lambda(x_\Lambda)$. That is $p(x_\Lambda) \propto \exp\{\beta_\Lambda \cdot t^\Lambda(x_\Lambda)\}$ for some choice of parameters β_Λ depending upon p. Then, $\eta^\Lambda = E_\theta\{t^\Lambda(\mathrm{x}_\Lambda)\}$ are the marginal moment coordinates uniquely specifying the marginal distribution $p(x_\Lambda)$.

Two fundamental problems of graphical modeling may be stated in terms coordinate transforms within the exponential family. We consider the *inference problem* as calculation of the moments $\eta^* = E_{\theta^*}\{t(\mathrm{x})\}$ given the parameters $\theta^* \in \Theta$. We consider the *modeling problem* as the inverse problem of recovering the parameters θ^* given the moments $\eta^* \in \eta(\Theta)$. In light of convex duality, both problems are endowed with the following dual variational interpretations.

$$\begin{aligned} \text{(Inference)}\ \ \eta^* &= \arg\min_\eta \{\varphi^*(\eta) - \theta^* \cdot \eta\} && (2.34) \\ \text{(Modeling)}\ \ \theta^* &= \arg\min_\theta \{\varphi(\theta) - \theta \cdot \eta^*\} && (2.35) \end{aligned}$$

These are both convex programming problems, minimizing a strictly convex function over a convex set, and each has a unique solution. We shall see that the modeling problem arises in the context of maximum-likelihood parameter estimation (Section 2.2.2). Also, in the context of information geometry, these variational problems correspond to dual notions of information projection (Section 2.2.3). Finally, we remark that calculation of the moments η is often considered as interchangeable with computation of the marginal distributions $(p(x_\Lambda), \Lambda \in \mathcal{H}^\phi_\Gamma)$. This is because we may either determine each marginal $p(x_\Lambda)$ from η^Λ (independently solving the modeling problem within each marginal exponential family) or we may determine η^Λ from $p(x_\Lambda)$ (inference in the marginal families).

Marginalizable exponential families of graphical models play a central role in the graphical modeling literature. We discuss this further both in the context of iterative modeling methods (Section 2.2.4) and recursive inference methods (Section 2.3).

2.1.5 Gauss-Markov Random Fields

A *Gauss-Markov random field* (GMRF) is a MRF $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ where the state $x \in R^n$ (suppressing the Γ subscript for notational brevity) is normally distributed $\mathrm{x} \sim \mathcal{N}(\hat{x}, P)$ having probability distribution

$$p(x) = \frac{1}{\sqrt{\det 2\pi P}} \exp\left\{-\frac{1}{2}(x-\hat{x})' P^{-1}(x-\hat{x})\right\} \tag{2.36}$$

This Gaussian distribution satisfies the expectation relations $E_p\{\mathrm{x}\} = \hat{x}$ and $E_p\{(\mathrm{x}-\hat{x})(\mathrm{x}-\hat{x})'\} = P$. Thus, the *moment parameters* $(\hat{x}, P)$ specify the *mean vector* $\hat{x}$ and the *covariance matrix* P of the random variable x. The covariance is a symmetric positive semidefinite matrix. We say that the GMRF is *regular* if the covariance is also positive definite. The marginal distributions of a Gaussian distribution are also Gaussian. Hence, the marginal distribution $p(x_\Lambda)$ of the state x_Λ of subfield $\Lambda \subset \Gamma$ is determined by the marginal moments $(\hat{x}_\Lambda, P_\Lambda)$ (the appropriate subvector and submatrix of the global mean $\hat{x}$ and covariance P).

The *information parameterization* $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$ is defined by:

$$\begin{aligned} h &= P^{-1}\hat{x} \quad &(2.37)\\ J &= P^{-1} \quad &(2.38) \end{aligned}$$

For regular GMRFs, the moment parameters may be recovered from the information parameters by $(\hat{x}, P) = (J^{-1}h, J^{-1})$. The information form of the Gaussian distribution is given by

$$p(x) = \exp\left\{-\frac{1}{2}x'Jx + h'x - \varphi(h, J)\right\} \tag{2.39}$$

where

$$\varphi(h, J) = \frac{1}{2}\left\{h'J^{-1}h - \log\det J + n\log 2\pi\right\} \tag{2.40}$$

This may be viewed as an exponential distribution with statistics $t(x) = (x, xx')$, parameters $\theta = (h, -J/2)$, moments $\eta = (\hat{x}, P + \hat{x}\hat{x}')$ and cumulant function $\varphi(\theta) = \varphi(h, J)$. The marginal distributions $\mathrm{x}_\Lambda \sim \mathcal{N}(\hat{x}_\Lambda, P_\Lambda)$ may also be represented in this information form $x_\Lambda \sim \mathcal{N}^{-1}(\hat{h}_\Lambda, \hat{J}_\Lambda)$ with marginal information parameters $(\hat{h}_\Lambda, \hat{J}_\Lambda) = ((P_\Lambda)^{-1}\hat{x}_\Lambda, (P_\Lambda)^{-1})$. Applying the matrix inversion lemma, marginal information parameters $(\hat{h}_\Lambda, \hat{J}_\Lambda)$ are related to global information parameters (h, J) by

$$\begin{aligned} \hat{h}_\Lambda &= h_\Lambda - J_{\Lambda,\backslash\Lambda}(J_{\backslash\Lambda})^{-1}h_{\backslash\Lambda} \quad &(2.41)\\ \hat{J}_\Lambda &= J_\Lambda - J_{\Lambda,\backslash\Lambda}(J_{\backslash\Lambda})^{-1}J_{\backslash\Lambda,\Lambda} \quad &(2.42) \end{aligned}$$

We shall see that this provides a fundamental basis for recursive inference in GMRFs. As is well known from the literature, the inverse covariance matrix J reflects the Markov structure of the field through its sparsity pattern (Speed & Kiiveri[124], Lauritzen[88]). We demonstrate this in the course of the following development.

Gibbsian Description. We now show the connection between the information form of the Gaussian distribution and the canonical potentials of a GMRF. We do so by deriving the canonical potentials for a GMRF working from the information form of the density. Here, we choose the zero ground state $x^*_\Gamma = 0$. Then the functions $U_\Lambda(x_\Lambda)$ (2.10) are given by

$$U_\Lambda(x_\Lambda) = -\frac{1}{2}x'_\Lambda J_\Lambda x_\Lambda + h'_\Lambda x_\Lambda \tag{2.43}$$

Now we calculate the canonical potentials $V_\Lambda(x_\Lambda)$ (2.11). For $\Lambda = \{\gamma\}$, we have the singleton canonical potentials $V_\gamma(x_\gamma) = U_\gamma(x_\gamma)$:

$$V_{\{\gamma\}}(x_\gamma) = -\frac{1}{2}x'_\gamma J_\gamma x_\gamma + h'_\gamma x_\gamma \tag{2.44}$$

For $\Lambda = \{\gamma, \lambda\}$ we have the pairwise canonical potentials:

$$\begin{aligned} V_{\{\gamma,\lambda\}}(x_\gamma, x_\lambda) &= U_{\{\gamma,\lambda\}}(x_\gamma, x_\lambda) - U_\gamma(x_\gamma) - U_\lambda(x_\lambda) \\ &= -x'_\gamma J_{\gamma,\lambda} x_\lambda \end{aligned} \tag{2.45}$$

for all $\{\gamma, \lambda\} \subset \Gamma$ such that $J_{\gamma,\lambda} \neq 0$.

All higher order potentials must vanish. This may be seen from the normalization property of the canonical potential as follows. The functions U_Λ, and hence V_Λ as well, consist entirely of linear and quadratic terms in the the variables x_γ for all $\gamma \in \Lambda$. Hence, V_Λ must have the form:

$$V_\Lambda(x_\Lambda) = \sum_{\gamma\in\Lambda} A_\gamma x_\gamma + \sum_{\gamma,\lambda\in\Lambda} x'_\gamma B_{\gamma,\lambda} x_\lambda \tag{2.46}$$

For $|\Lambda| > 1$, setting all states except x_γ to zero shows that we must have $A_\gamma = 0$ in order for the normalization to hold. Likewise for $|\Lambda| > 2$, setting all variables except x_γ and x_λ to zero shows that $B_{\gamma,\lambda} = 0$. Hence, $|\Lambda| > 2$ implies $V_\Lambda(x_\Lambda) = 0$ such that all interactions in the (sparse) canonical potential ϕ are either singleton effects or pairwise interactions among the sites of the field. This then provides for the following result.

Proposition 6 (Sparsity of J-matrix) *Let* $\mathrm{x}_\Gamma \sim \mathcal{N}^{-1}(h, J)$ *be a regular Gaussian random field. Then,* x_Γ *is Markov with respect to* $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ *with edges* $\mathcal{E}_\Gamma = \{\{\gamma, \lambda\} \subset \Gamma | J_{\gamma,\lambda} \neq 0\}$. *Conversely, if* x_Γ *is Markov with respect to some graph* $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ *then* $J_{\gamma,\lambda} = 0$ *for all* $\{\gamma, \lambda\} \notin \mathcal{E}_\Gamma$.

Proof. The first part of the theorem follows from the Proposition 2, since $\mathbf{G}_\Gamma$ is defined so as to insure that $p(x_\Gamma)$ factors with respect to $\mathbf{G}_\Gamma$ and is hence Markov with respect to $\mathbf{G}_\Gamma = \text{adj } \mathbf{G}_\Gamma$. The converse follows from sparsity of the canonical potentials (Proposition 4), $\phi_\Lambda(x_\Lambda)$ is zero for all x_Λ whenever Λ is not a clique in $\mathbf{G}_\Gamma$. As shown above, $\phi_{\gamma,\lambda}(x_\gamma, x_\lambda) = -x_\gamma J_{\gamma,\lambda} x_\lambda$ so that for $\{\gamma, \lambda\} \notin \mathcal{E}_\Gamma$ we must have $J_{\gamma,\lambda} = 0$. □

A more generic statement of the above result is given by the following corollary. This may be regarded as a stronger version of the Hammersley-Clifford theorem specific to Gaussian random fields.

Corollary 2 (Gauss-Markov Equivalence) *A regular Gaussian random field* x_Γ *is Markov with respect to* $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ *if and only if it has probability distribution which factors with respect to* $\mathbf{H}_\Gamma = (\Gamma, \mathcal{E}_\Gamma \cup \{\{\gamma\} | \gamma \in \Gamma\})$.

Remark. Note that the hypergraph $\mathbf{H}_\Gamma$ is just the graph $\mathbf{G}_\Gamma$ but augmented with singleton hyperedges (this is just a formality to accommodate those singleton potentials $\phi_{\{\gamma\}}$ in the case of isolated vertices, otherwise singleton potentials can be absorbed into pairwise potentials). The corollary is stronger than the converse of Hammersley-Clifford which, in general, only insures factorization with respect to the clique hypergraph $\mathbf{H}_\Gamma = \text{cliq } \mathbf{G}_\Gamma$. The specialization occurs because interactions among sites for Gaussian random fields are fundamentally pairwise.

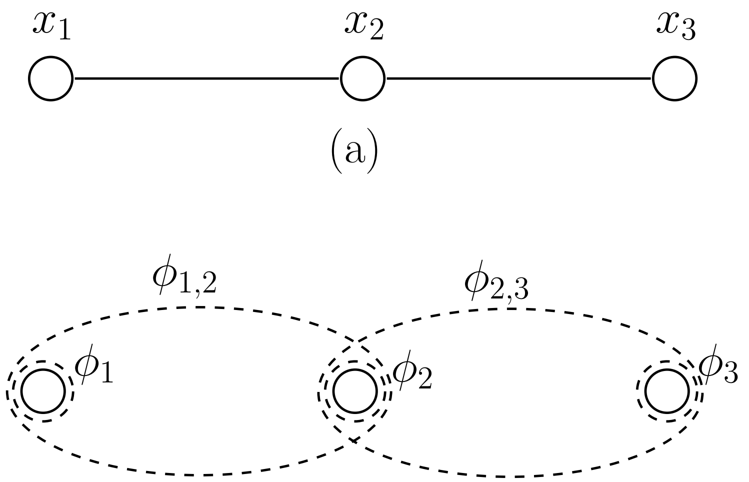
Thus, the information parameterization (h, J) of the GMRF provides a natural graphical model $(\mathbf{G}_\Gamma, \phi)$ of the GMRF. The graphical structure of the field $\mathbf{G}_\Gamma$, showing both the Markov structure and the interaction structure of the field, is determined by the non-zero off-diagonal entries of the matrix J. The Gibbsian canonical potential specification ϕ is based on the non-zero entries of h (which specify linear singleton effects) and J (where diagonal entries specify quadratic singleton effects and off-diagonal entries specify pairwise interactions). This interpretation of the information parameterization is illustrated in Figure 2-7 for a simple GMRF with only three sites.

In view of this Gibbsian interpretation of (h, J), conditional distributions of subfields (conditioned on the state on the boundary) are given by local potentials specified by the appropriate subset of information parameters. According to Proposition 3, to calculate the conditional probability distribution $p(x_\Lambda | x_{\partial\Lambda})$ we collect all potentials which depend upon the state of any site in Λ. This gives

$$\begin{aligned} p(x_\Lambda | x_{\partial\Lambda}) &\propto \exp\{-\frac{1}{2} x'_\Lambda J_\Lambda x_\Lambda + h'_\Lambda x_\Lambda - x'_{\partial\Lambda} J_{\partial\Lambda,\Lambda} x_\Lambda\} \\ &\propto \exp\{-\frac{1}{2} x'_\Lambda J_\Lambda x_\Lambda + h_{\Lambda|\partial\Lambda}(x_{\partial\Lambda})' x_\Lambda\} \end{aligned} \tag{2.47}$$

where $h_{\Lambda|\partial\Lambda}(x_{\partial\Lambda}) \equiv h_\Lambda - J_{\Lambda,\partial\Lambda} x_{\partial\Lambda}$ are the updated singleton effects after conditioning on the state of the boundary $x_{\partial\Lambda}$. This is seen to be a Gaussian distribution $\mathrm{x}_\Lambda | x_{\partial\Lambda} \sim \mathcal{N}^{-1}(h_{\Lambda|\partial\Lambda}(x_{\partial\Lambda}), J_\Lambda)$ with conditional mean $\hat{x}_{\Lambda|\partial\Lambda} \equiv E\{\mathrm{x}_\Lambda | x_{\partial\Lambda}\} = J_\Lambda^{-1} \hat{h}_\Lambda(x_{\partial\Lambda})$ and conditional covariance $P_{\Lambda|\partial\Lambda} \equiv E\{(\mathrm{x}_\Lambda - \hat{x}_{\Lambda|\partial\Lambda})(\mathrm{x}_\Lambda - \hat{x}_{\Lambda|\partial\Lambda})' | x_{\partial\Lambda}\} = J_\Lambda^{-1}$. Note that only the conditional mean depends upon the state of the boundary, the conditional covariance J_Λ^{-1} is constant with respect to $x_{\partial\Lambda}$. Also, the partial potential specification ϕ^Λ, parameterized by (h_Λ, J_Λ), gives the conditional distribution of the subfield assuming zero boundary conditions.

$$p(x_\Lambda | 0) \propto \exp \phi^\Lambda(x_\Lambda) = \exp\{-\frac{1}{2} x'_\Lambda J_\Lambda x_\Lambda + h'_\Lambda x_\Lambda\} \tag{2.48}$$



(a)

$$\begin{aligned}
\phi_1(x_1) &= -\frac{1}{2}x_1' J_{1,1} x_1 + h_1' x_1 \\
\phi_2(x_2) &= -\frac{1}{2}x_2' J_{2,2} x_2 + h_2' x_2 \\
\phi_3(x_3) &= -\frac{1}{2}x_3' J_{3,3} x_3 + h_3' x_3
\end{aligned}$$

$$\begin{aligned}
\phi_{1,2}(x_1, x_2) &= -x_1' J_{1,2} x_2 \\
\phi_{2,3}(x_2, x_3) &= -x_2' J_{2,3} x_3
\end{aligned}$$

(b)

$$h = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix}, \quad J = \begin{pmatrix} J_{1,1} & J_{1,2} & 0 \\ J_{1,2}' & J_{2,2} & J_{2,3} \\ 0 & J_{2,3}' & J_{3,3} \end{pmatrix}$$

(c)

Figure 2-7: Illustration of a simple 3-site GMRF: (a) Markov structure imposing conditional independency $p(x_1, x_3|x_2) = p(x_1|x_2)p(x_3|x_2)$, (b) Gibbsian potential specification $p(x) \propto \exp \sum_\Lambda \phi_\Lambda(x_\Lambda)$, and (c) information representation $p(x) \propto \exp\{-\frac{1}{2}x'Jx + h'x\}$. The probability distribution respects the Markov structure shown in (a) if and only if the interaction matrix respects the sparsity constraints shown in (c).

This is to be expected in view of the interpretation of (h, J) as the canonical potentials constructed relative to zero ground state $x^* = 0$.

Exponential Family Description. As our final task of the section, we describe the information parameterization (h, J) of a GMRF $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$ as a minimal representation of a regular exponential family. This follows very naturally from the canonical potentials derived previously based on the non-zero entries of (h, J). Without any loss of generality, we consider the case where all states x_γ are scalar-valued.[4]

Each singleton potential $\phi_\gamma(x_\gamma) = -\frac{1}{2}J_\gamma x_\gamma^2 + h_\gamma x_\gamma$ may be expressed as $\phi_\gamma(x_\gamma) = \theta_\gamma \cdot t_\gamma(x_\gamma)$ with statistics and parameters defined as

$$t_\gamma(x_\gamma) = (x_\gamma, x_\gamma^2) \tag{2.49}$$
$$\theta_\gamma = (h_\gamma, -J_\gamma/2) \tag{2.50}$$

for all $\gamma \in \Gamma$. Each pairwise potential $\phi_{\gamma,\lambda}(x_\gamma, x_\lambda) = -J_{\gamma,\lambda}x_\gamma x_\lambda$, defined for all edges $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$, may be expressed as $\phi_{\gamma,\lambda}(x_\gamma, x_\lambda) = \theta_{\gamma,\lambda}t_{\gamma,\lambda}(x_\gamma, x_\lambda)$ with statistics and parameters defined by

$$t_{\gamma,\lambda}(x_\gamma, x_\lambda) = x_\gamma x_\lambda \tag{2.51}$$
$$\theta_{\gamma,\lambda} = -J_{\gamma,\lambda} \tag{2.52}$$

for all $\{\gamma, \lambda\} \in \mathcal{G}_\Gamma$. Then, let $t(x)$ and θ denote the collection all such statistics and parameters such that $\theta \cdot t(x) = -\frac{1}{2}x'Jx + h'x$. The information form of the Gaussian distribution is then given by the exponential distribution $p(x) = \exp\{\theta \cdot t(x) - \varphi(\theta)\}$ with cumulant function $\varphi(\theta) = \varphi(h, J)$ as in (2.40). The cumulant function $\varphi(h, J)$ is finite provided the interaction matrix J is positive definite (the GMRF is regular). This condition determines the space of admissible parameters Θ and a corresponding exponential family $\mathcal{F}$ of normalizable probability distributions. The dimension of the family (number of linearly independent statistics) is $d = 2|\Gamma| + |\mathcal{E}_\Gamma|$. The dual moment parameters $\eta = E_p\{t(x)\}$ are given by

$$\eta_\gamma = (\hat{x}_\gamma, P_\gamma + \hat{x}_\gamma^2),\ \forall\gamma \in \Gamma \tag{2.53}$$
$$\eta_{\gamma,\lambda} = P_{\gamma,\lambda} + \hat{x}_\gamma\hat{x}_\lambda,\ \forall\{\gamma, \lambda\} \in \mathcal{E}_\Gamma \tag{2.54}$$

Note, these moments η directly specify a subset of the Gaussian moment parameters $(\hat{x}, P)$. The unspecified off-diagonal entries of P, corresponding to the zeros of J, are determined by the condition that the distribution is Markov with respect to $\mathbf{G}_\Gamma$. In principle, these remaining elements could be determined by maximum-entropy completion (see Frakt [55] and Tucker [128]).

This exponential family provides a minimal representation of the family of all regular Gaussian random fields x_Γ which are Markov with respect to $\mathbf{G}_\Gamma$. This identification is useful insofar as we may then apply methods of information geometry for

[4] Otherwise, for states $x_\gamma \in R^{n_\gamma}$ with $n_\gamma > 1$, replace site γ by n_γ sites, each having scalar-valued states, and couple these sites by pairwise interactions according to the non-zero entries of the $n_\gamma \times n_\gamma$ submatrix J_γ.

Gaussian random fields. Selected aspects of the information geometry of exponential families are discussed in the next section. We also remark that the preceding information description of GMRFs, as exponential families of graphical models based on $\mathbf{G}_\Gamma$, is *marginalizable* in the sense that the marginal distributions for any clique $\Lambda \in \mathcal{C}(\mathbf{G}_\Gamma)$ is contained in the exponential family $\mathcal{F}^\Lambda = \{p(x_\Lambda) \propto \exp\{-\frac{1}{2}x'_\Lambda \hat{J}_\Lambda x_\Lambda + \hat{h}_\Lambda \cdot x_\Lambda\}\}$ based on statistics $t^\Lambda(x_\Lambda) = (x_\Lambda, x_\Lambda x'_\Lambda)$. This has important implications both for modeling and inference in GMRFs.

2.2 Information Theory and Modeling

Among the key concepts of information theory are the related concepts of entropy, relative entropy (Kullback-Leibler divergence) and mutual information. A rigorous, systematic treatment of these information measures was provided by Shannon [121] in his development of an information-based theory of communication. The notion of entropy had previously played a fundamental role in thermodynamics and statistical mechanics as developed by physicists such as Boltzmann and Gibbs [20, 62]. This is closely related to recent applications of information theory for statistical modeling and approximate inference. We will review selected aspects of information theory and statistical modeling from this perspective.

2.2.1 Maximum-Entropy Principle

A key principle of statistical modeling introduced by physicists and generalized by statisticians is the *maximum-entropy principle* (Good [64], Jaynes [75], Cover and Thomas [31]). Given a probability model p for the outcome x of some random experiment, the entropy $h[p]$ is a real-valued measure of the randomness or uncertainty associated to that probability model.

Definition 5 (Entropy) *The* entropy *of probability distribution p on $\mathcal{X}$ is*

$$\begin{aligned} h[p] &= E_p\left\{\log \frac{1}{p(\mathrm{x})}\right\} \\ &= -\int_{\mathcal{X}} p(x)\log p(x)\,dx \end{aligned} \tag{2.55}$$

The maximum-entropy principle then simply asserts that among those probability models p consistent with whatever knowledge we have concerning the outcome x, we should adopt that probability model having maximum entropy $h[p]$. Intuitively, we should assume as little as possible about the process while still capturing what is known about the process. For instance, in the case that x is restricted to assuming only a finite number of values $x_1, x_2, \ldots, x_N$ (with no other prior knowledge given) the maximum-entropy principle requires that our model $p(x)$ assigns equal probability to each of these possible outcomes. This is the form that the maximum-entropy principle assumes in Boltzmann's formulation of statistical mechanics. Gibbs' formulation is related to the following result relating maximum-entropy subject to moment constraints to exponential families.

Proposition 7 (Maximum-Entropy Principle, Good [64]) *Let $\mathcal{P}$ denote the set of probability distributions on $\mathcal{X}$ which satisfy expectation constraints $\int_{\mathcal{X}} p(x)t(x)dx = \eta$ for $\eta \in \mathcal{R}^d$ and $t : \mathcal{X} \to \mathcal{R}^d$. Suppose that $\mathcal{P}$ contains a probability distribution p^* of the form*

$$p^*(x) = \exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{2.56}$$

where $\theta \in \mathcal{R}^d$ and $\varphi(\theta) = \log \int_{\mathcal{X}} \exp\{\theta \cdot t(x)\}dx < \infty$. Then, for all $p \in \mathcal{P}$,

$$h[p] \leq h[p^*] = \varphi(\theta) - \theta \cdot \eta \tag{2.57}$$

where equality occurs if and only if $p = p^*$. Hence, p^* is the maximum-entropy distribution in $\mathcal{P}$.

Proof. We illustrate the main idea by the method of Lagrange multipliers, we write the Lagrangian as a function of p as

$$\begin{aligned} L[p] &= h[p] + \theta_0 \int p(x)dx + \sum_k \theta_k \int p(x)t_k(x)dx \\ &= \int \{-p(x)\log p(x) + \theta_0 p(x) + \sum_k \theta_k p(x)t_k(x)\}\, dx \end{aligned} \tag{2.58}$$

where the θ_k parameters are Lagrange multipliers associated to moment constraints and θ_0 is the Lagrange multiplier associated with the normalization constraint $\int p(x)dx = 1$. The perturbation δL due to an infinitesimal variation δp is given by differentiating the integrand of L with respect to $p(x)$ and integrating with respect to the variation $\delta p = \delta p(x)dx$.

$$\delta L = \int \{-(1 + \log p(x)) + \theta_0 + \sum_k \theta_k t_k(x)\}\, \delta p \tag{2.59}$$

Stationarity of $L[p]$ with respect to arbitrary variations δp is forced by setting the integrand of δL to zero. Solving for p then gives the stationary distribution

$$p^*(x) = \exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{2.60}$$

where we have set $\varphi(\theta) = \theta_0 - 1$. The normalization condition gives $\varphi(\theta) = \log \int \exp\{\theta \cdot t(x)\}dx$. The remaining multipliers θ are determined by the moment constraints. Finally, evaluation of the objective function $h[p]$ for this stationary distribution gives $h[p] = -E_p\{\theta \cdot t(\mathrm{x}) - \varphi(\theta)\} = -\{\theta \cdot \eta - \varphi(\theta)\}$. That this stationary distribution is in fact the global maximizer follows from the strict concavity of the entropy $h[p]$ in the probability distribution $p(\cdot)$[5] and linearity of the constraints. □

This shows that an exponential model $p(x) \propto \exp\{\theta \cdot t(x)\}$ (based on $b(x) = 1$) with moment parameters η is the maximum-entropy probability distribution satisfying moment constraints $E_p\{t(\mathrm{x})\} = \eta$ and that the exponential parameters θ may be interpreted as Lagrange multipliers associated to those moment constraints. Also, the convex-conjugate of the cumulant function $\varphi^*(\eta) = \eta \cdot \theta - \varphi(\theta)$ is the negative

[5] See Cover and Thomas [31] for a simple concavity proof.

entropy $\varphi^*(\eta) = -h[p]$.

Kullback [84] suggested a generalized principle for model selection based on the quantity introduced by Kullback and Leibler [85] as the *mean information for discrimination*[6] between two probability distributions p and q defined below.

Definition 6 (Kullback-Leibler Divergence) *Let p and q be probability distributions on $\mathcal{X}$. The Kullback-Leibler (KL) divergence of probability distribution p relative to q is*

$$\begin{aligned} D(p\|q) &= E_p\left\{\log\frac{p(\mathrm{x})}{q(\mathrm{x})}\right\} \\ &= \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}\,dx \end{aligned} \tag{2.61}$$

It can be shown by Jensen's inequality that $D(p\|q) \geq 0$ with equality if and only if $p(x) = q(x)$ for essentially all x [31]. Hence, KL-divergence is considered as a measure of contrast between the two distributions. From the hypothesis-testing perspective emphasized by Kullback [84], $\log p(x)/q(x)$ measures the information provided by a specific observation x favoring the hypothesis '$\mathrm{x} \sim p$' over the hypothesis '$\mathrm{x} \sim q$'. The mean information for discrimination is then just the expected information gained in favor of '$\mathrm{x} \sim p$' over '$\mathrm{x} \sim q$' per observation when in fact $\mathrm{x} \sim p$. This provides the basis for the fundamental role KL-divergence plays as an error exponent in the asymptotic equipartition theorem (AEP), Shannon's communication theory [121], the method of types, and large-deviation theory [41] (e.g. Sanov's theorem). The text by Cover and Thomas [31] introduces this material and provides historical notes and references.

Kullback's principle, which may be considered as a *minimum-discrimination principle*, asserts that given a reference model $q(x)$ and some additional knowledge of the random outcome x, we should then adopt as our refined model that $p(x)$ which is consistent with our new knowledge of x but otherwise minimizes the KL-divergence $D(p\|q)$. We write this minimum principle as

$$p^* = \arg\min_{p\in\mathcal{P}} D(p\|q) \tag{2.62}$$

where $\mathcal{P}$ is the set of all probability distributions consistent with our knowledge of x. Intuitively, it should be made as difficult as possible to discriminate q from p given sample paths drawn from p. This then provides a perspective for the family of exponential models based on a normalized probability distribution q.

Proposition 8 (Minimum-Discrimination Theorem, Kullback [84]) *Let $\mathcal{P}$ denote the set of probability distributions on $\mathcal{X}$ which satisfy expectation constraints*

[6] But previously considered from a different perspective by Jeffreys. Jeffreys' goal was to construct an invariant measure of entropy. Several researchers consider KL as a generalization of the notion of entropy. Hence, KL is sometimes referred to as *relative entropy.*

$E_p\{t(\mathrm{x})\} = \eta$ for $t : \mathcal{X} \to \mathcal{R}^d$ and $\eta \in \mathcal{R}^d$. Suppose that $\mathcal{P}$ contains a probability distribution p^* given by

$$p^*(x) = q(x) \exp\{\theta \cdot t(x) - \varphi(\theta)\} \tag{2.63}$$

where q is a positive (non-vanishing) probability distribution on $\mathcal{X}$, $\theta \in \mathcal{R}^d$ and $\varphi(\theta) = \log \int_{\mathcal{X}} q(x) \exp\{\theta \cdot t(x)\} dx < \infty$. Then, for all $p \in \mathcal{P}$,

$$D(p\|q) \geq D(p^*\|q) = \varphi(\theta) - \theta \cdot \eta \tag{2.64}$$

where equality occurs if and only if $p = p^*$. Hence, p^* is the minimum-discrimination distribution in $\mathcal{P}$ relative to q.

Proof. This is a generalization of the maximum-entropy principle. The same variational argument as before applies but with an additional $-p(x) \log q(x)$ term so that solving for the stationary distribution of $D(p\|q)$, subject to constraints, gives $p(x) \propto q(x) \exp\{\theta \cdot t(x)\}$. This is the global minimum of $D(p\|q)$ due to the strict convexity of KL in the distribution $p(\cdot)$ (see Cover and Thomas [31] for the convexity proof). □

This shows that an exponential model $p(x) = q(x) \exp\{\theta \cdot t(x) - \varphi(\theta)\}$ (based on a normalized distribution $b(x) = q(x)$) with moment parameters η is the minimum-discrimination probability distribution relative to q satisfying the moment constraints $E_p\{t(\mathrm{x})\} = \eta$ and that the exponential parameters θ are Lagrange multipliers associated to those moment constraints. Also, the convex-conjugate of the cumulant function $\varphi^*(\eta) = \theta \cdot \eta - \varphi(\theta)$ gives the KL-divergence relative to q, i.e. $\varphi^*(\eta) = D(p\|q)$.

Note that when $q(x)$ is also an exponential model of the form $q(x) \propto \exp\{\theta \cdot t(x)\}$ then so is $p(x)$. In this case, the model $p(x)$ above is also a maximum-entropy model but subject to a new set of moment constraints. In particular, consider the case where the original model $q(x)$ is the maximum-entropy model associated with a set of moment constraints $Et_1(x) = \eta_1$ and we then impose an augmented set of constraints (containing these original constraints as a subset) $Et_2(x) = \eta_2$. In this case, the entropy is decreased by $h[q] - h[p] = D(p\|q)$. If we regard the negative entropy as the information content of the model, then we see that this information is very naturally increased by imposing additional moment constraints and the amount of this increase is given by the KL-divergence of the updated model relative to the prior model. This shows the connection between minimum-discrimination and maximum-entropy. Often, as in Della Pietra et al [106], minimum-discrimination is also referred to as a “maximum-entropy” method.

2.2.2 Duality with Maximum-Likelihood

Fisher [53, 54] introduced the *maximum-likelihood principle* for model selection among a parameterized family of models from data. For exponential families, it is well known that maximum-likelihood is dual to minimum-discrimination/maximum-entropy as we discuss below (Della Pietra et al [106], Jordan [77]).

Maximum-Likelihood Principle. The typical situation is as follows. Let x denote the outcome of some random experiment with unknown probability distribution $p(x)$. We observe this distribution through a collection of independent samples $x^N = (x_1, \ldots, x_N)$ each with distribution p. We then wish to select a model to best approximate $p(x)$ among a parameterized family of candidates $\mathcal{Q} = (q(x;\theta), \theta \in \Theta)$. The maximum-likelihood principle advises that we select the model which maximizes the likelihood of the data $q(x^N;\theta) = \prod_i q(x_i;\theta)$ (equivalently, the log-likelihood $\log q(x^N;\theta) = \sum_i \log q(x_i;\theta)$). This gives the *maximum-likelihood estimate* $\hat{\theta}_{ML}$ of the parameters θ.

$$\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \left\{ \sum_{i=1}^{N} \log q(x_i;\theta) \right\} \tag{2.65}$$

This also may be phrased in terms of information theory. Define the *empirical distribution* $\tilde{p}$ of data x^N as a scaled sum of Dirac δ-functions placed at the observed samples.

$$\tilde{p}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i) \tag{2.66}$$

This is defined such that expectations with respect to the empirical distribution gives sample averages.

$$E_{\tilde{p}}\{f(x)\} = \frac{1}{N} \sum_{i=1}^{N} f(x_i) \tag{2.67}$$

Essentially, $q(x^N;\theta) \propto \exp\{-ND(\tilde{p}\|q)\}$ such that the maximum-likelihood principle may be viewed as minimization of the KL-divergence $D(\tilde{p}\|q)$ over $\mathcal{Q}$. Denoting $\hat{q}_{ML} = q(\cdot;\hat{\theta}_{ML})$ we write

$$\hat{q}_{ML} = \arg\min_{q\in\mathcal{Q}} D(\tilde{p}\|q) \tag{2.68}$$

Note that the sense of KL-divergence is reversed in comparison to the minimum-discrimination principle (2.62). Also, the minimum-discrimination principle imposes expectation constraints while maximum-likelihood is over a parameterized family. However, in exponential families, maximum-likelihood and minimum-discrimination are dual problems. We state this in the following general form.

Proposition 9 (Minimum-Discrimination/Maximum-Likelihood Duality) *Let $\mathcal{Q}$ denote the exponential family based on $(\mathcal{X}, t(\cdot), b(\cdot))$. Let $\mathcal{P}$ denote the family of all probability distributions (not necessarily exponential) satisfying moment constraints $E_p\{t(\mathrm{x})\} = \eta$. Suppose that there exists a probability distribution $r = \mathcal{P} \cap \mathcal{Q}$. Then, r is both the minimum-discrimination distribution in $\mathcal{P}$ for any $q \in Q$,*

$$r = \arg\min_{p\in\mathcal{P}} D(p\|q), \tag{2.69}$$

and the maximum-likelihood distribution in $\mathcal{Q}$ for any $p \in \mathcal{P}$,

$$r = \arg\min_{q \in \mathcal{Q}} D(p\|q). \tag{2.70}$$

Moreover, either condition uniquely determines r such that $\mathcal{P} \cap \mathcal{Q} = \{r\}$.

Proof. By hypothesis, $r, q \in \mathcal{Q}$ such that $r(x)/q(x) \propto \exp\{\beta \cdot t(x)\}$ where $\beta = \theta(r) - \theta(q)$. Hence, by Kullback's minimum discrimination theorem, $r(x) \propto q(x)\exp\{\beta \cdot t(x)\}$ is the minimum-discrimination distribution relative to q subject to moment constraints $E\{t(\mathrm{x})\} = \eta$ which proves the first part of the proposition. Now, by hypothesis $r \in \mathcal{Q}$ so we may write $r(x) = b(x)\exp\{\theta(r) \cdot t(x) - \varphi(\theta(r))\}$ where $\varphi(\theta) = \int b(x)\exp\{\theta \cdot t(x)\}dx$. This minimizes $D(p\|q) = -h[p] - E_p\{\log q(x)\}$, a strictly convex function of $q(\cdot)$, over $q \in \mathcal{Q}$ if and only if the gradient in $\theta(q)$ vanishes at $\theta(r)$. A simple calculation shows that this occurs if and only if r and p give the same moments, i.e. $E_r\{t(\mathrm{x})\} = E_p\{t(\mathrm{x})\}$. By hypothesis $r, p \in \mathcal{P}$ so that this condition is satisfied and r is the maximum-likelihood distribution in $\mathcal{Q}$ for p. □

2.2.3 Information Geometry and Projection

We now consider the so-called *information geometry* of parameterized families of probability distributions with respect to the Fisher information metric. This framework provides a unifying perspective combining elements of information theory, differential geometry, convex duality and optimization theory. Many researchers have contributed to the development of this picture including Fisher [53], Rao [109], Chentsov [29, 28], Csiszár [34], Efron [46, 47], Barndorff-Nielsen [7], and Amari [3, 4]. The monograph of Amari and Nagaoka [5] is recommended for a systematic overview of the field and also for historical notes and references. Exponential families have played a central role in this development and will be our main concern. However, the dual differential-geometric framework of Amari applies more generally. We shall see that the duality between maximum-likelihood and minimum-discrimination carries over to information geometry and is then endowed with a very intuitive geometric interpretation.

General Statistical Manifolds. Consider a parameterized family of probability distributions $\mathcal{F} = \{p_\xi(x) | \xi \in \Xi\}$ on $\mathcal{X}$ with parameter space $\Xi \subset R^d$. That is, each $\xi \in \Xi$ specifies a probability distribution $p_\xi(x)$ on $\mathcal{X}$. Suppose that the following conditions hold:

1. For each $\xi \in \Xi$, the probability distribution p_ξ is positive (non-vanishing) on $\mathcal{X}$.

2. The mapping $\xi \to p_\xi$ is sufficiently smooth such that $p(\cdot; \xi)$ has partial derivatives of all orders (including mixed derivatives) with respect to ξ at every point in Ξ.

3. The parameterization is non-degenerate such that distinct parameters specify distinct probability distributions.

4. The parameter set Ξ is a simply-connected, open subset of $\mathcal{R}^d$.

We then say that $\mathcal{F}$ is a *statistical manifold* of dimension d with coordinates Ξ. We let $\xi(p)$ denote the (unique) coordinates of $p \in \mathcal{F}$. We also say that $\mathcal{G} \subset \mathcal{F}$ is a *submanifold* of $\mathcal{F}$ if $\mathcal{G}$ is a statistical manifold smoothly embedded in $\mathcal{F}$. Submanifolds are typically specified in one of two ways:

1. *Parametric Submanifold.* The submanifold $\mathcal{G}$ is explicitly specified by an injective map $\sigma : T \to \Xi$ where $T \subset \mathcal{R}^s$ is a simply-connected, open subset of $\mathcal{R}^s$ $(s < d)$.
$$\mathcal{G} = \{p \in \mathcal{F} | \exists t \in T : \sigma(t) = \xi(p)\} \tag{2.71}$$
For σ sufficiently smooth, this defines a submanifold of dimension s with coordinate system T.

2. *Implicit Submanifold.* The submanifold $\mathcal{G}$ is specified as the subset of $\mathcal{F}$ satisfying constraints $\rho(\xi) = 0$ where $\rho = (\rho_i : \Xi \to \mathcal{R}, i = 1, \ldots, m)$ is a collection of $m < d$ sufficiently smooth functions with linearly independent gradient vectors $\{\nabla \rho_i\}$ at each point in Ξ.
$$\mathcal{G} = \{p \in \mathcal{F} | \rho(\xi(p)) = 0\} \tag{2.72}$$
Provided the constraints are consistent, such that $\rho(\xi) = 0$ for some $\xi \in \Xi$, this defines a submanifold of dimension $s = d - m$. In this case, a coordinate system for the submanifold need not be explicitly specified.

We let $\xi(\mathcal{G}) \equiv \{\xi(p) | p \in \mathcal{G}\}$ denote the image of submanifold $\mathcal{G}$ in ξ-coordinates.

We now consider how to construct an invariant Riemannian metric on the statistical manifold $\mathcal{F}$ based on the Fisher information matrix of the parameterized family. In this construction, we consider two complementary viewpoints, where $\mathcal{F}$ is represented either as a set of probability distributions or as a set of log-probability distributions, and then construct a Riemannian metric as the inner product between tangents of these respective representations. This metric is related to KL-divergence and Fisher information and is invariant for sufficiently smooth reparameterization of the family.

M-Representation. First, consider representation of $\mathcal{F}$ as a set of probability distributions $\mathcal{F}^{(m)} = \{\mathbf{p}_\xi \equiv (p(x;\xi), x \in \mathcal{X}) | \xi \in \Xi\} \subset \mathcal{R}^{\mathcal{X}}$ where $\mathbf{p}_\xi$ denotes a point in function space $\mathcal{R}^{\mathcal{X}} = \{f : \mathcal{X} \to \mathcal{R}\}$. We call this the *m-representation* of $\mathcal{F}$.[7]

At each point $\mathbf{p}_\xi \in \mathbf{F}$, we define an *m-tangent* for each parameter differential $\Delta \in \mathcal{R}^d$ by
$$\mathbf{t}_\xi(\Delta) = \lim_{\lambda \downarrow 0} \left\{ \frac{\mathbf{p}_{\xi + \lambda\Delta} - \mathbf{p}_\xi}{\lambda} \right\} \tag{2.73}$$
This defines the *m-tangent space*
$$\mathbf{T}_\xi(\mathcal{F}) = \{\mathbf{t}_\xi(\Delta) | \Delta \in \mathcal{R}^d\} \tag{2.74}$$

[7] This terminology is adopted because *mixture families*, given by $\mathcal{F}^{(m)} = \{p(x;\xi) = (1 - \sum_i \xi_i) p_0(x) + \sum_i \xi_i p_i(x) | \xi_i > 0, \sum_i \xi_i < 1\}$ where $\{p_i\}$ are non-vanishing probability distributions on $\mathcal{X}$, are flat submanifolds of $\mathcal{R}^{\mathcal{X}}$ (Amari and Nagaoka, [5]).

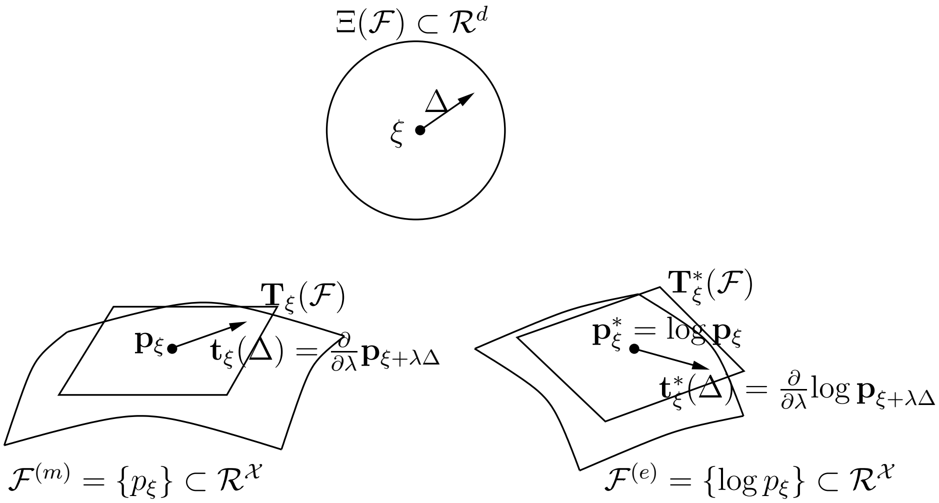

$\Xi(\mathcal{F}) \subset \mathcal{R}^d$
Δ
ξ
$\mathbf{T}_\xi(\mathcal{F})$
$\mathbf{p}_\xi$
$\mathbf{t}_\xi(\Delta) = \frac{\partial}{\partial\lambda}\mathbf{p}_{\xi+\lambda\Delta}$
$\mathcal{F}^{(m)} = \{p_\xi\} \subset \mathcal{R}^{\mathcal{X}}$
$\mathbf{T}^*_\xi(\mathcal{F})$
$\mathbf{p}^*_\xi = \log \mathbf{p}_\xi$
$\mathbf{t}^*_\xi(\Delta) = \frac{\partial}{\partial\lambda}\log \mathbf{p}_{\xi+\lambda\Delta}$
$\mathcal{F}^{(e)} = \{\log p_\xi\} \subset \mathcal{R}^{\mathcal{X}}$

Figure 2-8: Illustration depicting relationship between parameter differentials $\Delta \in \mathcal{R}^d$ (top center), m-tangents $\mathbf{t}_\xi(\Delta) \in \mathbf{T}_\xi(\mathcal{F})$ (bottom left), and e-tangents $\mathbf{t}^*_\xi(\Delta) \in \mathbf{T}^*_\xi(\mathcal{F})$ (bottom right).

at each point $\mathbf{p}_\xi \in \mathcal{F}^{(m)}$. For each $i = 1, \ldots, d$ we define a *basis vector* by

$$\mathbf{t}_{\xi,i} = \partial_i \mathbf{p}_\xi \tag{2.75}$$

where $\partial_i \equiv \frac{\partial}{\partial \xi_i}$ denotes partial differentiation with respect to the i-th parameter. We may express an arbitrary m-tangent $\mathbf{t}_\xi(\Delta)$ as a linear combination of these d basis vectors.

$$\mathbf{t}_\xi(\Delta) = \sum_i \Delta_i \mathbf{t}_{\xi,i} \tag{2.76}$$

This shows that, for each $\xi \in \Xi$, the m-tangent space $\mathbf{T}_\xi(\mathcal{F})$ is a d-dimensional vector space spanned by the basis vectors $(\mathbf{t}_{\xi,i}, i = 1, \cdots, d)$.

E-Representation. In an analogous manner, we may also consider representation of $\mathcal{F}$ as a set of log-probability distributions $\mathcal{F}^{(e)} = \{\mathbf{p}^*_\xi \equiv (\log p(x,\xi), x \in \mathcal{X}), \xi \in \Xi\}$. We call this the *e-representation* of $\mathcal{F}$.[8]

At each point $\mathbf{p}^*_\xi \in \mathcal{F}^{(e)}$, we define an *e-tangent* for each coordinate differential $\Delta \in \mathcal{R}^d$ by

$$\mathbf{t}^*_\xi(\Delta) = \lim_{\lambda \downarrow 0} \left\{ \frac{\mathbf{p}^*_{\xi+\lambda\Delta} - \mathbf{p}^*_\xi}{\lambda} \right\} \tag{2.77}$$

This defines the *e-tangent space*

$$\mathbf{T}^*_\xi(\mathcal{F}) = \{\mathbf{t}^*_\xi(\Delta) | \Delta \in \mathcal{R}^d\} \tag{2.78}$$

[8] This terminology is adopted because *exponential families*, in a *denormalized* representation $\tilde{\mathcal{F}}^{(e)} = \{\log \lambda p(x;\xi) = \sum_i \xi_i t_i(x) + \log \lambda | \xi \in \Xi, \lambda > 0\}$, are flat submanifolds of $\mathcal{R}^{\mathcal{X}}$ (Amari and Nagaoka, [5]).

at each point $\mathbf{p}^*_\xi \in \mathcal{F}^{(e)}$. We again define a collection of basis vectors by

$$\mathbf{t}^*_{\xi,i} = \partial_i \mathbf{p}^*_\xi \tag{2.79}$$

for $i = 1, \ldots, d$. Arbitrary e-tangents may then be expressed as liner combinations of these basis vectors.

$$\mathbf{t}^*_\xi(\Delta) = \sum_i \Delta_i \mathbf{t}^*_{\xi,i} \tag{2.80}$$

Hence, $\mathbf{T}^*_\xi(\mathcal{F})$ is a d-dimensional vector space spanned by the basis $(\mathbf{t}^*_{\xi,i}, i = 1, \ldots, d)$.

We consider these two complementary representations so as to illuminate the geometric interpretation of the Fisher information metric (Amari and Nagaoka, [5]) which may be constructed as follows.

Fisher Information Metric. Given any two points in function space $\mathbf{f}, \mathbf{g} \in \mathcal{R}^{\mathcal{X}}$, let $\mathbf{f} \cdot \mathbf{g}$ denote the usual inner-product of two functions:

$$\mathbf{f} \cdot \mathbf{g} = \int_{\mathcal{X}} f(x) g(x) dx \tag{2.81}$$

based on this inner-product of functions and the preceding definitions of m-tangent and e-tangent, let the *Fisher information metric* be defined by the bilinear form:

$$\langle \Delta_1, \Delta_2 \rangle_\xi = \mathbf{t}^*_\xi(\Delta_1) \cdot \mathbf{t}_\xi(\Delta_2) \tag{2.82}$$

for all $\xi \in \Xi$ and $\Delta_1, \Delta_2 \in \mathcal{R}^d$. This is the inner-product (in function space) of the e-tangent $\mathbf{t}^*_\xi(\Delta_1)$ with the m-tangent $\mathbf{t}_\xi(\Delta_2)$. This metric is closely related to KL-divergence. For all $\xi \in \Xi$ and $\Delta \in \mathcal{R}^d$, it holds that:

$$\frac{\partial D(p_\xi \| p_{\xi+\lambda\Delta})}{\partial \lambda} = \frac{1}{2} \langle \Delta, \Delta \rangle_\xi \tag{2.83}$$

To show the connection with Fisher information, we write

$$\begin{aligned} \langle \Delta_1, \Delta_2 \rangle_\xi &= \left(\sum_i \Delta_{1,i} \mathbf{t}^*_{\xi,i} \right) \cdot \left(\sum_j \Delta_{2,j} \mathbf{t}_{\xi,j} \right) && (2.84) \\ &= \sum_{i,j} \Delta_{1,i} \Delta_{2,j} \mathbf{t}^*_{\xi,i} \cdot \mathbf{t}_{\xi,j} \\ &= \Delta_1' G(\xi) \Delta_2 && (2.85) \end{aligned}$$

where $G(\xi)$ is the $d \times d$ matrix with ij-th element $g_{i,j}(\xi) = \mathbf{t}^*_{\xi,i} \cdot \mathbf{t}_{\xi,j}$. A simple calculation shows that

$$g_{i,j}(\xi) = E_\xi\{\partial_i \log p(\mathrm{x}; \xi) \partial_j \log p(\mathrm{x}; \xi)\} \tag{2.86}$$

This $G(\xi)$ is the *Fisher information* of the parameterized family $p(x; \xi)$. For each $\xi \in \Xi$, this is the covariance matrix of the random vector $\mathrm{v} = (\partial_i \log p(\mathrm{x}; \xi), i = 1, \ldots, d)$ and, hence, is a symmetric, positive-definite matrix. This also shows that, for each

$\xi \in \Xi$, $\langle \cdot, \cdot \rangle_\xi$ defines an *inner-product*[9] on $\mathcal{R}^d$. This defines a *Riemannian metric* on Ξ. Moreover, this Fisher information metric is *invariant* with respect to smooth reparameterization of the statistical manifold.[10]

Based on the Fisher information metric, we say that two parameter differentials $\Delta_1, \Delta_2 \in \mathcal{R}^d$ are *$\mathcal{I}$-orthogonal* at $\xi \in \Xi$ if

$$\langle \Delta_1, \Delta_2 \rangle_\xi \equiv \mathbf{t}^*_\xi(\Delta_1) \cdot \mathbf{t}_\xi(\Delta_2) = 0 \tag{2.87}$$

That is, if the e-tangent and m-tangent of the respective parameter differentials are orthogonal in function space. Accordingly, we say that two submanifolds $\mathcal{G}_1, \mathcal{G}_2 \subset \mathcal{F}$ are *$\mathcal{I}$-orthogonal* at $p_\xi \in \mathcal{G}_1 \cap \mathcal{G}_2$ when $\langle \Delta_1, \Delta_2 \rangle_\xi = 0$ for all Δ_1 and Δ_2 respectively tangent to $\xi(\mathcal{G}_1)$ and $\xi(\mathcal{G}_2)$ at the point ξ in parameter space Ξ.

Information Geometry of Exponential Families. With these ideas in mind, we now turn our attention back towards exponential families. The information geometry of exponential families enjoys some special properties arising due to the Legendre duality between the exponential and moment parameterizations of this family (discussed previously in Section 2.1.4).

Recall that an exponential family $\mathcal{F} = \{p_\theta(x) = b(x)\exp\{\theta \cdot t(x) - \varphi(\theta)\}\}$ is dually parameterized by moment coordinates $\eta = E_\theta\{t(\mathrm{x})\}$. Exponential and moment coordinates are related by the bijective coordinate transform:

$$\eta = \nabla\varphi(\theta) \;\Leftrightarrow\; \theta = \nabla\varphi^*(\eta) \tag{2.88}$$

where $\varphi(\theta)$ is the cumulant function, $\varphi^*(\eta)$ is the negative entropy function, and these are convex conjugate functions satisfying the Legendre relation

$$\varphi(\theta) + \varphi^*(\eta) = \theta \cdot \eta \tag{2.89}$$

for any dually-coupled pair (θ, η) satisfying (2.88). The parameter space of $\mathcal{F}$ in exponential and moment coordinates is respectively given by the effective domains of these cumulant and negative entropy functions; $\theta(\mathcal{F}) = \mathrm{dom}\ \varphi \equiv \{\theta | \varphi(\theta) < \infty\}$ and $\eta(\mathcal{F}) = \mathrm{dom}\ \varphi^* \equiv \{\eta | \varphi^*(\eta) < \infty\}$. Again, let $G(\theta)$ and $G^*(\eta)$ denote the Fisher information in these respective parameterizations of the family. These two symmetric positive definite matrices are also the Hessian curvature matrices respectively of the cumulant and entropy functions. Then, by virtue of (2.88), these are also the Jacobian matrices of coordinate transforms, $G(\theta) = \frac{\partial \eta}{\partial \theta}$ and $G^*(\eta) = \frac{\partial \theta}{\partial \eta}$, and are therefore

[9] That is, a symmetric, bilinear form having the property that $\langle \Delta, \Delta \rangle \geq 0$ for all $\Delta \in \mathcal{R}^d$ and $\langle \Delta, \Delta \rangle = 0$ if and only if $\Delta = 0$.

[10] To be precise, consider a second parameterization $\mathcal{F} = \{p(x; \xi^*) | \xi^* \in \Xi^*\}$ related to the the original ξ-parameterization by a diffeomorphism $\sigma : \Xi \to \Xi^*$, a sufficiently smooth, differentiable, bijective map. We may apparently define a second Fisher information metric in this reparameterized family by $\langle \Delta_1^*, \Delta_2^* \rangle^*_{\xi^*} \equiv (\Delta_1^*)' G^*(\xi^*) \Delta_2^*$ where $G^*(\xi^*)$ is the Fisher information of the ξ^* parameterization. Yet, for commensurate differentials $\Delta^* = \Sigma(\xi)\Delta$, related by the invertible Jacobian matrix $\Sigma(\xi) \equiv (\frac{\partial \sigma_i(\xi)}{\partial \xi_j})$, these metrics are actually equivalent. That is, $\langle \Delta_1, \Delta_2 \rangle_\xi = \langle \Delta_1^*, \Delta_2^* \rangle^*_{\sigma(\xi)}$ for all $\xi \in \Xi$ and $\Delta_1, \Delta_2 \in \mathcal{R}^d$.

inversely related $G^{-1}(\theta) = G^*(\eta)$ for any dually-coupled coordinate pair (θ, η).

These dual parameterizations give (equivalent) representations for the Fisher information metric:

$$\begin{aligned} \langle \Delta\theta_1, \Delta\theta_2 \rangle_\theta &= \Delta\theta_1' G(\theta) \Delta\theta_2 && (2.90) \\ \langle \Delta\eta_1, \Delta\eta_2 \rangle_\eta &= \Delta\eta_1' G^*(\eta) \Delta\eta_2 && (2.91) \end{aligned}$$

That is, for commensurate differentials $\Delta\eta_1 = G(\theta)\Delta\theta_1$ and $\Delta\eta_2 = G(\theta)\Delta\theta_2$, these are equal satisfying

$$\langle \Delta\theta_1, \Delta\theta_2 \rangle_\theta = \langle \Delta\eta_1, \Delta\eta_2 \rangle_\eta = \Delta\eta_1 \cdot \Delta\theta_2 \tag{2.92}$$

Compare this form of the Fisher information metric, as an inner product between differentials in dual parameter spaces, to our original definition (2.82), as an inner product (in function space) between tangents of the e- and m-representations of a statistical manifold. It appears that the dual exponential/moment parameterizations closely parallels the e- and m-representations of the family in function space.

To explore this connection a bit further, consider the e-tangent/m-tangent basis vectors in both exponential and moment coordinates:

$$\begin{aligned} \mathbf{t}_{\theta,i} = \frac{\partial}{\partial\theta_i}\mathbf{p}_\theta \quad &\Leftrightarrow \quad \mathbf{t}^*_{\theta,i} = \frac{\partial}{\partial\theta_i}\log \mathbf{p}_\theta && (2.93) \\ \mathbf{t}_{\eta,i} = \frac{\partial}{\partial\eta_i}\mathbf{p}_\eta \quad &\Leftrightarrow \quad \mathbf{t}^*_{\eta,i} = \frac{\partial}{\partial\eta_i}\log \mathbf{p}_\eta && (2.94) \end{aligned}$$

Each parameterization provides distinct bases for both the e- and m-tangent spaces. For instance, the two bases $(\mathbf{t}_{\theta,i}, i = 1, \ldots, d)$ and $(\mathbf{t}_{\eta,i}, i = 1, \ldots, d)$ each span $\mathbf{T}_p(\mathcal{F})$. Recalling the role Fisher information plays as the Jacobian matrix of the coordinate transform, then the chain-rule for partial differentiation yields the following invertible transformation law between bases;

$$\begin{aligned} \mathbf{t}_{\theta,i} &= \sum_j g_{ij}(\theta)\mathbf{t}_{\eta,j} && (2.95) \\ \mathbf{t}_{\eta,i} &= \sum_j g^*_{ij}(\eta)\mathbf{t}_{\theta,j} && (2.96) \end{aligned}$$

The same transformation law holds if we replace m-tangents by e-tangents in these two expressions. What is especially interesting about these dual bases is the following *biorthogonality principle* (Chentsov [29], Efron [47], Amari and Nagaoka [5]):

$$\mathbf{t}^*_{\theta,i} \cdot \mathbf{t}_{\eta,j} = \delta_{i,j} \tag{2.97}$$

where $\delta_{i,j}$ is the *Kronecker delta*.[11] This fundamental result points the way towards a more global characterization of information geometry in exponential families which we explore for the remainder of the section.

[11] The Kronecker delta $\delta_{i,j}$ is one whenever $i = j$ and is zero otherwise.

E-Flat and M-Flat Submanifolds. In exponential families, we will be especially concerned with two special types of submanifolds. In order to define these precisely, we first introduce some basic terminology. For $G \subset \mathcal{R}^d$, we define the set

$$\text{aff } G = \{u + \lambda(v - u) | u, v \in G, \lambda \in \mathcal{R}\}. \tag{2.98}$$

We say that G is *affine* if $G = \text{aff } G$. We say that G is an *affine restriction* of $U \subset \mathcal{R}^d$ if $G = U \cap \text{aff } G$ (equivalently, if there exists affine H such that $H \cap U = G$). The set $H = \text{aff } G$ is the smallest affine superset of G.

We consider two types of flat submanifolds within an exponential family:

- *E-flat Submanifold.* We say that $\mathcal{G} \subset \mathcal{F}$ is an *e-flat submanifold* of $\mathcal{F}$ if $\theta(\mathcal{G})$ is an affine restriction of $\theta(\mathcal{F})$. Intuitively, the submanifold is flat in exponential coordinates. The dimension of an e-flat submanifold equals the dimension of the hyperplane aff $\theta(\mathcal{G})$. A one-dimensional e-flat submanifold, a line segment in exponential coordinates, is also called an *e-geodesic*.

- *M-flat Submanifold.* We say that $\mathcal{G} \subset \mathcal{F}$ is an *m-flat submanifold* of $\mathcal{F}$ if $\eta(\mathcal{G})$ is an affine restriction of $\eta(\mathcal{F})$. Intuitively, the submanifold is flat in moment coordinates. The dimension of an m-flat submanifold equals the dimension of the hyperplane aff $\eta(\mathcal{G})$. A one-dimensional m-flat submanifold, a line segment in moment coordinates, is also called an *m-geodesic*.

E-flat submanifolds (of exponential families) enjoy the special status that these are themselves exponential families (which, in general, does not hold for m-flat submanifolds). M-flat submanifolds correspond to imposing linear expectation constraints on the exponential family.

Henceforth, we adopt the convention of letting $\mathcal{H}$ denote an e-flat submanifold, and letting $\mathcal{H}'$ denote an m-flat submanifold. Given such a pair of respectively e-flat/m-flat submanifolds, let us say that these are *biorthogonal submanifolds* if the image of the e-flat submanifold (in exponential coordinates) is perpendicular to the image of the m-flat submanifold (in moment coordinates). That is, if it holds that

$$(\eta(p_1) - \eta(p_2))'(\theta(q_1) - \theta(q_2)) = 0 \tag{2.99}$$

for all $q_1, q_2 \in \mathcal{H}$ and $p_1, p_2 \in \mathcal{H}'$. Furthermore, let us say that a pair of biorthogonal submanifolds $\mathcal{H}, \mathcal{H}' \subset \mathcal{F}$ are *complementary* if the respective dimensions of these submanifolds, s and s', are nonzero and $s + s' = d$, the dimension of the family $\mathcal{F}$.

We now indicate the relation between this *global* definition of biorthogonality to our previous *local* definition of $\mathcal{I}$-orthogonality with respect to the Fisher information metric. Suppose that $\mathcal{H}$ and $\mathcal{H}'$ are respectively e-flat and m-flat and have a common probability distribution $r \in \mathcal{H} \cap \mathcal{H}'$. Then, $\mathcal{H}$ and $\mathcal{H}'$ are biorthogonal if an only if they are $\mathcal{I}$-orthogonal at r. This is a consequence of the biorthogonality of the exponential and moment coordinate systems (2.97). We shall see that biorthogonal submanifolds intersect at *at most* one point. That is, if $r \in \mathcal{H} \cap \mathcal{H}'$ then $\mathcal{H} \cap \mathcal{H}' = \{r\}$. Furthermore, complementary biorthogonal submanifolds intersect at *exactly* one point. That is, there exist $r \in \mathcal{H} \cap \mathcal{H}'$ and, moreover, it is unique, i.e. $\mathcal{H} \cap \mathcal{H}' = \{r\}$. This is analogous to the fact that, in Euclidean geometry, two complementary perpendicular hyperplanes must intersect at exactly one point. We will show these results in the course of the following development. We approach this by considering the structure of KL-divergence in exponential families.

Properties of KL-Divergence in Exponential Families. The information geometry of exponential families is endowed with global structure which we now elucidate. As a point of departure, we consider KL-divergence as the *canonical divergence* (Amari and Nagaoka, [5]) induced by the convex-conjugate pair of functions $(\varphi(\theta), \varphi^*(\eta))$ defined as

$$K(\eta, \theta) = \varphi^*(\eta) + \varphi(\theta) - \eta \cdot \theta \qquad (2.100)$$

for all $\eta \in \mathrm{dom}\ \varphi^*$ and $\theta \in \mathrm{dom}\ \varphi$. This is a *convex bifunction* (Rockafellar [114]) as it is both convex in η (with θ held fixed) and convex in θ (with η held fixed). For an exponential family with cumulant function $\varphi(\theta)$ (negative entropy function $\varphi^*(\eta)$) this is precisely the KL-divergence $D(p\|q) = E_p\{\log q(\mathrm{x})\}$ between distributions $p, q \in \mathcal{F}$ with $\eta(p) = \eta$ and $\theta(q) = \theta$, i.e. $D(p\|q) = K(\eta(p), \theta(q))$ for all $p, q \in \mathcal{F}$.

Two related expressions for the KL-divergence are obtained by using the Legendre transform to specify both arguments in the same coordinate system. Let (θ, η) and (θ^*, η^*) each be a dually-coupled coordinate pair. We rewrite $K(\eta^*, \theta)$, applying the Legendre transform $\varphi^*(\eta^*) = \eta^* \cdot \theta^* - \varphi(\theta^*)$, to obtain:

$$K(\eta^*, \theta) = B(\theta; \theta^*) \equiv \varphi(\theta) - \{\varphi(\theta^*) + \eta^* \cdot (\theta - \theta^*)\} \qquad (2.101)$$

Here, we may view η^* as a function of θ^*, i.e. $\eta^* = \nabla\varphi(\theta^*)$. The function $B(\theta; \theta^*)$ is the *Bregman distance* (Bregman [24]) based on the convex function $\varphi(\theta)$. This may be written as $B(\theta; \theta^*) = \varphi(\theta) - \bar{\varphi}(\theta; \theta^*)$ where $\bar{\varphi}(\theta; \theta^*) = \varphi(\theta^*) + \nabla\varphi(\theta^*) \cdot (\theta - \theta^*)$ is the linear underestimate of $\varphi(\theta)$ based on the supporting hyperplane tangent to epi φ at θ^*. The corresponding geometric interpretation of KL-divergence, as the difference between a convex function $\varphi(\theta)$ and a supporting hyperplane, is illustrated in Figure 2-9.

Since the cumulant function is an essentially smooth, strictly convex function[12] this Bregman distance function then has some very useful properties:

1. For each $\theta^* \in \mathrm{dom}\ \varphi$, the function $b^*(\theta) = B(\theta; \theta^*)$ is itself an essentially smooth, strictly convex function with effective domain dom $b^* =$ dom φ.

2. For all $\theta, \theta^* \in \mathrm{dom}\ \varphi$ we have $B(\theta; \theta^*) \geq 0$. Moreover, $B(\theta; \theta^*) = 0$ if and only if $\theta = \theta^*$.

3. For each $\theta^* \in \mathrm{dom}\ \varphi$, the level sets $L_\delta = \{\theta \in \mathrm{dom}\ \varphi | b^*(\theta) \leq \delta\}$ are compact (closed and bounded), convex subsets of $\mathcal{R}^d$.

[12] That is, $\varphi(\theta)$ is both smooth and strictly convex over it's effective domain dom $\varphi = \{\theta | \varphi(\theta) < \infty\}$ and is also *steep* such that $\varphi(\theta) \to \infty$ for any sequence $\{\theta_k\} \subset \mathrm{dom}\ \varphi$ which converges to a limit point not contained in dom φ.

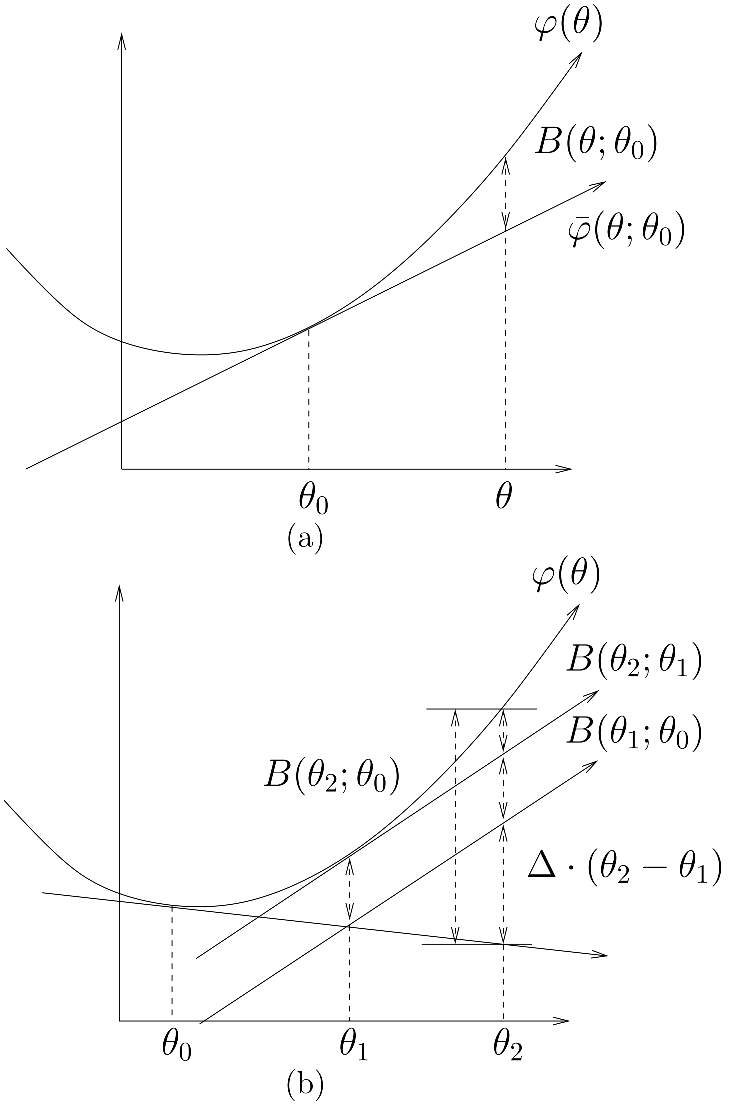


Figure 2-9: Illustration of Bregman distance and triangular relation. (a) Shows Bregman distance interpretation of KL-divergence in exponential coordinates, i.e. $D(p\|q)=B(\theta(q);\theta(p))$ where $B(\theta;\theta_0)=\varphi(\theta)-\bar{\varphi}(\theta;\theta_0)$ is the difference between the convex function $\varphi(\theta)$ and the linear underestimate $\bar{\varphi}(\theta;\theta_0)=\varphi(\theta_0)+\nabla\varphi(\theta_0)\cdot(\theta-\theta_0)$. (b) Shows geometric interpretation of the "triangular relation" satisfied by the Bregman distance, $B(\theta_2;\theta_0)=B(\theta_1;\theta_0)+B(\theta_2;\theta_1)+\Delta\cdot(\theta_2-\theta_1)$ where $\Delta=\nabla\varphi(\theta_1)-\nabla\varphi(\theta_0)=\eta_1-\eta_0$.

These results follow naturally from the geometric interpretation of the Bregman distance. In particular, the compactness of the level sets in (3) follows by showing that $B(\theta^{(k)}; \theta^*) \to \infty$ for any sequence $\{\theta^{(k)}\}$ in dom φ such that either: (i) $\theta^{(k)}$ converges to a limit point not contained in dom φ, or (ii) $\|\theta^{(k)}\| \to \infty$. Then, (i) and (ii) respectively show that the level sets are closed and bounded (and hence compact).

We may also derive an expression for the KL-divergence in terms of the moment coordinates of both arguments. Employing the Legendre transform $\varphi(\theta^*) = \theta^* \cdot \eta^* - \varphi^*(\eta^*)$, we obtain

$$K(\eta, \theta^*) = B^*(\eta; \eta^*) \equiv \varphi^*(\eta) - \{\varphi^*(\eta) + \theta^* \cdot (\eta - \eta^*)\} \tag{2.102}$$

This is the Bregman distance $B^*(\eta; \eta^*)$ based on the convex function $\varphi^*(\eta)$. This has a similar geometric interpretation as $B(\theta; \theta^*)$ and satisfies an analogous set of "dual" properties (obtained by reversing the sense of the KL-divergence and exchanging the roles played by exponential and moment coordinates).

We summarize this discussion as follows:

- *E-Balls.* We define the *e-ball* about $p \in \mathcal{F}$ of radius $\delta \geq 0$ as $\mathcal{L}_\delta(p) = \{q \in \mathcal{F} | D(p\|q) \leq \delta\}$. The image of an e-ball in *exponential coordinates* is a convex, compact subset of dom φ. Also, $\mathcal{L}_0(p) = \{p\}$ and $\mathcal{L}_\infty(p) \equiv \cup_{\delta>0}\mathcal{L}_\delta(p) =$ dom φ.

- *M-Balls.* We define the *m-ball* about $q \in \mathcal{F}$ of radius δ as $\mathcal{L}^*_\delta(q) = \{p \in \mathcal{F} | D(p\|q) \leq \delta\}$. The image of an m-ball in *moment coordinates* is a convex, compact subset of dom φ^*. Also, $\mathcal{L}^*_0(q) = \{q\}$ and $\mathcal{L}^*_\infty(q) \equiv \cup_{\delta>0}\mathcal{L}^*_\delta(q) =$ dom φ^*.

This tells us quite a bit about the topology of KL-divergence in exponential families. We will appeal to these properties to show the existence of certain information projections momentarily. But first, we present some very intuitive results for KL-divergence in exponential families which bear a close resemblance to analogous principles of Euclidean geometry.

Proposition 10 (Triangular Relation, Chentsov [29]) *Let $p, q, r \in \mathcal{F}$, an exponential family with dual coordinate systems (θ, η). Then,*

$$D(p\|q) = D(p\|r) + D(r\|q) + (\eta(p) - \eta(r)) \cdot (\theta(r) - \theta(q)) \tag{2.103}$$

Proof. This follows from the Bregman distance interpretation of KL. From (2.101), a simple calculation gives:

$$\begin{aligned} D(p\|q) - D(p\|r) - D(r\|q) &= \{\varphi(q) - \varphi(p) - \eta(p) \cdot (\theta(q) - \theta(p))\} \\ &\quad -\{\varphi(r) - \varphi(p) - \eta(p) \cdot (\theta(r) - \theta(p))\} \\ &\quad -\{\varphi(q) - \varphi(r) - \eta(r) \cdot (\theta(q) - \theta(r))\} \\ &= (\eta(p) - \eta(r)) \cdot (\theta(r) - \theta(q)) \end{aligned}$$

which proves the result. This has an intuitive geometric interpretation illustrated in Figure 2-9(b). □

This recalls the following triangle identity of Euclidean geometry for three points $p, q, r \in \mathcal{R}^n$,

$$\frac{1}{2}||p-q||^2 = \frac{1}{2}\|p-r\|^2 + \frac{1}{2}\|r-q\|^2 + (p-r)\cdot(r-q), \tag{2.104}$$

with KL-divergence playing an analogous role as half the squared distance.

For intersecting submanifolds $\mathcal{H}, \mathcal{H}' \subset \mathcal{F}$, which are respectively e-flat and m-flat, this gives the so-called "Pythagorean law" of information geometry characterizing $\mathcal{I}$-orthogonality of such e-flat/m-flat submanifolds.

Proposition 11 (Pythagorean Relation) *Let $\mathcal{H}$ and $\mathcal{H}'$ be respectively e-flat and m-flat submanifolds of $\mathcal{F}$ having a common probability distribution r. Then $\mathcal{H}$ and $\mathcal{H}'$ are $\mathcal{I}$-orthogonal at r (and, hence, biorthogonal) if and only if it holds that*

$$D(p\|q) = D(p\|r) + D(r\|q) \tag{2.105}$$

for all $p \in \mathcal{H}'$ and $q \in \mathcal{H}$. Moreover, if this condition holds, then $\mathcal{H} \cap \mathcal{H}' = \{r\}$.

Proof. Due to the biorthogonality of exponential and moment coordinates, the e-flat submanifold $\mathcal{H}$ and the m-flat submanifold $\mathcal{H}'$ are $\mathcal{I}$-orthogonal at r if and only if $(\eta(p)-\eta(r))\cdot(\theta(q)-\theta(r)) = 0$ for all $p \in \mathcal{H}'$ and $q \in \mathcal{H}$. Assuming $\mathcal{I}$-orthogonality, the Triangular relation then reduces to the Pythagorean relation. Conversely, if the Pythagorean relation holds for all $p \in \mathcal{H}'$ and $q \in \mathcal{H}$, then by the Triangular relation we must have $(\eta(p) - \eta(r)) \cdot (\theta(q) - \theta(r)) = 0$ for all $p \in \mathcal{H}'$ and $q \in \mathcal{H}$ so that $\mathcal{I}$-orthogonality also holds. To show uniqueness, suppose that $r_1, r_2 \in \mathcal{H} \cap \mathcal{H}'$. Then, by the result just shown, $D(r_1\|r_1) = D(r_1\|r_2) + D(r_2\|r_1) = 0$ such that $D(r_1\|r_2) = -D(r_2\|r_1)$. By positivity $D(p\|q) \geq 0$, such that $D(r_1\|r_2) = D(r_2\|r_1) = 0$ which occurs only when $r_1 = r_2$. Hence $\mathcal{H} \cap \mathcal{H}'$ cannot contain two distinct probability distributions and $\mathcal{H} \cap \mathcal{H}' = \{r\}$. □

This recalls the Pythagorean law for right triangles in Euclidean geometry. This viewpoint naturally leads to considering the following dual notions of "projection" in information geometry.

Information Projection. We now consider two related minimization problems of information geometry. Here, one seeks to minimize the KL-divergence $D(p\|q)$ with respect to q (alternatively p) over an e-flat (respectively m-flat) submanifold. Both of these "information projections" are shown to have unique solutions which may be characterized in terms of either $\mathcal{I}$-orthogonality or the Pythagorean relation. These projections are related to maximum-likelihood and minimum-discrimination and satisfy a corresponding duality principle.

Proposition 12 (M-Projection) *Let $\mathcal{H} \neq \varnothing$ be an* **e-flat** *submanifold of an exponential family $\mathcal{F}$ and let p be a given probability distribution in $\mathcal{F}$. Then, there exists a probability distribution $q^* \in \mathcal{H}$ satisfying the following (equivalent) conditions:*

(i) $D(p\|q^*) = \inf_{q\in\mathcal{H}} D(p\|q)$

(ii) $\forall q \in \mathcal{H}: \ (\eta(p) - \eta(q^*)) \cdot (\theta(q) - \theta(q^*)) = 0$

(iii) $\forall q \in \mathcal{H}: \ D(p\|q) = D(p\|q^*) + D(q^*\|q)$

Moreover, any one of these conditions uniquely determines q^*. We call this $q^* = \arg\min_{q \in \mathcal{H}} D(p\|q)$ the m-projection *of p to $\mathcal{H}$.*[13]

Proof. The crux of the proof lies in showing that the infimum in (i) is actually obtained by some $q^* \in \mathcal{H}$. The follows from the compactness of e-balls. By assumption, there exists some $q_0 \in \mathcal{H}$. Fix q_0 and set $\delta = D(p\|q_0)$. Define $\mathcal{D} \equiv \mathcal{L}_\delta(p) \cap \mathcal{H}$, the "disc" of all distributions q in the e-flat submanifold $\mathcal{H}$ with KL-divergence $D(p\|q) \le \delta$. In exponential coordinates, this disc is a compact subset of a hyperplane. Hence, the infimum of $D(p\|q)$ (a convex function of $\theta(q)$) over $\mathcal{D}$ is obtained by some point $q^* \in \mathcal{D}$, i.e. $D(p\|q^*) = \inf_{q \in \mathcal{D}} D(p\|q)$. Finally, since any point $q \in \mathcal{H} \setminus \mathcal{D}$ has $D(p\|q) > D(p\|q_0) \ge D(p\|q^*)$, this q^* actually achieves the infimum of $D(p\|q)$ over all of $\mathcal{H}$. Hence, there exists $q^* \in \mathcal{H}$ such that (i) holds as claimed. Moreover, since $D(p\|q)$ is strictly convex in $\theta(q)$, this q^* is unique.

Next, computing the gradient of $D(p\|q)$ with respect to $\theta(q)$ (with p held fixed) yields

$$\frac{\partial D(p\|q)}{\partial \theta(q)} = \eta(q) - \eta(p) \tag{2.106}$$

so that, in order for $q^* \in \mathcal{H}$ to be a stationary point of $D(p\|q)$ over $\mathcal{H}$, (ii) must hold, i.e. (i) ⇒ (ii). Conversely, since $D(p\|q)$ is strictly convex in $\theta(q)$, the stationarity condition (ii) actually shows q^* to be the (unique) global minimizer.

Finally, let $\mathcal{H}'$ be the m-geodesic connecting p and q^*. Assuming (ii), $\mathcal{H}$ and $\mathcal{H}'$ are shown to be $\mathcal{I}$-orthogonal at q^*. Hence, the Pythagorean relation (iii) holds, i.e. (ii) ⇒ (iii). Conversely, assuming (iii), Proposition 11 then asserts that $\mathcal{H}$ and $\mathcal{H}'$ are $\mathcal{I}$-orthogonal at r so that (ii) holds. □

M-projection may be considered as finding the maximum-likelihood probability distribution with respect to p over the exponential family determined by the e-flat submanifold $\mathcal{H}$.

Proposition 13 (E-Projection) *Let $\mathcal{H}' \neq \varnothing$ be an* **m-flat** *submanifold of an exponential family $\mathcal{F}$ and let q be a given probability distribution in $\mathcal{F}$. Then, there exists a probability distribution $p^* \in \mathcal{H}'$ satisfying the following (equivalent) conditions:*

(i) $D(p^*\|q) = \inf_{p \in \mathcal{H}'} D(p\|q)$

(ii) $\forall p \in \mathcal{H}': \ (\eta(p) - \eta(p^*)) \cdot (\theta(q) - \theta(p^*)) = 0$

(iii) $\forall p \in \mathcal{H}': \ D(p\|q) = D(p\|p^*) + D(p^*\|q)$

[13] Note that, although we are projecting to an *e-flat* submanifold, we follow Amari's convention of calling this *m-projection* because we may view the projection as following the *m-geodesic* containing p which is biorthogonal to the e-flat submanifold. This is called *reverse I-projection* by Csiszár.

Moreover, any one of these conditions uniquely determines p^*. We call this $p^* = \arg\min_{p\in\mathcal{H}'} D(p||q)$ the e-projection *of q to $\mathcal{H}'$.*[14]

Proof. The proof is a "dual" version of the argument given in Proposition 12 (just reverse the sense of KL-divergence and exchange the role played by exponential and moment coordinates). □

E-projection may be considered as finding the minimum-discrimination probability distribution with respect to q subject to linear moment constraints imposed by the m-flat submanifold $\mathcal{H}'$. That is, for some matrix $A \in \mathcal{R}^{m\times d}$ and vector $b \in \mathcal{R}^m$ $(m < s)$, where A is rank m and $s = d - m$ is the dimension of the m-flat submanifold, we may express the m-flat submanifold as $\mathcal{H}' = \{p \in \mathcal{F} | A\eta(p) - b = 0\}$. This may be seen as imposing expectation constraints $E_p\{\tilde{t}(\mathrm{x})\} = b$ with respect to statistics $\tilde{t}(x) = At(x)$. Then, by Kullback's minimum discrimination theorem (Proposition 8, Kullback [84]), the e-projection is given by

$$p^*(x) \propto q(x)\exp\{\lambda \cdot \tilde{t}(x)\} \tag{2.107}$$

where $\lambda \in \mathcal{R}^m$ is a vector of Lagrange multipliers chosen so as to satisfy the moment constraints $E_\lambda\{\tilde{t}(x)\} = b$. The e-projection is then specified in exponential coordinates by $\theta(p^*) = \theta(q) + A'\lambda$.[15]

The geometric interpretation of these two $\mathcal{I}$-projections is illustrated in Figure 2-10. The maximum-likelihood/minimum-discrimination duality between these information projections is shown by the following proposition:

Proposition 14 (Duality of $\mathcal{I}$-Projections) *Let $\mathcal{H}, \mathcal{H}' \subset \mathcal{F}$ be complementary biorthogonal submanifolds. Then, there exists $r \in \mathcal{H} \cap \mathcal{H}'$ and this is both the m-projection to $\mathcal{H}$ for any $p \in \mathcal{H}'$,*

$$r = \arg\min_{q\in\mathcal{H}} D(p||q) \tag{2.108}$$

and the e-projection to $\mathcal{H}'$ for any $q \in \mathcal{H}$,

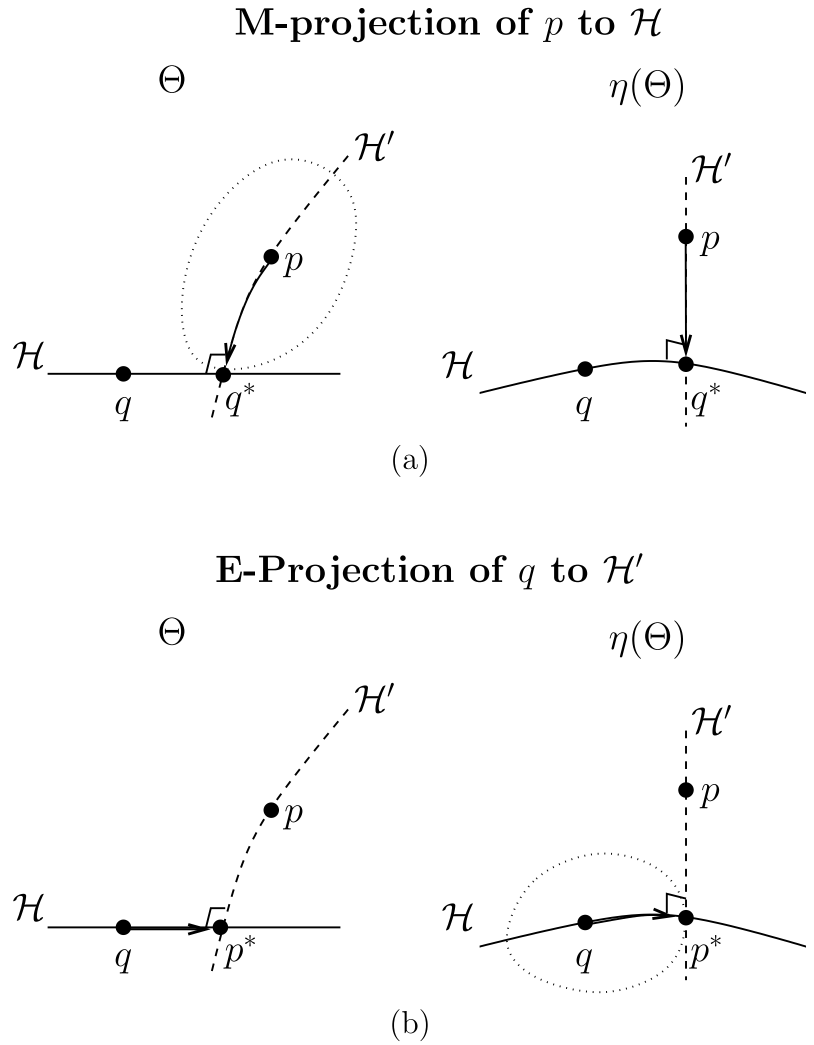
$$r = \arg\min_{p\in\mathcal{H}'} D(p||q) \tag{2.109}$$

Moreover, either condition uniquely determines r such that $\mathcal{H} \cap \mathcal{H}' = \{r\}$.

Proof. The complementary biorthogonal submanifolds $\mathcal{H}, \mathcal{H}'$ correspond to hyperplanes $E, M \subset \mathcal{R}^d$, given by $E = \text{aff } \theta(\mathcal{H})$ and $M = \text{aff } \eta(\mathcal{H}')$. These are orthogonal complements of one another such that any vector orthogonal to E is parallel to M. Pick an arbitrary point $p \in \mathcal{H}'$ and let r be the m-projection of p to

[14] Note that, although we are projecting to an *m-flat* submanifold, we follow Amari's convention of calling this *e-projection* because we may view the projection as following the *e-geodesic* containing p which is biorthogonal to the m-flat submanifold. This is called *I-projection* by Csiszár.

[15] This shows that the e-projection is located within an m-dimensional e-flat submanifold $\mathcal{H} = \{p \in \mathcal{F} | \exists\lambda \in \mathcal{R}^m : \theta(p) = \theta(q) + A'\lambda\}$. This is the complementary biorthogonal submanifold to $\mathcal{H}'$ containing p.



M-projection of p to $\mathcal{H}$

(a)

E-Projection of q to $\mathcal{H}'$

(b)

Figure 2-10: Illustration of dual $\mathcal{I}$-projections in both exponential coordinates θ (left) and moment coordinates $\eta = E_\theta\{t(\mathrm{x})\}$ (right): (a) m-projection of p to the e-flat submanifold $\mathcal{H}$, (b) e-projection of q to the m-flat submanifold $\mathcal{H}'$. Note that $\mathcal{H}$ is flat in exponential coordinates while $\mathcal{H}'$ is flat in moment coordinates. These are biorthogonal submanifolds since $\mathcal{H}$ drawn in exponential coordinates is perpendicular to $\mathcal{H}'$ drawn in moment coordinates. Given $(p, \mathcal{H})$, the m-projection q^* is determined by tracing the straight line in moment coordinates biorthogonal to $\mathcal{H}$. Given $(q, \mathcal{H}')$, the e-projection p^* is determined by tracing the straight line in exponential coordinates biorthogonal to $\mathcal{H}'$. These are dual problems and both give the same solution, i.e. $p^* = q^*$.

$\mathcal{H}$. By Proposition 12(ii), $\Delta\eta = \eta(r) - \eta(p)$ is orthogonal to E (parallel to M). Hence, $\eta(r) = \eta(p) + \Delta\eta \in M$. Since $\theta(\mathcal{H}') = M \cap \theta(\mathcal{F})$, this shows the existence of $r \in \mathcal{H} \cap \mathcal{H}'$.

Then, by Proposition 11, $\mathcal{H} \cap \mathcal{H}' = \{r\}$ and the Pythagorean relation holds: $D(p\|q) = D(p\|r) + D(r\|q)$ for all $p \in \mathcal{H}'$ and $q \in \mathcal{H}$. Then, by the positivity of KL-divergence, $D(p\|q) \geq D(p\|r)$ for all $q \in \mathcal{H}$ so that r is the m-projection of $p \in \mathcal{H}'$ to $\mathcal{H}$. Likewise, $D(p\|q) \geq D(r\|q)$ for all $p \in \mathcal{H}'$ so that r is the e-projection of $q \in \mathcal{H}$ to $\mathcal{H}'$. $\square$

As a consequence of this duality, we may perform m-projection by a dual e-projection. For instance, if we wish to obtain the m-projection of a given $p \in \mathcal{F}$ to an e-flat submanifold $\mathcal{H}$, this is also given by the e-projection of an arbitrary $q \in \mathcal{H}$ to the complementary biorthogonal m-flat submanifold $\mathcal{H}'$ containing p. This may be desirable as iterative scaling algorithms are available for performing e-projection in exponential family graphical models. We discuss these algorithms momentarily but first wish to clarify the relevance of m-projection for maximum-likelihood model thinning.

Model Thinning. Our main interest in information projections is for thinning of graphical models. Suppose we have a graphical model specified by an exponential family of Gibbs distributions $\mathcal{F} = \{p(x;\theta) \propto \exp \sum \phi_\Lambda(x_\Lambda;\theta_\Lambda)\}$ with linearly parameterized potential specification $\phi = (\phi_\Lambda(x_\Lambda;\theta_\Lambda) = \theta_\Lambda \cdot t_\Lambda(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma^\phi)$ as discussed in Section 2.1. This potential specification gives a hypergraph $\mathbf{H}_\Gamma^\phi$ describing the structure of interactions determining the Markov structure $\mathbf{G}_\Gamma^\phi = \text{adj } \mathbf{H}_\Gamma^\phi$.

A natural approach to model reduction then is to consider omitting some of the sufficient statistics $t(x)$ from the model by forcing the corresponding exponential parameters to zero. This gives an embedded family of exponential family models based on a reduced set of sufficient statistics. Pruning the associated hyperedges from $\mathbf{H}_\Gamma^\phi$ then gives a "thinned" adjacency graph $\mathbf{G}_\Gamma$ so that this may be considered as thinning of a graphical model.

From the perspective of information geometry, this embedded family is regarded as an e-flat submanifold $\mathcal{H} \subset \mathcal{F}$. Given an initial model $p \in \mathcal{F}$, we then wish to select $q \in \mathcal{H}$ to best approximate p. The maximum-likelihood principle advises that we select q so as to maximize the expected log-likelihood of samples drawn from q. Equivalently, this may be posed as minimizing the KL-divergence $D(p\|q)$ over $q \in \mathcal{H}$ which is precisely the m-projection problem posed earlier. In view of duality, we may also pose this as e-projection.

Let us partition the statistics as $t(x) = (t_\mathcal{H}(x), t'_\mathcal{H}(x))$ where $t_\mathcal{H}(x)$ are the sufficient statistics of the embedded family and $t'_\mathcal{H}(x)$ are those statistics to be neglected. Accordingly, partition the exponential coordinates $\theta = (\theta_\mathcal{H}, \theta'_\mathcal{H})$ and the moment coordinates $\eta = (\eta_\mathcal{H}, \eta'_\mathcal{H})$. The e-flat submanifold $\mathcal{H}$ may be specified by restriction of the exponential coordinates as:

$$\mathcal{H} = \{q \in \mathcal{F} \mid \theta'_\mathcal{H}(q) = 0\} \tag{2.110}$$

For any $p \in \mathcal{F}$ we may define an m-flat submanifold $\mathcal{H}'(p)$ by restriction of moment coordinates:

$$\mathcal{H}'(p) = \{r \in \mathcal{F} \mid \eta_{\mathcal{H}}(r) = \eta_{\mathcal{H}}(p)\} \tag{2.111}$$

These are complementary biorthogonal submanifolds since for arbitrary $p \in \mathcal{H}'$, $q \in \mathcal{H}$ and $r \in \mathcal{H} \cap \mathcal{H}'$ we have:

$$\begin{aligned}(\eta(p) - \eta(r)) \cdot (\theta(r) - \theta(q)) &= (\eta_{\mathcal{H}}(p) - \eta_{\mathcal{H}}(r)) \cdot (\theta_{\mathcal{H}}(r) - \theta_{\mathcal{H}}(q)) \\ &\quad + (\eta'_{\mathcal{H}}(p) - \eta'_{\mathcal{H}}(r)) \cdot (\theta'_{\mathcal{H}}(r) - \theta'_{\mathcal{H}}(q)) \\ &= (\eta_{\mathcal{H}}(p) - \eta_{\mathcal{H}}(p)) \cdot (\theta_{\mathcal{H}}(r) - \theta_{\mathcal{H}}(q)) \\ &\quad + (\eta'_{\mathcal{H}}(p) - \eta'_{\mathcal{H}}(r)) \cdot (0 - 0) \\ &= 0 \end{aligned} \tag{2.112}$$

Also, in this context, minimum-discrimination reduces to maximum-entropy. This is shown in that, for $p, \tilde{p} \in \mathcal{H}'$ and $q \in \mathcal{H}$, we have:

$$\begin{aligned} D(\tilde{p}\|q) - D(p\|q) &= (\varphi^*(\tilde{p}) - \varphi^*(p)) + (\eta(\tilde{p}) - \eta(p)) \cdot \theta(q) \\ &= (\varphi^*(\tilde{p}) - \varphi^*(p)) + (\eta_{\mathcal{H}}(p) - \eta_{\mathcal{H}}(p)) \cdot \theta_{\mathcal{H}}(q) \\ &= \varphi^*(\tilde{p}) - \varphi^*(p) \end{aligned} \tag{2.113}$$

Hence, maximizing $h[\tilde{p}] = -\varphi^*(\tilde{p})$ over $\tilde{p} \in \mathcal{H}'$ also minimizes $D(\tilde{p}\|q)$ over $\tilde{p} \in \mathcal{H}'$ for any fixed $q \in \mathcal{H}$. These observations provide for the following proposition summarizing the pertinent information geometry of model thinning:

Proposition 15 (Model Thinning) *Let $\mathcal{H} \neq \varnothing$ be an embedded e-flat submanifold of an exponential family $\mathcal{F}$ as in (2.110). Let $\mathcal{H}'$ be the complementary biorthogonal m-flat submanifold containing $p \in \mathcal{F}$ as in (2.111). Then there exists a probability distribution $r \in \mathcal{H} \cap \mathcal{H}'$ satisfying the following (equivalent) conditions:*

(i) $r = \arg\min_{q \in \mathcal{H}} D(p\|q)$

(ii) $r = \arg\max_{\tilde{p} \in \mathcal{H}'} h[\tilde{p}]$

(iii) $\theta'_{\mathcal{H}}(r) = 0 \ \wedge \ \eta_{\mathcal{H}}(r) = \eta_{\mathcal{H}}(p)$

Moreover, $\mathcal{H} \cap \mathcal{H}' = \{r\}$ and this r is uniquely determined by any one of these conditions. Also, $D(p\|r) = h[r] - h[p]$.

Proof. This follows from the preceding discussion and Proposition 14. Also, $D(p\|r) = D(p\|q) - D(r\|q) = \varphi^*(p) - \varphi^*(r) = h[r] - h[p]$. □

This gives dual perspectives for model thinning: (i) maximum-likelihood over $\mathcal{H}$ (m-projection), (ii) maximum-entropy over $\mathcal{H}'$ (e-projection). Necessary and sufficient conditions are provided by (iii) which may be regarded as either;

1. *Moment matching.* Find $r \in \mathcal{H}$ solving $\eta_{\mathcal{H}}(r) = \eta_{\mathcal{H}}(p)$.

2. *Parameter annihilation.* Find $r \in \mathcal{H}'$ solving $\theta'_{\mathcal{H}}(r) = 0$.

While these are equivalent insofar as they give the same solution, they suggest different approaches to model thinning. In the next section, we discuss a moment matching approach based on iterative e-projections. We also remark that model thinning may be performed inductively by successive m-projection to embedded families:

Proposition 16 (Inductive M-Projection, Amari [4]) *Let $\{\mathcal{H}_k\}$ be a sequence of embedded e-flat submanifolds of $\mathcal{F}$ such that $\mathcal{H}_0 \equiv \mathcal{F} \supset \mathcal{H}_1 \supset \ldots \supset \mathcal{H}_K$. For probability distribution $p \in \mathcal{F}$ define $\hat{p}^{(k)} = \arg\min_{q \in \mathcal{H}_k} D(p||q)$ the m-projection of p to $\mathcal{H}_k$. Then, for $k = 1, \ldots, K$ we have*

$$\hat{p}^{(k)} = \arg\min_{q \in \mathcal{H}_k} D(\hat{p}^{(k-1)} \| q) \tag{2.114}$$

so that m-projection may be performed inductively. Also, the cumulative KL-divergence is additive:

$$D(p||\hat{p}^{(K)}) = \sum_{k=1}^{K} D(\hat{p}^{(k-1)} \| \hat{p}^{(k)}) \tag{2.115}$$

Proof. The result follows by inductive application of Propositions 11 and 12. □

Hence, we may perform optimal (maximum-likelihood) model thinning by incrementally releasing moment constraints (reducing graphical structure) and maximizing entropy. This may be viewed as a selective "forgetting" algorithm which, in a sense, is the inverse of the iterative scaling techniques (usually employed for model identification) we discuss next.

2.2.4 Iterative Scaling and Covariance Selection

Csiszár [34] formalized the iterative scaling (IS) procedure for evaluating the e-projection of a probability distribution $q \in \mathcal{F}$ to an m-flat submanifold $\mathcal{H}'$.

$$p^* = \arg\min_{p \in \mathcal{H}'} D(p\|q) \tag{2.116}$$

This procedure consists of alternating e-projections to a collection of m-flat submanifolds $\{\mathcal{H}'_i\}_{i \in \mathcal{I}}$ with intersection $\cap_{i \in \mathcal{I}} \mathcal{H}'_i = \mathcal{H}'$. Let $(i(k) \in \mathcal{I}, k = 1, 2, 3, \ldots)$ be a sequence of indices drawn from $\mathcal{I}$ such that each index is included infinitely often. Then, the IS procedure generates a sequence of probability distributions given by alternating e-projections

$$q^{(k+1)} = \arg\min_{p \in \mathcal{H}'_{i(k+1)}} D(p||q^{(k)}) \tag{2.117}$$

where the sequence is initialized by $q^{(0)} = q$. As shown by Csiszár, this sequence of e-projections asymptotically converges to the desired $p^* \in \mathcal{H}'$ with the KL-divergence $D(p^* \| q^{(k)})$ monotonically decreasing to zero. This idea is illustrated in Figure 2-11. We also remark that this IS procedure is a special case of Bregman's relaxation method [24]. See Bauschke and Borwein [11] for discussion of the method of random Bregman projections.

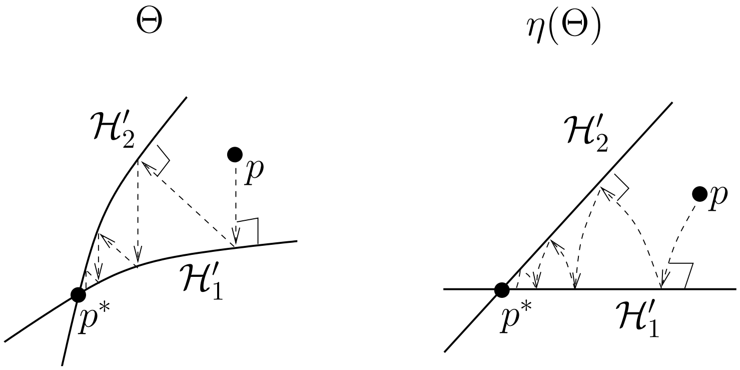


Figure 2-11: Illustration of iterative scaling procedure for finding the point p^* in the intersection of a set of m-flat submanifolds nearest to the starting point p in KL-divergence $D(p^* \| p)$. The procedure generates a sequence of e-projections to individual submanifolds approaching the desired p^*. Often, as in this example, the m-flat submanifolds are chosen so that the intersection uniquely determines the point p^* consistent with the linear moment constraints collectively imposed by these m-flat submanifolds. In that case, the starting point p may be chosen arbitrarily.

In order for this procedure to be of practical use, we must choose the submanifolds $\{\mathcal{H}'_i\}$ such that the alternating e-projections are given by a tractable calculation. Next, we consider such a case in the context of graphical models.

Iterative Proportional Fitting. The iterative proportional fitting (IPF) procedure of Ireland and Kullback [71] is an iterative procedure for adjusting the parameters of a graphical model to give prescribed marginal distributions. Consider a graphical model based on a graph $\mathbf{G}_\Gamma$ with probability distribution $p(x_\Gamma) \propto \prod_{\Lambda \in \mathcal{C}(\mathbf{G}_\Gamma)} \psi_\Lambda(x_\Lambda)$. IPF adjusts the compatibility functions $\psi_\Lambda(x_\Lambda)$ to give a prescribed set of (consistent) marginal distributions $(p^*(x_\Lambda), \Lambda \in \mathcal{C})$ specified over some subset of the cliques $\mathcal{C} \subset \mathcal{C}(\mathbf{G}_\Gamma)$. If these selected cliques $\mathcal{C}$ contain the maximal cliques of the graph $\mathcal{C}^*(\mathbf{G}_\Gamma)$, then IPF solves for the maximum entropy distribution subject to those marginal constraints. More generally, if we only impose marginal constraints on a subset of the maximal cliques, then IPF may be seen as solving for the e-projection (minimizing KL-divergence) of the given graphical model to the subfamily of graphical models on $\mathbf{G}_\Gamma$ consistent with these prescribed marginal distributions.

IPF operates by iterating over the selected cliques $\Lambda \in \mathcal{C}$ and updating the associated compatibility function $\psi_\Lambda(x_\Lambda)$ by the ratio of the desired marginal $p^*(x_\Lambda)$ to the actual marginal $p(x_\Lambda)$.

$$\hat{\psi}_\Lambda(x_\Lambda) = \psi_\Lambda(x_\Lambda) \times \frac{p^*(x_\Lambda)}{p(x_\Lambda)} \tag{2.118}$$

Upon replacing ψ_Λ by $\hat{\psi}_\Lambda$, this gives an updated graphical model with probability distribution $\hat{p}(x_\Gamma)$ such that $\int \hat{p}(x_\Gamma) dx_{\setminus \Lambda} = p^*(x_\Lambda)$. Iterating this update procedure over cliques in $\mathcal{C}$ gives a sequence of graphical models on $\mathbf{G}_\Gamma$ with marginal distributions approaching the prescribed marginals. Note that, at each step, computation of the marginal distribution of the selected clique is required. In exponential families, this IPF procedure may be seen as a special case of the iterative scaling procedure (Jordan [77]) as we now discuss.

Consider an exponential family $\mathcal{F}$ of graphical models based on statistics $t(x) = (t_\Lambda(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$ defined relative to hypergraph $\mathbf{H}_\Gamma$. The family $\mathcal{F}$ is Markov with respect to graph $\mathbf{G}_\Gamma = \text{adj } \mathbf{H}_\Gamma$. Recall the exponential family modeling problem, where we wish to determine the exponential coordinates $\theta^* \in \Theta$ which give a prescribed set of moment coordinates $\eta^* \in \eta(\Theta)$. That is, we wish to solve $E_\theta\{t(\mathrm{x})\} = \eta^*$ for the unique[16] solution in Θ. The variational formulation (2.35) may be regarded as m-projection to the exponential family $\mathcal{F}$. Dually, this may be posed as e-projection of any $q \in \mathcal{F}$ to the family $\mathcal{H}' = \{p | E_p\{t(\mathrm{x})\} = \eta^*\}$. In any case, the solution is given by $p^* \in \mathcal{F} \cap \mathcal{H}'$, the unique member of the exponential family with moment coordinates η^*. We may perform the e-projection by iterative scaling as follows. For each hyperedge Λ of $\mathbf{H}_\Gamma$ (a clique in $\mathbf{G}_\Gamma$), define an associated m-flat submanifold by

$$\mathcal{H}'_\Lambda = \{p | E_p\{t^\Lambda(x_\Lambda)\} = (\eta^*)^\Lambda\} \tag{2.119}$$

which imposes just the moments constraints $(\eta^*)^\Lambda$ defined within subfield Λ.[17] The intersection of these submanifolds, taken over all hyperedges of $\mathbf{H}_\Gamma$, is precisely $\mathcal{H}'$. Given any initial starting point $q^{(0)} = f(\cdot; \theta^{(0)}) \in \mathcal{F}$, alternating e-projections to these m-flat submanifolds gives a sequence of probability distributions $q^{(k)} = f(\cdot; \theta^{(k)}) \in \mathcal{F}$ which asymptotically converges to the desired p^* with moment coordinates η^*. Thus, the sequence $\theta^{(k)}$ converges to the desired θ^* such that $E_{\theta^*}\{t(\mathrm{x})\} = \eta^*$.

Each of the requisite alternating e-projections is performed as follows. By Kullback's minimum-discrimination theorem (Proposition 8), the IS update for hyperedge Λ (e-projection to $\mathcal{H}'_\Lambda$) corresponds to multiplication by an exponential factor based on just the statistics $t^\Lambda(x_\Lambda)$.

$$q^{(k+1)}(x) \propto q^{(k)}(x) \exp\{\delta\theta^\Lambda \cdot t^\Lambda(x_\Lambda)\} \tag{2.120}$$

This shows that each e-projection stays within the exponential family $\mathcal{F}$ and that just the exponential parameters θ^Λ are updated by e-projection to $\mathcal{H}'_\Lambda$,

$$(\theta^{(k+1)})^\Lambda = (\theta^{(k)})^\Lambda + \delta\theta^\Lambda \tag{2.121}$$

In general, the parameter update $\delta\theta^\Lambda$ is determined by the condition $E_{\theta^{(k+1)}}\{t^\Lambda(\mathrm{x}_\Lambda)\} = (\eta^*)^\Lambda$ and some method must be given to solve this nonlinear system of equations. If the exponential family $\mathcal{F}$ of graphical models is marginalizable,[18] this IS update

[16] Assuming minimal representation of the exponential family in terms of linearly independent statistics. Otherwise, η^* might specify a degenerate manifold in Θ satisfying $E_\theta\{t(\mathrm{x})\} = \eta^*$ and we wish to determine some point in this degenerate manifold.

[17] We could also perform iterative scaling with respect to just the *maximal* hyperedges of $\mathbf{H}_\Gamma$ (not a subset of some other hyperedge) or, alternatively, the maximal cliques of $\mathbf{G}_\Gamma$.

[18] Such that the marginal distributions $p(x_\Lambda)$ on hyperedges $\Lambda \in \mathcal{H}_\Gamma$ are exponential distributions

is equivalent to iterative proportional fitting. Then, the moment coordinates $(\eta^*)^\Lambda$ uniquely determine the marginal distribution $p^*(x_\Lambda) \propto \exp\{\beta^*_\Lambda \cdot t^\Lambda(x_\Lambda)\}$ satisfying $E_{\beta^*_\Lambda}\{t^\Lambda(\mathrm{x}_\Lambda)\} = (\eta^*)^\Lambda$. Given these marginal distributions $p^*(x_\Lambda)$, the IS update above simplifies to the earlier IPF update. This is shown by integrating both sides of (2.120) over $\mathcal{X}_{\setminus\Lambda}$ which gives

$$p^*(x_\Lambda) \propto q^{(k)}(x_\Lambda)\exp\{\delta\theta^\Lambda \cdot t^\Lambda(x_\Lambda)\} \tag{2.122}$$

such that the exponential update factor is proportional to the IPF update $p^*(x_\Lambda)/q^{(k)}(x_\Lambda)$. This shows the connection between iterative scaling and iterative proportional fitting in exponential families. The parameter update may then by calculated as

$$\delta\theta^\Lambda = \beta^*_\Lambda - \beta^{(k)}_\Lambda \tag{2.123}$$

where β^*_Λ and $\beta^{(k)}_\Lambda$ are respectively the marginal exponential coordinates of $p^*(x_\Lambda)$ and $q^{(k)}(x_\Lambda)$. Hence, inference is required to compute the marginal distribution $q^{(k)}(x_\Lambda)$.

The main disadvantage of IPF is that many iterations over the collection of m-flat submanifolds may be required before an acceptable level of convergence is achieved. Each such e-projection requires a global inference operation to calculate the current marginal distribution $q^{(k)}(x_\Lambda)$ of the subfield being updated. Several modifications (discussed next) of this IS/IPF approach have been developed which attempt to reduce the number of requisite global inference operations by updating all parameters of the model at each iteration.

Generalized Iterative Scaling. Darroch and Radcliff [38] developed the generalized iterative scaling (GIS) procedure to accelerate the convergence of IS by performing updates "in parallel". This is related to IPF by the update formula:

$$q^{(k)}(x) \propto q^{(k-1)}(x) \times \prod_{\Lambda\in\mathcal{C}} \left(\frac{p^*(x_\Lambda)}{q(x_\Lambda)}\right)^{c_\Lambda} \tag{2.124}$$

where the coefficients $(c_\Lambda, \Lambda \in \mathcal{C})$ are positive and sum to one. In exponential coordinates, this gives GIS updates which are convex combinations of IS updates,

$$\delta\theta = \sum_{\Lambda\in\mathcal{C}} c_\Lambda \delta\theta^\Lambda \tag{2.125}$$

where the IS updates $\delta\theta^\Lambda$ are zero-padded in taking the sum. In view of the strict convexity of the KL-divergence $D(p||q)$ in the exponential coordinates of q, we see that the KL-divergence is monotonically decreased by these GIS updates. See Csiszár for convergence analysis from the perspective of information geometry [35].

We also mention the improved iterative scaling (IIS) procedure of Della Pietra et al [106]. This is similar to GIS but where the convexity constraint is relaxed. This may be viewed as adding a positive "gain" parameter scaling the coefficients c_Λ.

based on statistics $t^\Lambda(x_\Lambda)$.

Optimizing this gain parameter to minimize a certain tractable convex upper-bound of the KL-divergence then gives a simple optimality condition which may be solved efficiently by a one-parameter search. This is shown to converge by analysis similar to that as for the expectation-maximization algorithm.

Application for Covariance Selection. We now consider iterative scaling for Gaussian graphical models (GMRFs). Dempster [43] introduced the fundamental covariance selection problem. Here, one considers a Gaussian random vector of known mean but unknown covariance. Observing that this may be viewed as an exponential family parameterized by the inverse covariance, Dempster proposes estimation of the covariance while positing some zero elements of the inverse covariance matrix. It is interesting to note that Dempster's motivation for this approach was purely from the perspective of parameter reduction rather than any assumption of Markov structure:

> "Two main currents of thought underlie the covariance fitting technique...The first is the principle of parsimony in parametric model fitting, which suggests that parameters should be introduced sparingly and only when the data indicate they are required." (Dempster [43])

It was shown by Speed and Kiiveri [124] that these zeros in the inverse covariance matrix describe conditional independencies (such as we have seen by Proposition 6). Once the zero-pattern has been specified, maximum-likelihood reduces to adjusting the (nonzero) exponential parameters to give the prescribed marginals. Dempster proposed two iterative approaches. One approach fixes parameters to zero while iteratively adjusting moments, the other fixes moments while iteratively driving parameters to zero. Speed and Kiiveri define corresponding cyclic methods interpreted as alternating information projections. One of these is iterative scaling as we now discuss. Here, we specify iterative scaling updates for the information parameterization of the GMRF. This is a minor extension of the covariance selection problem obtained by viewing the information form of the Gaussian density as an exponential family with unknown means as well as unknown covariance.

Consider the family of regular GMRFs $(\mathrm{x}_\Gamma, \mathbf{G}_\Gamma)$. Recall that this may be represented as an exponential family $\mathcal{F}$ with information parameters (h, J) where J is structured to respect the conditional independencies imposed by $\mathbf{G}_\Gamma$. Suppose that we wish to impose a collection of (consistent) marginal moment constraints $(\hat{x}^*_\Lambda, P^*_\Lambda)$ for $\Lambda \in \mathcal{H}_\Gamma \subset \mathcal{C}(\mathbf{G}_\Gamma)$. Within the Gaussian family, this is equivalent to imposing Gaussian marginal distributions $\mathrm{x}_\Lambda \sim \mathcal{N}(\hat{x}^*_\Lambda, P^*_\Lambda)$. In the exponential family description, this is equivalent to specifying subsets of moment coordinates $(\eta^*)^\Lambda = (\hat{x}^*_\Lambda, P^*_\Lambda + \hat{x}^*_\Lambda(\hat{x}^*_\Lambda)')$. If $\mathcal{H}_\Gamma$ covers $\mathbf{G}_\Gamma$[19], then these constraints collectively specify the moment coordinates η^* of some distribution $p^* \in \mathcal{F}$. We may then determine the exponential coordinates $\theta^* = (h^*, -\frac{1}{2}J^*)$ of p^* by iterative scaling. The *marginal* information parameters, defined as $(\hat{h}_\Lambda, \hat{J}_\Lambda) = (P_\Lambda^{-1}\hat{x}_\Lambda, P_\Lambda^{-1})$, correspond to

[19] Such that each vertex $\gamma \in \Gamma$ and each edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ of $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ is contained in some $\Lambda \in \mathcal{H}_\Gamma$. For instance, $\mathcal{H}_\Gamma = \mathcal{C}^*(\mathbf{G}_\Gamma)$ or $\mathcal{H}_\Gamma = \mathcal{E}_\Gamma \cup \{\{\gamma\} | \gamma \in \Gamma\}$ both satisfy this condition.

marginal exponential parameters $\beta_\Lambda = (\hat{h}_\Lambda, -\frac{1}{2}\hat{J}_\Lambda)$. Hence, the IS parameter update, given by (2.121) and (2.123), may be seen as updating the information parameters (h_Λ, J_Λ) according to

$$\begin{aligned} h_\Lambda^{(k+1)} &= h_\Lambda^{(k)} + (\hat{h}_\Lambda^* - \hat{h}_\Lambda^{(k)}) && (2.126) \\ J_\Lambda^{(k+1)} &= J_\Lambda^{(k)} + (\hat{J}_\Lambda^* - \hat{J}_\Lambda^{(k)}) && (2.127) \end{aligned}$$

where $(\hat{h}_\Lambda^*, \hat{J}_\Lambda^*) = ((P_\Lambda^*)^{-1}\hat{x}_\Lambda^*, (P_\Lambda^*)^{-1})$ and $(\hat{h}_\Lambda^{(k)}, \hat{J}_\Lambda^{(k)}) = ((P_\Lambda^{(k)})^{-1}\hat{x}_\Lambda^{(k)}, (P_\Lambda^{(k)})^{-1})$. Note that this requires computation of the current marginal moments $(\hat{x}_\Lambda^{(k)}, P_\Lambda^{(k)})$. Equivalently, this may be viewed as iterative proportional fitting.

$$\begin{aligned} q^{(k+1)}(x) &\propto q^{(k)}(x) \times \frac{p^*(x_\Lambda)}{q^{(k)}(x_\Lambda)} \\ &\propto q^{(k)}(x) \times \frac{\exp\{-\frac{1}{2}x'_\Lambda \hat{J}_\Lambda^* x_\Lambda + \hat{h}_\Lambda^* \cdot x_\Lambda\}}{\exp\{-\frac{1}{2}x'_\Lambda \hat{J}_\Lambda^{(k)} x_\Lambda + \hat{h}_\Lambda^{(k)} \cdot x_\Lambda\}} \\ &\propto q^{(k)}(x) \exp\{-\frac{1}{2}x'_\Lambda(\hat{J}_\Lambda^* - \hat{J}_\Lambda^{(k)})x_\Lambda + (\hat{h}_\Lambda^* - \hat{h}_\Lambda^{(k)}) \cdot x_\Lambda\} && (2.128) \end{aligned}$$

After the update we have that $(\hat{x}_\Lambda^{(k+1)}, P_\Lambda^{(k+1)}) = (\hat{x}_\Lambda^*, P_\Lambda^*)$. The corresponding GIS updates are convex combinations of these IPF updates,

$$\begin{aligned} h^{(k+1)} &= h^{(k)} + \sum_{\Lambda \in \mathcal{C}} c_\Lambda(\hat{h}_\Lambda^* - \hat{h}_\Lambda^{(k)}) && (2.129) \\ J^{(k+1)} &= J^{(k)} + \sum_{\Lambda \in \mathcal{C}} c_\Lambda(\hat{J}_\Lambda^* - \hat{J}_\Lambda^{(k)}) && (2.130) \end{aligned}$$

where it is understood that the local updates are zero-padded in evaluating the sum. The (appropriately scaled) IIS updates are given similarly. Note that the Gaussian IS update formula (2.126-2.127), and hence the GIS/IIS update as well, respect the sparsity constraints imposed on J by the graph $\mathbf{G}_\Gamma$ such that these updates stay within the family $\mathcal{F}$ of GMRFs on $\mathbf{G}_\Gamma$.

2.2.5 Extensions of Maximum-Likelihood

In this last subsection we consider extensions of maximum-likelihood for selecting an approximate model from among a collection of variable-order parameterized families. This is related to Dempster's motivation for considering thinned covariance selection models. We mainly focus upon the perspective of Akaike.

Akaike's Information Criterion. Akaike [1, 2] developed a generalization of maximum likelihood parameter estimation which addresses the issue of order estimation. Akaike's information criterion (AIC) is also based on the KL-divergence but is more general than maximum-likelihood in that it accounts for variable-order models. As in Section 2.2.2, we observe a set $x^N = (x_1, \ldots, x_N)$ of independent, identically distributed samples from the (unknown) true generative distribution $g(x)$ which we wish to model. But now we must select our working model of g from among a set of candidate families of parameterized models of variable order.

Let $\{f_k(x;\theta^k)\}$ denote this set of families indexed by k where θ^k indicates the parameters of the k-th family. Also, let r_k denote the order of each family which is just the number of parameters θ^k. The idea behind the AIC is to specify a model selection metric which minimizes an approximately unbiased estimate of the expected KL-divergence $E\{D(g||f_k(\cdot;\hat{\theta}^k_{ML}(\mathrm{x}^N)))\}$, based upon the data x^N, where $\hat{\theta}^k_{ML}(x^N)$ is the maximum likelihood parameter estimate of the k-th family. This generalizes the notion of m-projection performed by maximum likelihood parameter estimation by selecting how rich a family of models to use so as to minimize this estimate of the modeling error. Intuitively, higher-order families of models are richer and could in principle better approximate $g(x)$ but embedded lower-order families of models have fewer parameters to estimate and are less prone to estimation error.

Akaike's argument is based on several asymptotic approximations. See Pawitan [103] for a concise derivation. The resulting criterion then selects the family of models which maximizes the following model metric.

$$-\frac{1}{2}AIC = \log f_k(x^N; \hat{\theta}^k(x^N)) - r_k \tag{2.131}$$

Thus we see that the AIC is equivalent to maximum likelihood except that a penalty is assessed to more complex (higher-order) models thus addressing the problem of over-fitting of the data.

Thus the AIC approach indicates a fundamental connection between the information theoretic notion of maximum-entropy modeling and selecting the "best" model among a class of order restricted models. In the case of exponential models this corresponds to choosing as few sufficient statistics as possible while still providing a faithful model. For GMRFs this may be posed as limiting the number of pairwise interactions between sites within the field by setting selected off-diagonal elements of the inverse covariance matrix to zero.

Related Criteria. It should also be remarked that several other model-selection criteria have since been developed which lead to similar conclusions as in the AIC and also have connections to information theory. This includes the Bayesian information criterion (BIC) developed by Schwarz [120] and the minimum description length principle (MDL) developed by Rissanen [111, 112, 113]. The BIC is developed from the Bayesian inference philosophy where one introduces a prior distribution over the space of candidate models and then chooses the model which maximizes the conditional probability of the model given the data. The BIC, however, does not actually depend upon the choice of prior as it corresponds to a 2nd order asymptotic expansion of this posterior model probability which proves to be independent of the prior model distribution under certain conditions. The BIC differs from the AIC in that it introduces a factor of $\frac{1}{2}\log N$, where N is the number of samples, in front of the model order such that (for large N) it favors lower-order models than in the AIC. The

BIC, however, is shown to be asymptotically optimal in the Bayesian sense. Rissanen developed the related MDL criterion from a coding perspective as MDL selects the model which minimizes the joint description length of both the data and the model (the minimum number of bits required to encode both the model and the data). By Shannon's coding theorem, this is closely related to the KL-minimization underlying the AIC since the KL-divergence between the true distribution and the model used to encode the data gives the expected penalty in the description length of each sample-path of the data. Minimizing the description length of the model then generalizes Akaike's notion of model complexity which may be regarded as a naive estimate of the model description length. The AIC, however, provides sufficient motivation for our present purpose.

This concludes the theoretical discussion concerning the intersection of information theory and statistical modeling. We now turn our attention towards inference.

2.3 Recursive Inference

In this section we consider recursive approaches to inference (both exact and approximate) appropriate for MRFs. The fundamental inference problem we address is evaluation (or approximation) of the marginal distributions of a graphical model. We have described very general families of graphical models given by exponential families of Gibbs distributions. These give representation of the probability distribution $p(x_\Gamma)$ of a random field x_Γ in terms of local interaction potentials $\phi(x_\Lambda; \theta_\Lambda) = \theta_\Lambda \cdot t_\Lambda(x_\Lambda)$. The structure of these interactions describes a hypergraph $\mathbf{H}_\Gamma^\phi$ which determines the Markov structure of the random field by the associated adjacency graph $\mathbf{G}_\Gamma^\phi = \text{adj } \mathbf{H}_\Gamma^\phi$. Given such a graphical model, we then wish to evaluate the marginal distributions

$$p(x_\Lambda) = \int_{\mathcal{X}_{\setminus\Lambda}} p(x_\Lambda) dx_{\setminus\Lambda} \tag{2.132}$$

for selected subfields specified by $\Lambda \subset \Gamma$. Often, we wish to evaluate these marginal distributions "in parallel" for a collection of subfields $C = \{\Lambda_i \subset \Gamma | i \in \mathcal{I}\}$. These might be chosen as just the sites of the field with $C = \{\{\gamma\} | \gamma \in \Gamma\}$. In this case, inference may be regarded as evaluating the m-projection of $p(x_\Gamma)$ to the family of "fully factored" distributions (corresponding to completely disconnected graphical models having only singleton interaction potentials). This m-projection is nothing but the product of marginals distributions, $\hat{p}(x_\Gamma) = \prod_{\gamma \in \Gamma} p(x_\gamma)$. Alternatively, we may choose these subfields C to be the hyperedges $\mathcal{H}_\Gamma^\phi$ of the graphical model. This is desirable, for instance, when we wish to calculate the moment coordinates $\eta = E_\theta t(\mathrm{x})$ of an exponential family model for specified exponential coordinates θ. The moment parameters may then be evaluated by $\eta_\Lambda = \int p(x_\Lambda) t_\Lambda(x_\Lambda) dx_\Lambda$ for all $\Lambda \in \mathcal{H}_\Gamma^\phi$.

Recursive inference procedures are best illustrated in the simple case of acyclic graphical models. These are discussed in Section 2.3.1. Later sections show how these inference procedures have been extended for inference in MRFs defined on graphs with cycles. Two approaches, junction trees (Section 2.3.2) and multiscale models (Section

2.3.3), are discussed. Both rely upon the idea of grouping sites of the random field so as to provide an equivalent Markov tree description (to be discussed).

2.3.1 Markov Trees

We first focus on recursive inference techniques for acyclic graphical models. In this subsection, we focus mainly on the picture developed by Pearl [105], but also briefly touch upon some variational extensions of this method (Yedidia et al [136], Wainwright [129], Minka [96]). In later subsections (Sections 2.3.2 and 2.3.3), we discuss structured versions of this approach appropriate for more general MRFs and also show the connection to Kalman filtering (Section 2.3.3).

We say that a graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ is a *tree* if it is connected and acyclic. Several examples of trees are illustrated in Figure 2-12. Trees are *singly-connected* since every pair of vertices $\gamma, \lambda \in \Gamma$ are connected by exactly one path. In trees, we say that a vertex $\gamma \in \Gamma$ is a *leaf* if it is adjacent to exactly one other vertex. A *chain* is a tree where every vertex is adjacent to at most two other vertices such as in Figure 2-12(a). More generally, we say that a graph is a *forest* when the subgraphs induced by the connected components of the graph are trees. This describes the most general situation for an acyclic graph.

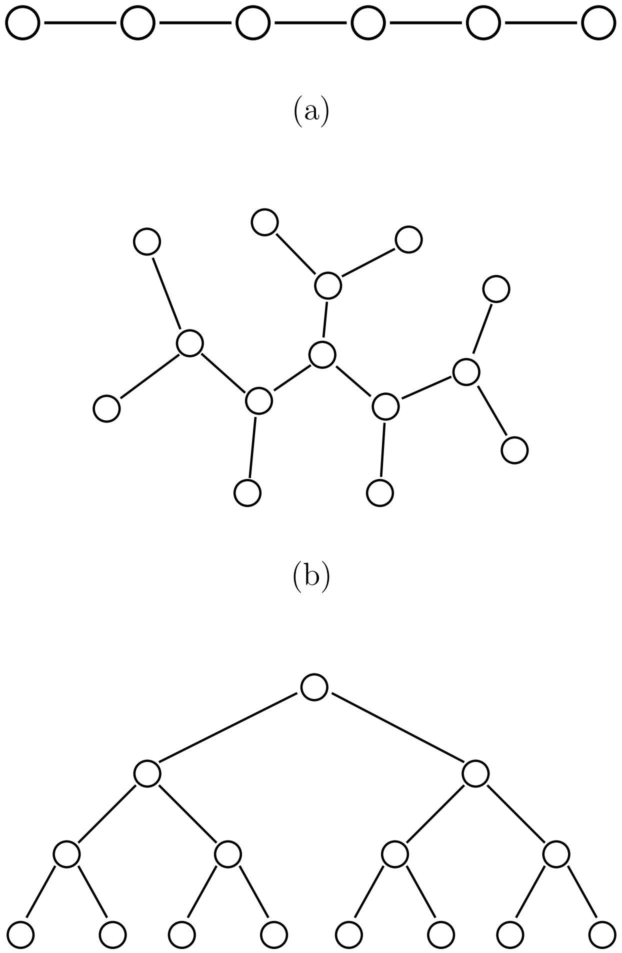
A *Markov tree* is a Markov random field $(x_\Gamma, \mathbf{G}_\Gamma)$ where the graph $\mathbf{G}_\Gamma$ is a tree. This is a *Markov chain* if $\mathbf{G}_\Gamma$ is also a chain. Recursive inference techniques were first developed in the context of Markov chains. Yet, we focus on recursive inference techniques for Markov trees as these apply for Markov chains as well. The more general situation of a forest is treated by applying the following inference procedures for each connected component.

Decimation. By the Hammersley-Clifford theorem, the probability distribution for a Markov tree $(x_\Gamma, \mathbf{G}_\Gamma)$ may be factored into the form

$$p(x_\Gamma) \propto \prod_{\gamma \in \Gamma} \psi(x_\gamma) \prod_{\{\gamma,\lambda\} \in \mathcal{E}_\Gamma} \psi(x_\gamma, x_\lambda) \tag{2.133}$$

in terms of singleton compatibility functions ψ_γ at each site $\gamma \in \Gamma$ and pairwise compatibility functions $\psi_{\gamma,\lambda}$ between adjacent sites. Given such a graphical model, we may calculate the marginal distribution $p(x_{\gamma_0})$ of some arbitrary site $\gamma_0 \in \Gamma$ by the following *decimation* procedure (Jaakkola [74]). Pick any leaf $\lambda \neq \gamma_0$ and *eliminate* λ from the graphical model by marginalizing (integrating or summing) over the state x_λ yielding a graphical model for $p(x_{\Gamma \setminus \lambda})$. Let $\pi(\lambda)$ denote the site adjacent to leaf λ in $\mathbf{G}_\Gamma$. Elimination deletes the vertex λ and the edge $\{\lambda, \pi(\lambda)\}$ from the graph, along with associated compatibility functions ψ_λ and $\psi_{\pi(\lambda),\lambda}$, and replaces $\psi_{\pi(\lambda)}$ by the following updated compatibility function

$$\hat{\psi}(x_{\pi(\lambda)}) = \psi(x_{\pi(\lambda)}) \times \int_{\mathcal{X}_\lambda} \psi(x_{\pi(\lambda)}, x_\lambda)\psi(x_\lambda)dx_\lambda \tag{2.134}$$



(a)

(b)

(c)

Figure 2-12: Diagrams of several connected acyclic graphs (trees). Graph (a) is also a chain. Note that these are singly-connected such that every pair of vertices has a unique path connecting them. Deleting any non-leaf vertex (and its associated edges) separates the graph into two or more connected components.

In exponential family graphical models, compatibility functions correspond to exponents of the form:

$$\psi(x_\gamma) = \exp\{\theta_\gamma \cdot t_\gamma(x_\gamma)\} \tag{2.135}$$

$$\psi(x_\gamma, x_\lambda) = \exp\{\theta_{\{\gamma,\lambda\}} \cdot t_{\{\gamma,\lambda\}}(x_\gamma, x_\lambda)\} \tag{2.136}$$

Hence, if the family is marginalizable, then the elimination step must reduce to the update formula $\hat{\theta}_{\pi(\lambda)} = \theta_{\pi(\lambda)} + \Delta\theta_{\pi(\lambda)}$ where $\Delta\theta_{\pi(\lambda)}$ is determined by

$$\exp\{\Delta\theta_{\pi(\lambda)} \cdot t_{\pi(\lambda)}(x_{\pi(\lambda)})\} \propto \int_{\mathcal{X}_\lambda} \psi(x_{\pi(\lambda)}, x_\lambda)\psi(x_\lambda)dx_\lambda \tag{2.137}$$

For instance, in the information representation of GMRFs, this update formula is given by

$$\hat{h}_{\pi(\lambda)} = h_{\pi(\lambda)} - J_{\pi(\lambda),\lambda}J_\lambda^{-1}h_\lambda \tag{2.138}$$

$$\hat{J}_{\pi(\lambda)} = J_{\pi(\lambda)} - J_{\pi(\lambda),\lambda}J_\lambda^{-1}J_{\lambda,\pi(\lambda)} \tag{2.139}$$

In any case, once λ is eliminated, the resulting graphical model describes a Markov tree $(\mathrm{x}_{\Gamma\setminus\lambda}, \mathbf{G}_{\Gamma\setminus\lambda})$ where $\mathbf{G}_{\Gamma\setminus\lambda} = (\Gamma \setminus \lambda, \mathcal{E}_\Gamma \setminus \{\lambda, \pi(\lambda)\})$ is the subtree of $\mathbf{G}_\Gamma$ induced by $\Gamma \setminus \lambda$. Hence, we may iterate this leaf elimination procedure with respect to the modified graphical model until just site γ_0 remains thereby computing the marginal distribution $p(x_{\gamma_0})$. In a similar manner, the probability distribution $p(x_{\gamma_1}, x_{\gamma_2})$ for a pair of adjacent sites $\{\gamma_1, \gamma_2\} \in \mathcal{E}_\Gamma$ is computed by eliminating all sites except γ_1 and γ_2.

Two-Pass Belief Propagation on Trees. The decimation approach may also be reformulated as a message passing procedure on $\mathbf{G}_\Gamma$. Such message passing inference procedures are generally referred to as *belief propagation* (Pearl [105]).

Let us say that γ_0 is the *root* of the tree, $\pi(\lambda)$ is the *parent* of λ, and $\Lambda(\gamma) = \{\lambda|\pi(\lambda) = \gamma\}$ are the *children* of γ. Then the decimation procedure may be viewed as an "upward" message passing procedure where messages are passed from children to parents towards the root. This procedure begins with leaves passing messages to their parents.

$$\mu_{\lambda\to\pi(\lambda)}(x_{\pi(\lambda)}) = \int_{\mathcal{X}_\lambda} \psi(x_{\pi(\lambda)}, x_\lambda)\psi(x_\lambda)dx_\lambda \tag{2.140}$$

In marginalizable exponential families, this message correspond to the parameter update $\Delta\theta_{\pi(\lambda)}$ (an information parameter update in GMRFs). The upward message passing proceeds once a non-leaf γ has received messages from all of its children. Then, γ passes a message to its parent incorporating information from each of its children.

$$\mu_{\gamma\to\pi(\gamma)}(x_{\pi(\gamma)}) = \int_{\mathcal{X}_\gamma} \psi(x_{\pi(\gamma)}, x_\lambda)\left\{\psi(x_\lambda)\prod_{\lambda\in\Lambda(\gamma)}\mu_{\lambda\to\gamma}(x_\gamma)\right\}dx_\gamma \tag{2.141}$$

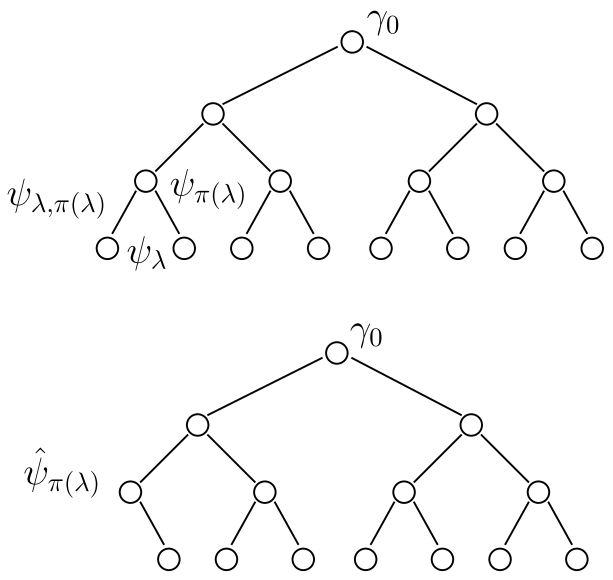


Figure 2-13: Illustration of tree decimation to compute marginal distribution $p(\gamma_0)$. A leaf λ is selected for elimination (top). This elimination step marginalizes over x_λ which just replaces $\psi_{\pi(\lambda)}\psi_{\pi(\lambda),\lambda}\psi_\lambda$ by $\hat{\psi}_{\pi(\lambda)} = \psi_{\pi(\lambda)} \int \psi_{\pi(\lambda),\lambda}\psi_\lambda dx_\lambda$ (bottom). This elimination step is iterated until only site γ_0 remains.

Once γ_0 has received messages from all of its children, the marginal distribution is calculated as $p(x_{\gamma_0}) \propto \psi(x_{\gamma_0}) \prod_{\lambda \in \Lambda(\gamma_0)} \mu_{\lambda \to \gamma_0}(x_{\gamma_0})$.

The advantage of this message passing formulation is that we may define a complementary "downward" procedure which reuses the upward messages to efficiently calculate the marginal distribution at every site of the Markov tree. This procedure begins with the root γ_0 passing a message down to each of its children $\gamma \in \Lambda(\gamma_0)$ which incorporates information from the other children.

$$\mu_{\gamma_0 \to \gamma}(x_\gamma) = \int_{\mathcal{X}_{\gamma_0}} \psi(x_\gamma, x_{\gamma_0}) \left\{ \psi(x_{\gamma_0}) \prod_{\lambda \in \Lambda(\gamma_0) \setminus \gamma} \mu_{\lambda \to}(x_{\gamma_0}) \right\} dx_{\gamma_0} \tag{2.142}$$

As each site γ receives a message from its parent, the downward message-passing continues by passing a message to each child of γ incorporating information from the parent and other children of γ.

$$\mu_{\gamma \to \lambda}(x_\lambda) = \int_{\mathcal{X}_\gamma} \psi(x_\lambda, x_\gamma) \left\{ \psi(x_\gamma) \mu_{\pi(\gamma) \to \gamma}(x_\gamma) \prod_{\lambda' \in \Lambda(\gamma) \setminus \lambda} \mu_{\lambda' \to \gamma}(x_\gamma) \right\} dx_\gamma \tag{2.143}$$

Message passing terminates once every leaf has received a message from its parent.

Fuse-Predict Description of Messages. Note that each message passing step, in both the upward and downward procedures, has the following "sum-product" structure consisting of two-steps: First, information is fused at a given site γ by multi-

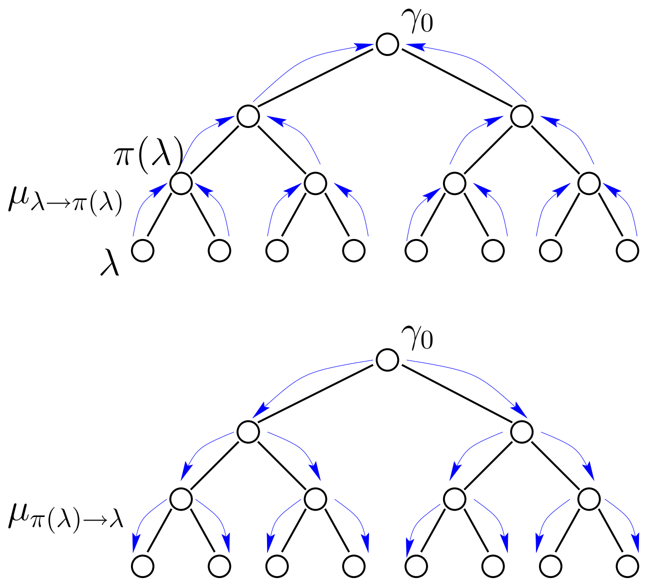

γ_0
$\pi(\lambda)$
$\mu_{\lambda\to\pi(\lambda)}$
λ
γ_0
$\mu_{\pi(\lambda)\to\lambda}$

Figure 2-14: Illustration of two-pass belief propagation. First, an upward message-passing procedure (top) propagates messages from parent to child towards the root. Each non-leaf must wait until it receives messages from all of its children before sending a message to its parent. Second, a downward message-passing procedure (bottom) propagates messages from parent to child until the leaves are reached.

plication of ψ_γ by the product of messages from all but one of the sites adjacent to γ.

$$\hat{\psi}^{\setminus\lambda}(x_\gamma) = \psi(x_\gamma) \prod_{\lambda' \in \partial\gamma \setminus \lambda} \mu_{\lambda' \to \gamma}(x_\gamma) \tag{2.144}$$

Second, this fused information is predicted towards the remaining adjacent site where prediction involves multiplication by the pairwise compatibility function and integration (summation) over x_γ.

$$\mu_{\gamma\to\lambda}(x_\lambda) = \int_{\mathcal{X}_\gamma} \psi(x_\lambda, x_\gamma) \hat{\psi}_{\setminus\lambda}(x_\gamma) dx_\gamma \tag{2.145}$$

In marginalizable exponential families, where messages $\mu_{\lambda\to\gamma}(x_\gamma)$ correspond to parameter updates $\Delta\theta_{\lambda\to\gamma}$, the fusion step sums parameter updates

$$\hat{\theta}_\gamma^{\setminus\lambda} = \theta_\gamma + \sum_{\lambda' \in \partial\gamma \setminus \lambda} \Delta\theta_{\lambda' \to \gamma} \tag{2.146}$$

and the prediction step calculates $\Delta\theta_{\gamma\to\lambda}$ such that

$$\exp\{\Delta\theta_{\gamma\to\lambda} \cdot t_\lambda(x_\lambda)\} \propto \int_{\mathcal{X}_\gamma} \psi(x_\lambda, x_\gamma) \exp\{\hat{\theta}_\gamma^{\setminus\lambda} \cdot t_\gamma(x_\gamma)\} dx_\gamma. \tag{2.147}$$

Marginal Computations. Once any given site γ has received messages from all adjacent sites, the marginal distribution may be calculated as

$$p(x_\gamma) \propto \psi_\gamma(x_\gamma) \prod_{\lambda \in \partial\gamma} \mu_{\lambda\to\gamma}(x_\gamma) \tag{2.148}$$

In exponential families, the marginal distribution $p(x_\gamma) \propto \exp\{\beta_\gamma \cdot t_\gamma(x_\gamma)\}$ is given by summing parameter updates from all adjacent sites.

$$\beta_\gamma = \theta_\gamma + \sum_{\lambda \in \partial\gamma} \Delta\theta_{\lambda\to\gamma} \tag{2.149}$$

The marginal distribution of a pair of adjacent sites $\Lambda = \{\gamma_1, \gamma_2\} \in \mathcal{E}_\Gamma$ is given by

$$p(x_\Lambda) \propto \psi(x_{\gamma_1})\psi(x_{\gamma_2})\psi(x_\Lambda) \prod_{\lambda_1 \in \partial\gamma_1 \setminus \gamma_2} \mu_{\lambda_1\to\gamma_1}(x_{\gamma_1}) \prod_{\lambda_2 \in \partial\gamma_2 \setminus \gamma_1} \mu_{\lambda_2\to\gamma_2}(x_{\gamma_2}) \tag{2.150}$$

and by a similar calculation (adding parameters) in exponential families.

This two-pass message passing procedure was originally developed by Pearl for finite-state Markov trees [105]. Essentially, this is just a recursively structured version of the decimation procedure where stored messages (computed by elimination of subtrees) allows all marginals to be computed in parallel without redundant computation.

Belief-Propagation as Refactorization. A slightly modified description of belief propagation may be viewed as refactoring the graphical model (2.133), into the form

$$p(x_\Gamma) = \prod_{\gamma \in \Gamma} p(x_\gamma) \prod_{\{\gamma,\lambda\} \in \mathcal{E}_\Gamma} \frac{p(x_\gamma, x_\lambda)}{p(x_\gamma)p(x_\lambda)} \tag{2.151}$$

This is accomplished by the same two-pass message passing procedure as before except that the compatibility functions are now adjusted in the course of the procedure as follows. Whenever a message is passed, from λ to γ, the message $\mu_{\lambda\to\gamma}$ is absorbed into $\psi_\gamma \leftarrow \psi_\gamma \times \mu_{\lambda\to\gamma}$ and the inverse message is absorbed into $\psi_{\gamma,\lambda} \leftarrow \psi_{\gamma,\lambda}/\mu_{\lambda\to\gamma}$ (such that the product of compatibility functions is preserved). In this approach, once a site has received messages form all but one of the adjacent sites, it predicts its updated compatibility function (which now contains the appropriate product of messages). Hence, upward messages are the same as before. Likewise, in the downward pass, when we predict from the parent γ back down to λ, the extra factor $\mu_{\lambda\to\gamma}$ in ψ_γ cancels with the inverse factor in $\psi_{\gamma,\lambda}$ so that the downward messages are also the same as before.

Message passing terminates once each site has received messages from all adjacent sites. Then, each ψ_γ contains a message from each of the adjacent sites such that $\psi_\gamma(x_\gamma) \propto p(x_\gamma)$. Furthermore, each $\psi_{\lambda,\gamma}$ has been divided by $\mu_{\lambda\to\gamma}\mu_{\gamma\to\lambda}$ such that the product of updated compatibility functions $\psi_{\gamma,\lambda}\psi_\gamma\psi_\lambda$ gives, after cancellation, $p(x_\gamma, x_\lambda)$ as in (2.150). Hence, $\psi(x_\lambda, x_\gamma) \propto \frac{p(x_\gamma,x_\lambda)}{p(x_\gamma)p(x_\lambda)}$. This shows, after normalization, that the probability distribution $p(x_\Gamma)$ of any Markov tree may be factored as in (2.151). Thus, belief propagation may be viewed as a refactorization procedure for manipulating an arbitrary factorization into this canonical form where marginal distributions are specified locally. Also, this version of belief propagation is more compact in that the computation is performed "in place" requiring storage of just the compatibility functions.

Variational Methods. We also mention some approximate inference methods deriving from belief propagation on trees. The most well known of these methods is *loopy belief propagation* (Yedidia et al [136]). The simplest version of this approach is essentially a "parallel" version of belief propagation but extended to operate in loopy graphical models. This procedure is initiated by each vertex γ of the graphical model sending a message $\mu^{(0)}_{\gamma\to\lambda}(x_\lambda)$ to each of the adjacent vertices $\lambda \in \partial\gamma$.

$$\mu^{(0)}_{\gamma\to\lambda}(x_\lambda) = \int \psi(x_\lambda, x_\gamma)\psi(x_\gamma)dx_\gamma \tag{2.152}$$

Then, on later iterations, we employ the same message-passing structure as in belief propagation on trees. That is, each vertex *fuses* messages from all but one of it's neighbors with the local compatibility function and then *predicts* this information to that neighbor by passing the message:

$$\mu^{(k+1)}_{\gamma\to\lambda}(x_\lambda) = \int \psi(x_\lambda, x_\gamma)\left\{\psi(x_\gamma)\prod_{\lambda'\in\partial\gamma\setminus\lambda}\mu^{(k)}_{\lambda'\to\gamma}(x_\gamma)\right\}dx_\gamma \tag{2.153}$$

At any step of this iteration, we may compute a *pseudo-marginal* $\tilde{p}^{(k)}(x_\gamma)$ for each vertex γ by fusing the local compatibility function $\psi(x_\gamma)$ with messages from all adjacent vertices $\lambda \in \partial\gamma$,

$$\tilde{p}^{(k)}(x_\gamma) = \frac{1}{Z^{(k)}}\psi(x_\gamma)\prod_{\lambda\in\partial\gamma}\mu^{(k)}_{\lambda\to\gamma}(x_\gamma) \tag{2.154}$$

where $Z^{(k)}$ is just the normalization constant. In trees, this iteration is equivalent to two-pass belief propagation such that, after a finite number of iterations, these pseudo-marginals agree with the true marginal distributions. However, it must be emphasized that, in loopy graphs, the interpretation of this method as performing decimation (i.e. marginalization) no longer holds. In general, the loopy approach is not even guaranteed to converge to a fixed point. Even when the method does converge to a fixed point, the pseudo-marginals computed by the method need not agree with the actual marginal distributions.

Nevertheless, loopy belief propagation often does converge to a fixed point and, in some cases, gives very good approximation of the marginal distributions. Based on ideas from mean field theory, Yedidia introduced a variational interpretation of loopy belief propagation as attempting to minimize the *Bethe free energy* over a family of structured approximations of the graphical model [135]. Essentially, this may be understood as minimizing a tractable approximation of KL-divergence.[20] Heskes has shown that any local minimum of this Bethe free energy is a stable fixed point of loopy belief propagation [69]. Also, Weiss and Freeman [131] have shown that, in Gaussian models, stable fixed points of loopy belief propagation are correct insofar as the means of the pseudomarginals then agree with the means of the true marginal distributions (but covariances need not agree). In related work, Wainwright [129] has developed a refactorization formulation of loopy belief propagation and, based on this viewpoint, developed a variant of belief propagation which performs two-pass belief propagation on embedded trees. We should also remark that a variety of other variational methods have been developed with the aim of either: (i) incorporating higher order structure into the approach so as to improve the quality of approximation (Yedidia et al [136], Kappen and Wiegerinck [80]), or (ii) guaranteeing convergence to a local minimum of the Bethe free energy (Yuille [139]).

Finally, we remark that Minka has developed an extension of the loopy belief propagation method for *non-marginalizable* exponential families [96, 95]. Essentially, his method may be seen as performing m-projection back to the exponential family after each prediction step which reduces to a local moment matching procedure. Hence, Minka's method is called *expectation propagation*. Minka has also shown that expectation propagation has a similar variational interpretation as in loopy belief propagation. We point out that expectation propagation also provides a tractable, iterative alternative to exact belief propagation in non-marginalizable exponential families of Markov trees. This non-loopy version of expectation propagation (for trees) is actually closely related to the recursive cavity modeling approach developed in this thesis. However, we develop a structured approach applicable for more general MRFs. Also, in RCM, we develop an adaptive model thinning approach to select which statistics should be included in our exponential family model so as to allow accurate approximations.

2.3.2 Junction Trees

We now consider a recursive inference approach for general MRFs $(x_\Gamma, \mathbf{G}_\Gamma)$, where $\mathbf{G}_\Gamma$ is a "loopy" graph (i.e., there exists cycles in the graph). The main idea is that we may apply the preceding recursive inference methods (belief propagation on trees) to infer marginal distributions of a MRF provided we first convert the MRF into an equivalent Markov tree. This may be accomplished by clustering sites of the MRF, corresponding to separators of the field, and by appeal to the global Markov property. An example illustrating this approach is shown in Figure 2-15. We first discuss this general idea and then show how this leads to the idea of *junction trees*, and the associated inference procedures, which are prevalent in the graphical modeling literature (Shenoy and Shafer [122], Dawid [40], Jordan [77]). The idea of converting a MRF into a Markov tree has also been developed (in parallel with the junction tree perspective favored in the graphical model literature) in the multiscale modeling

[20] However, unlike the e-projection and m-projection problems minimizing KL-divergence, the Bethe free energy is not, in general, convex and may have many local minima.

literature. This author's view of inference in MRFs, and of the role of junction trees, has been especially influenced by this latter perspective (Taylor and Willsky [127], Luettgen et al [92], Daniel [36]). We later consider the role Markov trees play in those methods (Section 2.3.3).

Describing Markov Random Fields as Markov Trees. Suppose that we can partition Γ into $K+1$ disjoint subfields $\mathcal{P} = (S, \Lambda_1, \ldots, \Lambda_K)$ where S separates Λ_i and Λ_j for all i, j.[21] Then, by the global Markov property, $(\mathrm{x}_S, \mathrm{x}_{\Lambda_1}, \ldots, \mathrm{x}_{\Lambda_K})$ forms a Markov tree based on vertices $\mathcal{P}$ with edges $(\{S, \Lambda_k\}, k = 1, \ldots, K)$. This initialization step is illustrated in Figure 2-15(a).

This decomposition procedure may be iterated on subfields to recursively decompose the field as a Markov tree. For instance, suppose that we can further partition Λ_i into $K_i + 1$ disjoint subsets $(S', \Lambda_{i,1}, \ldots, \Lambda_{i,K_i})$ where S' separates $\Lambda_{i,j}$ and $\Lambda_{i,k}$ in subgraph $\mathbf{G}_{\Lambda_i}$ for all $j \neq k$. Then, $S_i \equiv S' \cup \partial\Lambda_i$ likewise decomposes Γ into K_{i+1} disjoint components $(\Lambda_{i,0} \equiv \Gamma \setminus \bar{\Lambda}_i, \Lambda_{i,1}, \ldots \Lambda_{i,K_i})$ separated by S_i. Note that S_i augments S' with $\partial\Lambda_i \subset S$ so as to separate the subfields $\Lambda_{i,k}$ from $\Lambda_{i,0}$. Then, as illustrated in Figure 2-15(b), we may replace leaf node Λ_i in our Markov tree representation by a new subtree with $K_i + 1$ nodes $\{S_i, \Lambda_{i,1}, \ldots, \Lambda_{i,K_i}\}$ where the root S_i takes the place of Λ_i in our Markov tree but we now have K_i new leaf nodes $\Lambda_{i,j}$ each linked to S_i.

In this manner we may recursively decompose the field, "growing" a corresponding Markov tree, until the subfields corresponding to leaves of the tree are sufficiently small so as to be tractable by direct (non-recursive) inference methods. This gives a tree $\mathbf{T}_{\mathcal{S}} = (\mathcal{S}, \mathcal{E}_{\mathcal{S}})$, based on a collection $\mathcal{S}$ of subsets of Γ (the separators and leaves in the preceding decomposition), such that the random field $(x_\Lambda, \Lambda \in \mathcal{E}_{\mathcal{S}})$ is Markov with respect $\mathbf{T}_{\mathcal{S}}$. For instance, in our example, this gives the Markov tree depicted in Figure 2-15(c).

Hence, we may now employ the previously discussed two-pass belief propagation procedure, with some minor modifications[22], to perform inference on the Markov tree thereby inferring marginal distributions of the original MRF. The computational structure of this inference is clarified by considering how decimation on the Markov tree relates to variable elimination in the MRF. We illustrate this approach in Figure 2-16. We let $\pi(S)$ denote both the parent of node S in the Markov tree and the corresponding subfield of the MRF. Eliminating a node $S \in \mathcal{S}$ of the Markov tree corresponds to elimination of a subfield $\Lambda = S \setminus \pi(S) \subset \Gamma$ of the MRF, that is integration of $p(x_\Gamma)$ over $\mathcal{X}_\Lambda$. With respect to subfield Λ, the graphical model of the MRF may be factored into the form

$$p(x_\Gamma) \propto \psi_{\setminus\Lambda}(x_{\setminus\Lambda}) \psi_{\partial\Lambda,\Lambda}(x_{\partial\Lambda}, x_\Lambda) \psi_\Lambda(x_\Lambda) \tag{2.155}$$

[21] Such that any path connecting Λ_i and Λ_j must pass through S.

[22] To accommodate the fact that, in the Markov tree description of the field, we have allowed some states of the MRF to be duplicated at adjacent nodes of the tree. This requires that when a message is predicted from nodes α to β in the tree, we only perform integration (summation) over those states at α which are not duplicates of states at β.

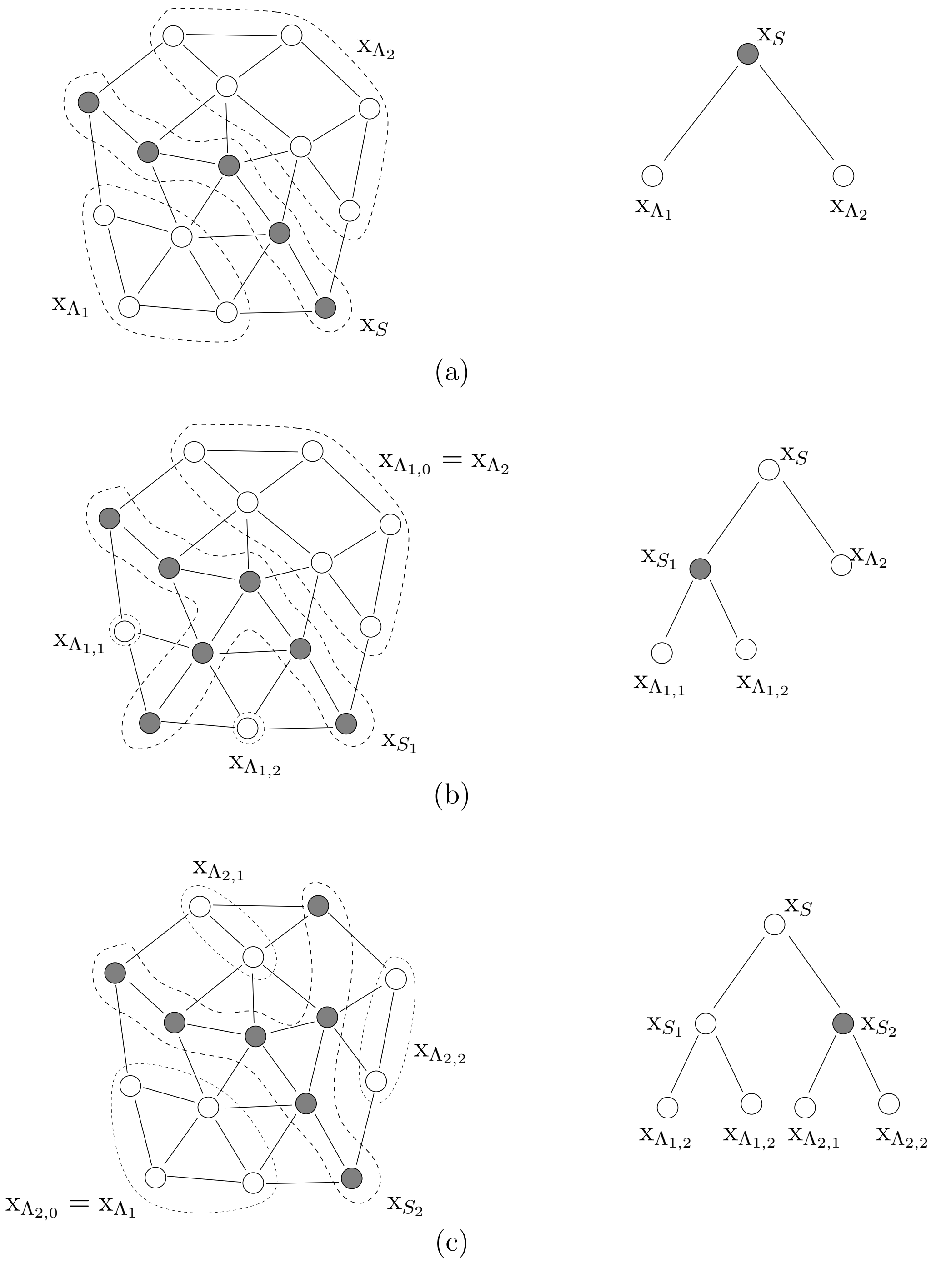

x_{Λ_2}
x_{Λ_1}
x_S
x_S
x_{Λ_1}
x_{Λ_2}
(a)
$\mathrm{x}_{\Lambda_{1,0}} = \mathrm{x}_{\Lambda_2}$
$\mathrm{x}_{\Lambda_{1,1}}$
x_{S_1}
$\mathrm{x}_{\Lambda_{1,2}}$
x_S
x_{S_1}
x_{Λ_2}
$\mathrm{x}_{\Lambda_{1,1}}$
$\mathrm{x}_{\Lambda_{1,2}}$
(b)
$\mathrm{x}_{\Lambda_{2,1}}$
$\mathrm{x}_{\Lambda_{2,2}}$
$\mathrm{x}_{\Lambda_{2,0}} = \mathrm{x}_{\Lambda_1}$
x_{S_2}
x_S
x_{S_1}
x_{S_2}
$\mathrm{x}_{\Lambda_{1,2}}$
$\mathrm{x}_{\Lambda_{1,2}}$
$\mathrm{x}_{\Lambda_{2,1}}$
$\mathrm{x}_{\Lambda_{2,2}}$
(c)

Figure 2-15: Illustration showing decomposition of MRF on a loopy graph (left) as a Markov tree (right).

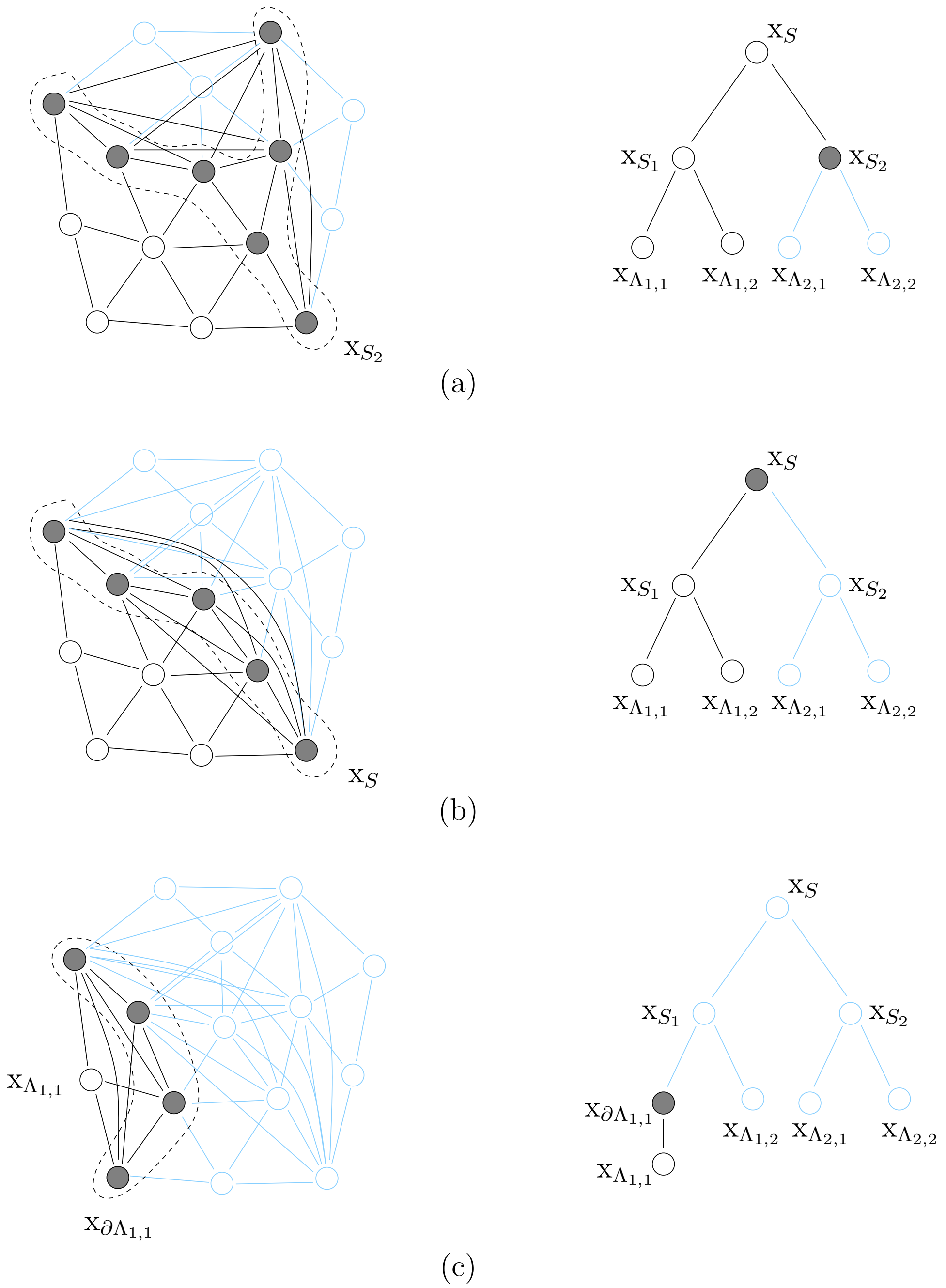

x_{S_2}
x_S
x_{S_1}
x_{S_2}
$\mathrm{x}_{\Lambda_{1,1}}$
$\mathrm{x}_{\Lambda_{1,2}}$
$\mathrm{x}_{\Lambda_{2,1}}$
$\mathrm{x}_{\Lambda_{2,2}}$
(a)
x_S
x_S
x_{S_1}
x_{S_2}
$\mathrm{x}_{\Lambda_{1,1}}$
$\mathrm{x}_{\Lambda_{1,2}}$
$\mathrm{x}_{\Lambda_{2,1}}$
$\mathrm{x}_{\Lambda_{2,2}}$
(b)
$\mathrm{x}_{\Lambda_{1,1}}$
$\mathrm{x}_{\partial\Lambda_{1,1}}$
x_S
x_{S_1}
x_{S_2}
$\mathrm{x}_{\partial\Lambda_{1,1}}$
$\mathrm{x}_{\Lambda_{1,2}}$
$\mathrm{x}_{\Lambda_{2,1}}$
$\mathrm{x}_{\Lambda_{2,2}}$
$\mathrm{x}_{\Lambda_{1,1}}$
(c)

Figure 2-16: Illustration of variable elimination procedure (decimation of the Markov tree) for calculation of $p(x_{\Lambda_{1,1}})$: (a) after elimination of $\Lambda_{2,1}$ and $\Lambda_{2,2}$, (b) after elimination of $S_2 \setminus S$, (c) after elimination of $S \setminus S_1$, $\Lambda_{1,2}$ and $S_1 \setminus \partial\Lambda_{1,1}$. Elimination of $\partial\Lambda_{1,1}$ will then yield $p(x_{\Lambda_{1,1}})$.

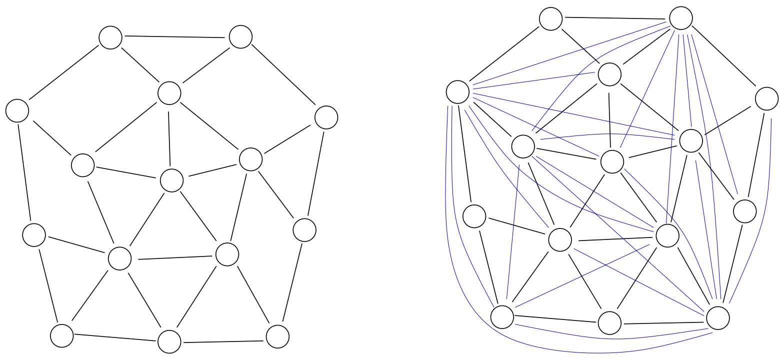


Figure 2-17: Illustration showing triangulated graphical model. Edges are added to the original graph (left) producing a chordal graph (right). These extra "fill" edges reflect the additional degrees of freedom we are allowing in adopting the Markov tree description of the MRF. That is, the family of MRFs on this chordal graph is precisely the family of MRFs which also respect the Markov tree. In a parametric representation of the graphical model, these extra degrees of freedom are required in order to implement consistent belief propagation on the Markov tree.

Hence, elimination of subfield Λ corresponds to the computation

$$\begin{aligned} p(x_{\setminus\Lambda}) &= \int_{\mathcal{X}_\Lambda} p(x_\Gamma) dx_\Lambda && (2.156) \\ &\propto \psi_{\setminus\Lambda}(x_{\setminus\Lambda}) \mu_{\Lambda\to\partial\Lambda}(x_{\partial\Lambda}) && (2.157) \end{aligned}$$

where

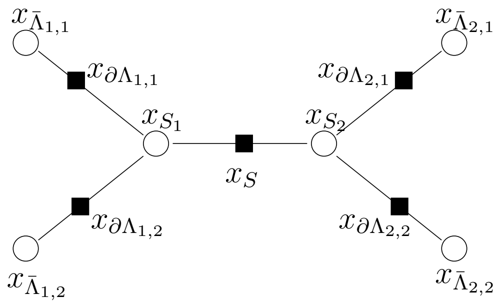
$$\mu_{\Lambda\to\partial\Lambda}(x_{\partial\Lambda}) = \int_{\mathcal{X}_\Lambda} \psi_{\partial\Lambda,\Lambda}(x_{\partial\Lambda}, x_\Lambda)\psi_\Lambda(x_\Lambda) dx_\Lambda. \qquad (2.158)$$

In general, this message will couple all sites in the Markov blanket $\partial\Lambda$ of the eliminated subfield Λ. For instance, in GMRFs this corresponds to the calculation

$$\begin{aligned} \hat{h}_{\partial\Lambda} &= h_{\partial\Lambda} - J_{\partial\Lambda,\Lambda} J_\Lambda^{-1} h_\Lambda && (2.159) \\ \hat{J}_{\partial\Lambda} &= J_{\partial\Lambda} - J_{\partial\Lambda,\Lambda} J_\Lambda^{-1} J_{\Lambda,\partial\Lambda} && (2.160) \end{aligned}$$

which (typically) will cause the interaction matrix $\hat{J}_{\partial\Lambda}$ to become a full (non-sparse) matrix.

Iterating this procedure to solve for the marginal distribution of a single node of the tree, such as illustrated in Figure 2-16, tends to have the affect of "filling out" the MRF by coupling sites within each subfield corresponding to a node or edge of the Markov tree. In this regard, inference on the Markov tree (by either decimation or belief propagation) need not respect the Markov structure of the original MRF. Alternatively, we could instead first introduce the necessary interactions into our model and then perform belief propagation with respect to this augmented model. This is where junction trees enter the picture.



$x_{\bar{\Lambda}_{1,1}}$
$x_{\partial\Lambda_{1,1}}$
x_{S_1}
x_{S_2}
$x_{\partial\Lambda_{2,1}}$
$x_{\bar{\Lambda}_{2,1}}$
x_S
$x_{\partial\Lambda_{1,2}}$
$x_{\partial\Lambda_{2,2}}$
$x_{\bar{\Lambda}_{1,2}}$
$x_{\bar{\Lambda}_{2,2}}$

Figure 2-18: Illustration of one possible junction tree representation $\mathbf{J}_{\mathcal{C}} = (\mathcal{C}, \mathcal{E}_{\mathcal{C}})$ for the MRF depicted in Figure 2-15. The junction tree is based on cliques $\mathcal{C} = \{\bar{\Lambda}_{1,1}, \bar{\Lambda}_{1,2}, S_1, S_2, \bar{\Lambda}_{2,1}, \bar{\Lambda}_{2,2}\}$ (shown as circular nodes) in a triangulated version of the interaction graph $\mathbf{G}_\Gamma$. These are linked by edges $\mathcal{E}_{\mathcal{C}} = \{\{S_1, S_2\}, \{S_1, \bar{\Lambda}_{1,1}\}, \{S_1, \bar{\Lambda}_{1,2}\}, \{S_2, \bar{\Lambda}_{2,1}\}, \{S_2, \bar{\Lambda}_{2,2}\}\}$. Also, each edge corresponds to a separator $\mathcal{S} = \{\partial\Lambda_{1,1}, \partial\Lambda_{1,2}, S, \partial\Lambda_{2,1}, \partial\Lambda_{2,2}\}$ (shown as square nodes "splitting" each edge).

Relation between Markov Trees and Junction Trees. Essentially, the junction tree representation of a MRF is a Markov tree built upon a triangulated representation of the MRF. Such a representation may be constructed as follows. First, the interaction graph $\mathbf{G}_\Gamma$, describing the Markov structure of the MRF, is *triangulated.* This means that we add edges to the graph until the graph becomes *chordal.* That is, every cycle of length four or greater has a chord (an edge not contained in the cycle linking two vertices of the cycle). For instance, the graph shown on the right in Figure 2-17 is chordal. The representation of the graphical model is likewise augmented to include additional interactions (as appropriate for the family under consideration) to accommodate any MRF which respects this triangulated graph. This corresponds to relaxing some of the conditional independencies satisfied by the given model.

Next, a Markov tree is constructed with respect to this triangulated graph. Formally, this is specified by a *junction tree.* This is a clique tree $\mathbf{J}_{\mathcal{C}} = (\mathcal{C}, \mathcal{E}_{\mathcal{C}})$, an acyclic graph based on a collection of cliques $\mathcal{C}$ covering the triangulated graph, which also satisfies the so-called *running intersection property.* That is, for all $\Lambda_1, \Lambda_2 \in \mathcal{C}$, the intersection $\Lambda_1 \cap \Lambda_2$ is contained in every clique Λ along the (unique) path connecting Λ_1 and Λ_2 in the junction tree. It is known that such a junction tree exists for any chordal graph (see Jordan [77]). Then, $\mathbf{J}_{\mathcal{C}}$ describes a Markov tree with states $(\mathrm{x}_\Lambda, \Lambda \in \mathcal{C})$. Hence, we may implement belief propagation with respect to this junction tree representation of the MRF and this may be posed as passing messages between overlapping cliques in the triangulated representation of the MRF. Computation of the marginal distributions $(p(x_\Lambda), \Lambda \in \mathcal{C})$ then provides for subsequent computation of the marginal distributions for each site $(p(x_\gamma), \gamma \in \Gamma)$ and each pair of adjacent sites $(p(x_\gamma, x_\lambda), \{\gamma, \lambda\} \in \mathcal{E}_\Gamma)$ in the original MRF. For instance, the MRF depicted in Figure 2-15 may be represented by the junction tree shown in Figure 2-18 once we have triangulated the interaction graph as shown in Figure 2-17.

To relate this to our earlier discussion, it is helpful to note that the junction tree representation has redundancy built into the state description. For instance, if $\mathbf{G}_\Gamma$ is itself a tree (and hence chordal), then a junction tree $\mathbf{J}_\mathcal{C}$ is constructed by taking the *vertices* of the junction tree $\mathcal{C}$ to be the *edges* of the Markov tree $\mathbf{G}_\Gamma$ and then linking those vertices in the junction tree which are adjacent edges in the Markov tree. Similarly, for each edge of the junction tree $\{\Lambda_1, \Lambda_2\} \in \mathcal{E}_\mathcal{C}$ we may define a separator $S = \Lambda_1 \cap \Lambda_2$ and this is also a separator of the MRF corresponding to some non-leaf node of the Markov tree. More generally, given any description of a MRF $(x_\Gamma, \mathbf{G}_\Gamma)$ as a Markov tree $(\mathrm{x}_\Lambda, \Lambda \in \mathcal{S})$ on $\mathbf{T}_\mathcal{S}$, we may construct a corresponding junction tree representation of that MRF as follows. For every subfield $\Lambda \subset \Gamma$ corresponding either to a node or edge of the Markov tree, we add edges to $\mathbf{G}_\Gamma$ such that Λ is completely connected. Then, the Markov tree $\mathbf{T}_\mathcal{S} = (\mathcal{S}, \mathcal{E}_\mathcal{S})$ determines a junction tree $\mathbf{J}_\mathcal{C} = (\mathcal{C}, \mathcal{E}_\mathcal{C})$ for $\mathbf{G}_\Gamma$ based on the edges of $\mathbf{T}_\mathcal{S}$. In this manner, we may convert any Markov tree representation of the MRF into a junction tree representation.

Due to the redundancy of states in junction trees, some specialization of belief propagation on junction trees is warranted. In general, given a junction tree representation of the MRF we may "split" each edge $\{\Lambda_1, \Lambda_2\}$ into two edges $(\{\Lambda_1, S\}, \{\Lambda_2, S\})$ and a separator node $S = \Lambda_1 \cap \Lambda_2$ while preserving the Markov property (see Figure 2-18). Hence, by Hammersley-Clifford, we may factor the probability distribution with respect to the junction tree as

$$p(x_\Gamma) \propto \frac{\prod_{C \in \mathcal{C}} \phi(x_C)}{\prod_{S \in \mathcal{S}} \phi(x_S)} \tag{2.161}$$

where $\mathcal{S}$ is the set of separators (corresponding to edges) of the junction tree. This is the representation most often considered in the graphical modeling literature. In this representation, belief propagation may again be posed as "refactorization" where passing a message from cliques A to B, via the separator $S = A \cap B$, may be reformulated as

$$\phi_B(x_B) \leftarrow \phi_B(x_B) \times \frac{\int \phi_A(x_A) dx_{A \setminus B}}{\phi_S(x_S)} \tag{2.162}$$

$$\phi_S(x_S) \leftarrow \int \phi_A(x_A) dx_{A \setminus B} \tag{2.163}$$

Performing a two-pass message passing procedure then refactors this representation into the form

$$p(x_\Gamma) = \frac{\prod_{C \in \mathcal{C}} p(x_C)}{\prod_{S \in \mathcal{S}} p(x_S)} \tag{2.164}$$

thus calculating the requisite marginal distributions on the junction tree. This is the usual form of belief propagation discussed for junction trees (Dawid [40], Lauritzen [88], Cowell [33, 32], Jordan [77]).

Hence, in principle at least, belief propagation for Markov trees may be extended (in various ways) to perform exact inference in MRFs. Yet, the need to triangulate the graphical model in order to implement consistent belief propagation in MRFs exposes the potential intractability of the method. Mainly, when the graphical structure of the

MRF is such that any Markov tree (junction tree) representation must have some tree nodes corresponding to large subfields (cliques) in the MRF, then the computational complexity of this approach may become computationally infeasible. In such cases, there is need to develop tractable approximate inference procedures. We discuss some work along these lines in the remainder of the chapter. This is also the intent of the RCM approach discussed in Chapter 4.

2.3.3 Multiscale Models

In this subsection we review the basic multiscale modeling approach introduced in [14], [9] and [8]; and further developed in [30], [91], [92], [93], [72], and [55]. A recent survey article by Willsky [132] provides a unified perspective of this field. The basic paradigm here is to provide a tractable model for a multi-dimensional signal by providing a sequence of "coarse-scale" descriptions of the signal and modeling this multiscale description of the signal by a tree-structured MRF (see Figure 2-19). The levels of the tree are thought of as corresponding to coarse-scale representations of the signal of interest at various levels of resolution. The signal of interest corresponds to the fine-scale process which is represented by the states at the leaf nodes of the tree. The state-space associated to the internal levels of the tree then provide progressively coarser descriptions of the process.

MAR Models. The majority of this literature has been aimed at linear problems where the process is Gaussian or where we are constrained to perform linear least-squares estimation of some arbitrary process based on just the first and second-order statistics of that process. These linear systems may be formulated as *multiscale autoregressive* (MAR) models. The state at the root node γ_0 is then modeled as normally distributed with specified mean and covariance $\mathrm{x}_{\gamma_0} \sim \mathcal{N}(\hat{x}_{\gamma_0}, P_{\gamma_0})$. The statistics of the rest of the process are then determined by linear dynamics recursing down the tree. For each node γ beneath the root node γ_0, the state of that node x_γ is modeled as being a linear functional of the state $\mathrm{x}_{\pi(\gamma)}$ of the parent node $\pi(\gamma)$ (the node immediately "above" γ at the next coarser scale in the tree) but corrupted by additive Gaussian driving noise $\mathrm{w}_\gamma \sim \mathcal{N}(0, Q_\gamma)$.

$$\mathrm{x}_\gamma = A_\gamma \mathrm{x}_{\pi(\gamma)} + \mathrm{w}_\gamma \tag{2.165}$$

The various driving noise terms $\{\mathrm{w}_\gamma\}$ are taken as mutually independent and also independent of the state at the root node x_{γ_0}. This defines a Markov tree $(x_\Gamma, \mathbf{T}_\Gamma)$ with Gaussian probability distribution which factors on the tree as

$$p(x_\Gamma) = p(x_\gamma) \prod_{\gamma \in \Gamma \setminus \gamma_0} p(x_\gamma | x_{\pi(\gamma)}) \tag{2.166}$$

in terms of the Gaussian distribution at the root node

$$p(x_{\gamma_0}) = \frac{1}{\sqrt{\det 2\pi P_{\gamma_0}}} \exp\{-\frac{1}{2}(x_{\gamma_0} - \hat{x}_{\gamma_0})' P_{\gamma_0}^{-1} (x_{\gamma_0} - \hat{x}_{\gamma_0})\} \tag{2.167}$$

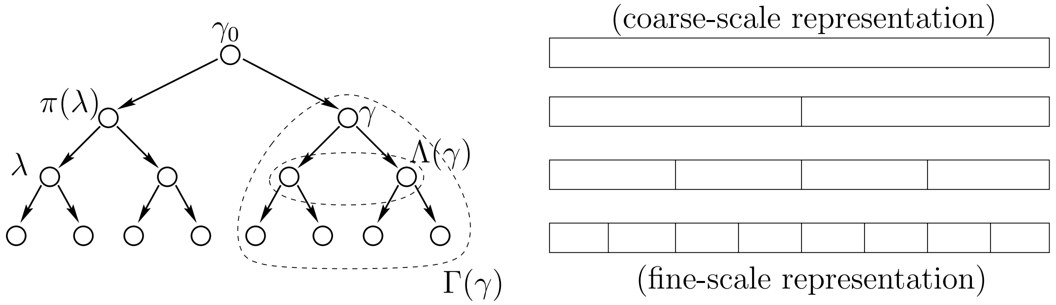


Figure 2-19: Depiction of multiscale model for a 1-D signal. A causal Markov tree (left) provides a scale recursive coarse-to-fine causal model of the signal (right).

and conditional Gaussian distributions on the edges of the tree

$$p(x_\gamma|x_{\pi(\gamma)}) = \frac{1}{\sqrt{\det 2\pi Q_\gamma}} \exp\{-\frac{1}{2}(x_\gamma - A_\gamma x_{\pi(\gamma)})' Q_\gamma^{-1}(x_\gamma - A_\gamma x_{\pi(\gamma)})\} \tag{2.168}$$

This may be viewed as a graphical model with compatibility functions $\psi(x_{\gamma_0}) = p(x_{\gamma_0})$ and $(\psi(x_\gamma, x_{\pi(\gamma)}) = p(x_\gamma|x_{\pi(\gamma)}), \gamma \in \Gamma \setminus \gamma_0)$. This is an example of a *causal* graphical model.

We may also incorporate measurements into this graphical model. Consider measurements in the form of linear functionals of the state at each node γ corrupted by additive Gaussian measurement noise $\mathrm{v}_\gamma \sim \mathcal{N}(0, R_\gamma)$.

$$\mathrm{y}_\gamma = C_\gamma \mathrm{x}_\gamma + \mathrm{v}_\gamma \tag{2.169}$$

The various measurement noise terms $\{\mathrm{v}_\gamma\}$ are taken to be mutually independent and also independent of both the driving noise and the state at the root node. The conditional probability distribution $p(x_\Gamma|y_\Gamma) \propto p(y_\Gamma|x_\Gamma)p(x_\Gamma)$ also factors on the tree as

$$p(x_\Gamma|y_\Gamma) \propto p(x_{\gamma_0}) \prod_{\gamma\in\Gamma} p(y_\gamma|x_\gamma) \prod_{\gamma\in\Gamma\setminus\gamma_0} p(x_\gamma|x_{\pi(\gamma)}) \tag{2.170}$$

where

$$p(y_\gamma|x_\gamma) = \frac{1}{\sqrt{\det 2\pi R_\gamma}} \exp\{-\frac{1}{2}(y_\gamma - C_\gamma x_\gamma)' R_\gamma^{-1}(y_\gamma - C_\gamma x_\gamma)\} \tag{2.171}$$

This may be viewed as a graphical model with compatibility functions $\psi(x_{\gamma_0}) = p(y_{\gamma_0}|x_{\gamma_0})p(x_{\gamma_0})$, $(\psi(x_\gamma) = p(y_\gamma|x_\gamma), \gamma \in \Gamma \setminus \gamma_0)$, and $(\psi(x_\gamma, x_{\pi(\gamma)}) = p(x_\gamma|x_{\pi(\gamma)}), \gamma \in \Gamma \setminus \gamma_0)$.

We may calculate the conditional distributions at each node of the tree by two-pass belief propagation. With respect to this causal specification, this may be implemented by a two-pass recursive filtering and smoothing algorithm which generalizes the Kalman filter and Rauch-Tung-Striebel Smoother for 1D time-series models [78, 79, 110, 58]. The upward pass of this procedure calculates likelihood functions

$\hat{\psi}(x_\gamma) = p(y_{\Gamma(\gamma)}|x_\gamma)$ where $\Gamma(\gamma)$ denotes the set of all nodes in the subtree rooted at γ (Figure 2-19). The downward pass propagates messages $\mu_{\pi(\gamma)\to\gamma}(x_\gamma) = p(x_\gamma|y_{\setminus\Gamma(\gamma)})$, the conditional distribution of x_γ given all observations not on the subtree rooted at node γ. Fusing this downward message with $\hat{\psi}(x_\gamma)$ gives the desired conditional distribution $p(x_\gamma|y_\Gamma)$ conditioned on all measurements.

Relation to Information Form of BP. We now relate this causal inference approach to the corresponding information form. The probability distribution of x_Γ is Gaussian and may be put into the information form $p(x_\Gamma) \propto \exp\{-\frac{1}{2}x_\Gamma J x_\Gamma + h' x_\Gamma\}$ where the information parameters (h, J) are calculated from the parameters of the MAR model as follows. The parameters h are given by $h_{\gamma_0} = P_{\gamma_0}^{-1}\hat{x}_{\gamma_0}$ at the root node and are zero elsewhere. The diagonal blocks of J are given by

$$J_{\gamma,\gamma} = Q_\gamma^{-1} + \sum_{\lambda\in\Lambda(\gamma)} A'_\lambda Q_\lambda^{-1} A_\lambda \tag{2.172}$$

where $Q_{\gamma_0} = P_{\gamma_0}$ and $\Lambda(\gamma)$ are the children of γ (Figure 2-19). The off-diagonal blocks $J_{\gamma,\pi(\gamma)} = J'_{\pi(\gamma),\gamma}$ are given by

$$J_{\gamma,\pi(\gamma)} = -2A'_\gamma Q_\gamma^{-1} \tag{2.173}$$

for all $\gamma \neq \gamma_0$. The remaining entries of J are zero so as to respect the conditional independencies dictated by $\mathbf{T}_\Gamma$. Thus, it is straight-forward to convert from the causal model to the (directionless) information form. However, recovering the causal model given the information model is not so simple. Essentially, this would require that we infer the marginal distributions under the information model (a global calculation) in order to calculate $p(x_{\gamma_0})$ at the root and $p(x_\gamma|x_{\pi(\gamma)}) = p(x_\gamma, x_{\pi(\gamma)})/p(x_{\pi(\gamma)})$ at every other node.

The conditional distribution $p(x_\Gamma|y_\Gamma)$ may also be expressed in the information form $p(x_\Gamma|y_\gamma) \propto \exp\{-\frac{1}{2}x'_\Gamma \hat{J} x_\Gamma + \hat{h}'(y_\Gamma)x_\Gamma\}$ where

$$\begin{aligned} \hat{h}_\gamma(y_\Gamma) &= h_\gamma + C'_\gamma R_\gamma^{-1} y_\gamma & (2.174)\\ \hat{J}_{\gamma,\gamma} &= J_{\gamma,\gamma} + C'_\gamma R_\gamma^{-1} C_\gamma & (2.175)\end{aligned}$$

and $\hat{J}_{\gamma,\lambda} = J_{\gamma,\lambda}$ for all $\gamma \neq \lambda$. In this manner, we may incorporate measurements into the information representation. This gives a graphical model with compatibility functions $(\psi(x_\gamma) = \exp\{-\frac{1}{2}x'_\gamma \hat{J}_\gamma x_\gamma + \hat{h}'_\gamma x_\gamma\}, \gamma \in \Gamma)$ and $(\psi(x_\gamma, x_\lambda) = \exp\{-x'_\gamma \hat{J}_{\gamma,\lambda} x_\lambda\}, \{\gamma,\lambda\} \in \mathcal{E}_\Gamma)$. These have the interpretation

$$\begin{aligned} \psi(x_\gamma) &\propto p(y_\gamma|x_\gamma, \mathrm{x}_{\partial\gamma} = 0) & (2.176)\\ \psi(x_\gamma, x_\lambda) &\propto \frac{p(y_{\{\gamma,\lambda\}}|x_{\{\gamma,\lambda\}}, \mathrm{x}_{\partial\{\gamma,\lambda\}} = 0)}{p(y_\gamma|x_\gamma, \mathrm{x}_{\partial\gamma} = 0)p(y_\lambda|x_\lambda, \mathrm{x}_{\partial\lambda} = 0)} & (2.177)\end{aligned}$$

which differs from the interpretation of compatibility functions in the causal model. Consequently, the representation, calculation and interpretation of messages and intermediate beliefs arising in belief propagation likewise differ. For instance, the upward pass of belief propagation (in the information form) calculates $\hat{\psi}(x_\gamma) \propto p(y_{\Gamma(\gamma)}|x_\gamma, \mathrm{x}_{\pi(\gamma)} = 0)$. The downward messages may be interpreted as $\mu_{\pi(\gamma)\to\gamma}(x_\gamma) \propto p(y_{\setminus\Gamma(\gamma)}|x_\gamma, \mathrm{x}_{\Lambda(\gamma)} = 0)$. Yet, fusing these again gives the conditional marginals $p(x_\gamma|y_\Gamma)$.

Hence, while these two approaches closely parallel one another and ultimately give the same results, these are nevertheless distinct versions of belief propagation where the messages and intermediate beliefs computed under the two methods cannot be related until *after* we have gathered all available evidence (messages) at a given node.

"Cutset" Construction of Markov Trees. For the purpose of this discussion, the main aspect of the multiscale modeling approach (relevant to RCM) is the manner in which exact tree models may be constructed for multidimensional GMRFs. This technique, developed by Luettgen for image processing applications (Luettgen [91], Luettgen et al [92], Luettgen and Willsky [93]), entails recursively partitioning the Markov random field into quadrants forming a quad-tree structured hierarchical decomposition of the field. This approach is illustrated in Figure 2-20. By defining states at the coarse scales of the Markov tree as the joint state on separator sets between partitions (given by the surfaces of those adjacent partitions) one is able to then construct a causal tree-structured realization of the underlying MRF which is regarded as residing at the finest scale of the tree model. This is closely related to our previous discussion, in Section 2.3.1, where we also decomposed the MRF by recursive specification of separators. Yet, Luettgen's recursive partitioning approach has the effect of assuring that each site of the MRF is reproduced at a leaf node in the Markov tree representation (this is important for the state-reduction method to be discussed). This partitioning approach is also used in our RCM approach and is discussed further in Chapter 4. The computation of the associated MAR model parameters given specification of an arbitrary MRF in terms of local interactions is a nontrivial inference problem. This may be posed as belief propagation in the information representation of the Markov tree to yield a causal representation of the Markov tree.

The main disadvantage of employing this type of exact realization approach is that it yields large state dimensions at coarse scales of the tree. Then, performing belief propagation with respect to this Markov tree may become computationally infeasible. For instance, given a $N = W \times W$ 2D nearest-neighbor GMRF, the state dimension at the root node is order W and the associated complexity of exact inference on this tree model is order W^3 or $N^{3/2}$ which does not scale linearly with the number of sites N of the field.

Reduced-State Cutset Approximations. In order to provide a tractable inference approach, much work has been performed to develop approximate realization techniques which attempt to construct good approximations to a given fine-scale process by carefully designing the state-space of coarse-scale nodes to do the best job possible of decorrelating disjoint partitions subject to dimension constraints on the state space of these hidden nodes. The general idea is illustrated in Figure 2-21.

This has included the work of Irving and Frakt concerning the realization of Gaus-

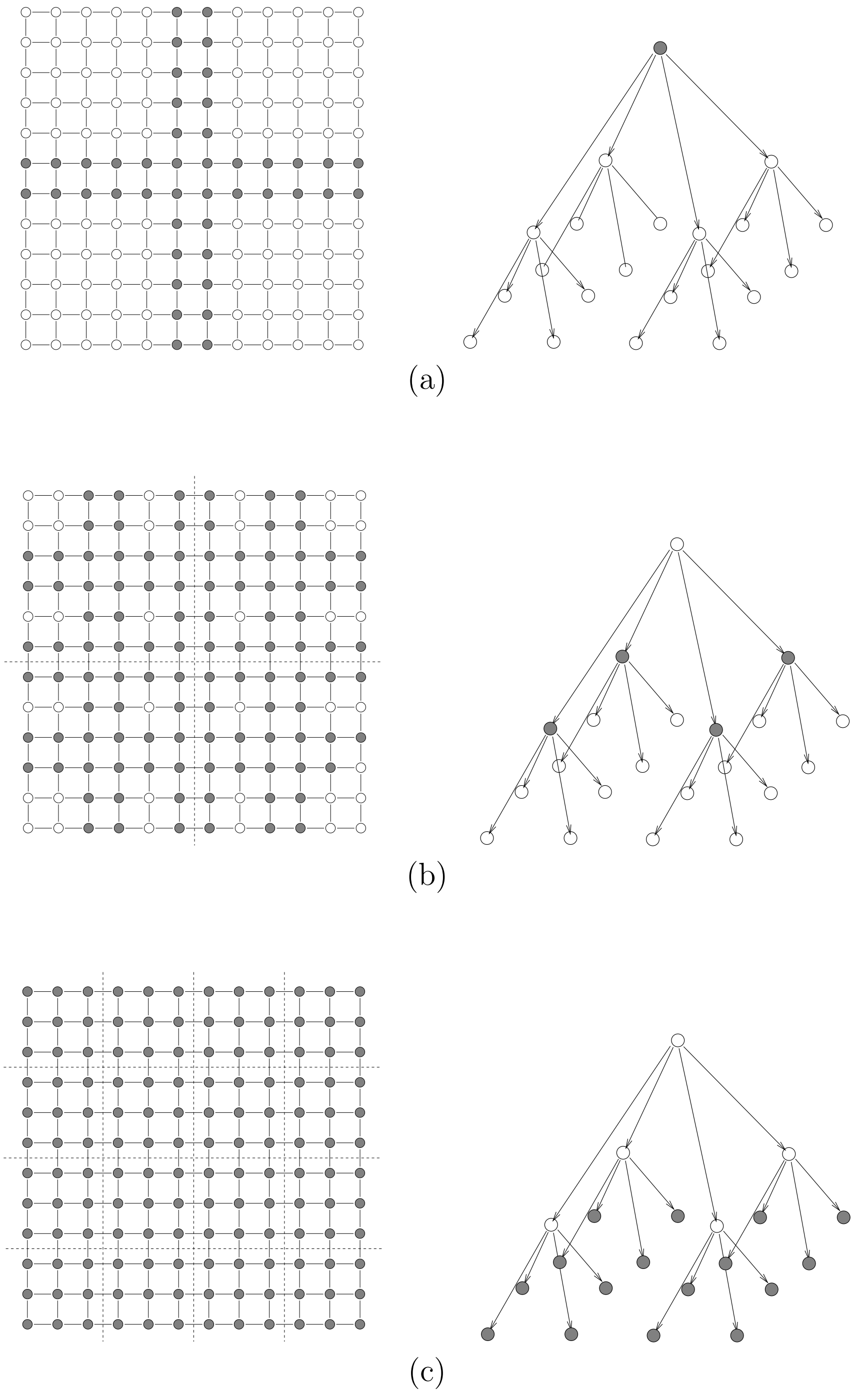


Figure 2-20: Illustration of multiscale quadtree model for a 12 × 12 MRF. (a) The root node specifies the state on a "cross-hair" cutset of the MRF. (b) The four nodes at the second level of the quadtree correspond to four partitions of the MRF. Each specifies the state of the surface of the corresponding partition as well as on a cutset within the partition. (c) The sixteen leaf nodes each correspond to a 3 × 3 subfield.

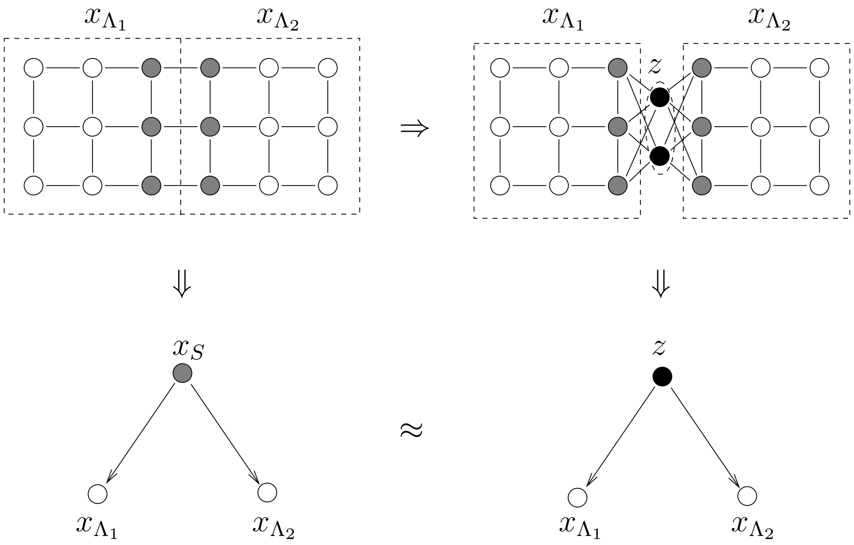


Figure 2-21: Illustration of idea underlying state-reduction approximations.

sian processes given specification of the covariance structure of the fine-scale process. Irving employs methods based on information theory and canonical correlations (Irving [72], Irving and Willsky [73]) while Frakt employs the estimation theoretic notion of predictive efficiency (Frakt [55], Frakt and Willsky [57]). Frakt has also extended his techniques to allow for partial covariance specifications based on the maximum entropy principle (Frakt [55], Frakt et al [56]). Currently, Tucker [128] is working to extend these techniques to the data-driven case.

One problem which has become apparent in image-processing applications in regards to such state-reduced approximate models is the so-called phenomena of "blocky artifacts." The fact remains that a state-reduced multiscale model does not store sufficient information at coarse scale states to fully render the subordinate processes conditionally independent. This leads to degradation in the quality of estimates along those quadrantal boundaries associated with the quad-tree decomposition of the field. These irregularities can become more apparent at coarser scales as the inadequate state-dimension becomes more problematic.

The occurrence of these artifacts is a primary consideration motivating the RCM approach. An important distinction is that RCM performs approximate inference with respect to an exact model instead of performing exact inference with respect to an approximate model. This is achieved by combining the inference and modeling procedures. As we shall see, RCM does not impose restrictions on the state-dimension of these "cutset" states but rather imposes restrictions upon the complexity of the inferred information models associated with those decorrelating states. This is seemingly a less heavy-handed type of approximation and hopefully will avoid the problem of blocky artifacts perhaps leading to an overall reduction in estimation error at comparable levels of computation.

Domain Decomposition. Also arising in the multiscale modeling literature are certain domain decomposition approaches (Taylor and Willsky [127], Taylor [126], Daniel [36]) for estimation and filtering of 2D linear systems which are perhaps more closely related to RCM than those techniques based on reduced-state MAR models. These methods are closely related to the nested dissection algorithm [61] and other domain decomposition techniques employed in the solution of sparse linear systems such as arise in the numerical solution of multidimensional partial differential equations. The methods of both Taylor and Daniel begin by recursively decomposing the field precisely as in Luettgen's approach (Figure 2-20). We focus on Taylor's approach which more closely resembles RCM in that it takes a model based view of the local processing performed within each subfield.

Taylor considers two-dimensional linear systems specified as a *nearest neighbor model* (NNM). This describes a collection of random variables $(\mathrm{x}_\Gamma, \mathrm{y}_\Gamma)$ arranged on a 2D square grid $\Gamma = \{(i,j) | i = 1,\ldots,N, j = 1,\ldots,M)\}$ with interactions given by a system of noisy constraints and measurements

$$\begin{aligned} \mathrm{x}_{i,j} &= N_{(i,j)}\mathrm{x}_{i,j+1} + S_{(i,j)}\mathrm{x}_{i,j-1} + E_{(i,j)}\mathrm{x}_{i+1,j} + W_{(i,j)}\mathrm{x}_{i-1,j} + \mathrm{w}_{i,j} && (2.178)\\ \mathrm{y}_{i,j} &= C_{(i,j)}\mathrm{x}_{i,j} + \mathrm{v}_{i,j} && (2.179) \end{aligned}$$

together with some appropriate set of boundary conditions[23] where $\mathrm{v}_{i,j} \sim \mathcal{N}(0, R_{(i,j)})$ is measurement noise, independent from site to site, and where the driving noise $w \sim \mathcal{N}(0, Q)$ has a sparse covariance matrix designed so as to insure that the process is Markov with respect to the graph defined by the neighborhoods $\partial(i,j) = \{(i,j+1),(i,j-1),(i+1,j),(i-1,j)\}$ (Woods [133, 134]). Consider the problem of computing the conditional marginal distributions $p(x_\gamma | y_\Gamma)$ at each site γ given the measurements y_Γ. This may be posed in the information form as we have discussed previously. Taylor considers a more general approach calculating the minimum-norm least-squares estimate $\hat{x}_\Gamma(y_\Gamma)$ minimizing

$$\begin{aligned} l(x_\Gamma) &= (y_\Gamma - Cx_\Gamma)' R^{-1}(y_\Gamma - Cx_\Gamma) + x_\Gamma'(I-K)'Q^{-1}(I-K)x_\Gamma && (2.180)\\ &= x_\Gamma(C'R^{-1}C + (I-K)'Q^{-1}(I-K))x_\Gamma + \mathrm{const} && (2.181) \end{aligned}$$

This is interpreted as maximum-likelihood estimation, since $l(x_\Gamma) = -2\log p(y_\Gamma; x_\Gamma)$ viewing x_Γ as (non-random) parameters of the model. This more general formulation is well-posed even when neither the covariance nor the inverse covariance of x_Γ is well-defined. We give a high-level description of Taylor's approach.

The basic strategy again involves a two-pass procedure following the tree-structured decomposition of the field as shown in Figure 2-20. The upward pass begins within the smallest squares subfields corresponding to leaves of the tree. Within each subfield, a causal Markov chain description of the interior of the field is adopted where the states of the Markov chain correspond to a sequence of square "rings" beginning at the center of the subfield and advancing radially outwards towards the surface of the subfield. This Markov chain is implicitly conditioned on zero state outside of that

[23] Along the edges of the grid, we may either (i) neglect contributions from sites outside the grid, (ii) introduce periodic boundary conditions, or (iii) explicitly specify state values along the boundary.

subfield. Taylor constructs this Markov chain from just the local NNM parameters defined within that subfield. His approach specifies the dynamics of this Markov chain in the form of a *separable two-point boundary value descriptor system* which generalizes the autoregressive type dynamics we discussed previously for MAR models. In the upward pass, an outwards filtering step is performed with respect to this Markov chain producing a description of the statistics at the surface of the field conditioned on all data within that subfield (and also conditioned on zero state outside of that subfield). The upward pass then proceeds up the tree by merging adjacent subfields and performing a similar outwards filtering step along the common boundary of the two subfields (which is a separator of the conditional subfield assuming zero state outside of the merged subfield). A complimentary downward pass then reverses this processing by performing an inwards smoothing operation with respect to each of these Markov chains recursing down the tree. Once the leaves of the tree are reached, this yields the desired state estimates and corresponding error covariances at each site of the field.

This processing very closely parallels a corresponding variable elimination approach within the information representation of the field (when this is well-defined). In particular, Taylor's interpretation of the outwards and inwards message passing as respectively calculating the zero-input response and zero-state response of his (linear) maximum-likelihood estimator corresponds precisely to the interpretation of messages in the information form of belief propagation as being conditioned upon zero boundary conditions either outside or inside each subfield.

As Taylor discusses, the complexity of his approach is dominated by the cost of the filtering/smoothing calculations associated with the outer boundary of the largest domain. This is fundamentally related to the problem of "fill" in the corresponding variable elimination approach which is also most costly at the largest separator of the field. Taylor considers suboptimal filtering and smoothing procedures to adjust the computational complexity of his method by truncating either the covariance or the inverse covariance matrices propagated in his method. This models the outer-most "ring" in his radial filtering approach as either a 1D periodic autoregressive or moving average model going around the boundary. The approximation is imposed by simply setting unwanted elements of the matrix, outside some specified bandwidth w, to zero so as to only retain interactions between the w nearest neighbors to a given site in the boundary. This allows a lower-order Markov chain to be constructed at the next level up in the tree thus reducing the computational complexity of his method.

Daniel's [36] approach has similar structure as in Taylors method, but is posed as performing a partial LU decomposition of the inverse covariance J in a tree structured manner which exploits the sparsity of J (precisely as in nested dissection). This then supports solution of $J\hat{x} = h$ by a two-pass estimation procedure which first solves $Ly = h$ and then solves $U\hat{x} = y$. His method also introduces a thinning operation which yields a more tractable albeit approximate decomposition of J. Yet, the interpretation of this thinning step is less clear, from a modeling point of view, than in Taylor's method.

As we will see in Chapter 4, RCM is also a domain decomposition approach which aims to provide tractable inference by reducing the complexity of the intermediate cutset models arising in domain decomposition. RCM differs from Taylor's approach in that all processing is done in the information representation of the GMRF (local conditional subfields are never converted into a causal representation) and we employ the variable elimination approach to inference throughout. RCM also adopts thinned Markov models of the surfaces of subfields corresponding to Taylor's periodic autoregressive models going around the boundary of each subfield. An important distinction, however, is that RCM employs the machinery of information geometry to give a principled approach for selecting such approximations (both the Markov structure and the parameters of the model). Also, we will give iterative procedures to refine these approximations. Finally, the general RCM framework should prove applicable for far more general families of MRFs than just GMRFs.

Chapter 3

Model Thinning

This chapter focuses on the fundamental model thinning problem. The techniques developed here are derived from the information geometric perspective for model selection among exponential families of Gibbs random fields as discussed in the previous chapter. These methods lie at the heart of the recursive cavity modeling approach for tractable inference of graphical models to be discussed in Chapter 4.

The model thinning problem (introduced in Section 2.2.3) is now summarized. We are given a graphical model specified by a set of interaction potentials $\phi = (\phi_\Lambda(x_\Lambda; \theta_\Lambda) = \theta_\Lambda \cdot t_\Lambda(x_\Lambda), \Lambda \in \mathcal{H}^\phi_\Lambda)$ relative to a hypergraph $\mathbf{H}^\phi_\Gamma = (\Gamma, \mathcal{H}^\phi_\Gamma)$. This describes a MRF $(x_\Gamma, \mathbf{G}^\phi_\Gamma)$ with probability distribution $p(x_\Gamma) \propto \exp\{\sum \theta_\Lambda \cdot t_\Lambda(x_\Lambda)\}$ and with Markov structure determined by the adjacency graph $\mathbf{G}^\phi_\Gamma = \text{adj } \mathbf{H}^\phi_\Gamma$. We also assume that the graphical model is tractable by exact inference methods such that it is feasible to calculate the moment parameters $\eta = E_\theta\{t(x_\Gamma)\}$. For instance, the graphical model is either sufficiently small such that simple brute-force inference methods apply or has sufficiently low tree-width such that efficient recursive inference methods apply (for instance, by decomposing the graphical model as a Markov chain or tree with low state dimensions).[1]

Our objective is to determine a *thinned* graphical model which provides a more compact yet faithful approximation of the original. We pose this as thinning of an exponential family model by neglecting some of the statistics $t(x_\Gamma)$ (forcing the associated exponential parameters θ to zero). This may be seen as pruning hyperedges from $\mathbf{H}^\phi_\Gamma$ or as pruning edges from $\mathbf{G}^\phi_\Gamma$ so as to give a thinned graphical model. The model thinning problem then has two components: (i) selection of which statistics to omit from the model (thereby selecting an embedded exponential family and corresponding Markov structure), and (ii) adjustment of the remaining (non-zero) exponential parameters. Subject to the choice of (i), the optimal (maximum-likelihood) choice of exponential parameters is given by m-projection of the given graphical model to the selected exponential family of graphical models.

The model thinning procedures developed here are based on an information criterion inspired by the Akaike information criterion (AIC) for model selection from data [1, 2]. This criterion balances the competing objectives of model fidelity (measured by

[1] In Chapter 4, where we consider intractable models, these assumptions are met as we then only consider model thinning for tractable subfields.

KL-divergence) and model complexity (measured by model order). We discuss this further in Section 3.1. Over a given exponential family, minimization of this criterion reduces to the m-projection problem. In Section 3.2, we discuss our approach to m-projection based on a novel adaptation of the iterative scaling technique inspired by the cluster variation method for approximate inference (Kappen and Wiegerinck [80]) and also Minka's method of m-projection to trees (Minka [96]). In Section 3.3, we then specify model thinning procedures which incrementally prune statistics from the exponential family by inductive m-projection (such as in Proposition 16). We say that this approach is *inductive* because later decisions as to which statistics ought to be neglected are based on observing the effect of earlier m-projections.

3.1 Information Criterion

This section introduces the information criterion we employ for model thinning. Let μ denote a given graphical model we wish to thin and let ν denote a thinned version of μ which we wish to assess as a possible substitute for μ. We wish to select ν to provide a compact yet faithful approximation of μ. We require that the candidate ν remain faithful to the original μ in the sense of having low KL-divergence $D(\mu\|\nu) = E_\mu \log \mu(\mathrm{x})/\nu(\mathrm{x})$ where $\mu(x)$ and $\nu(x)$ denote the probability distributions of x under the two models. We require that the candidate ν is compact in the sense of having low model-order $K(\nu)$, i.e. the number of independent model parameters. These competing objectives are balanced against one another by the cost function

$$V(\mu;\nu) = D(\mu\|\nu) - \delta(K(\mu) - K(\nu)) \tag{3.1}$$

where $\delta > 0$ is an adjustable parameter controlling how strongly we favor compactness over fidelity. The first term $D(\mu\|\nu) \geq 0$ measures the *modeling error* incurred by replacing μ by ν. The second term reduces this modeling error by an amount proportional to the *order reduction* $\Delta K = K(\mu) - K(\nu)$ resulting from the substitution. This is the number of statistics in the model μ which are neglected in the model ν by forcing the corresponding parameters to zero. The parameter $\delta > 0$ scales the order reduction relative to the modeling error and hence indicates the amount of modeling error we are prepared to accept per removed model parameter. This gives the *adjusted modeling error* which we wish to minimize by our choice of ν. Larger values of δ favor more compact models allowing substitutions which incur greater modeling error. Note also that $V(\mu;\mu) = 0$ so that model thinning is recommended only if $V(\mu;\nu) < 0$ indicating a reduction of the adjusted modeling error favoring the substitution of ν for μ.

Inductive Decomposition. We also remark that the adjusted modeling error $V(\mu;\nu)$ inherits the "Pythagorean" decomposition of the KL-divergence it is based on. Consider a sequence of embedded exponential families $\mathcal{H}_0 = \mathcal{F} \supset \mathcal{H}_1 \supset \ldots \supset \mathcal{H}_K$ with $\mu \in \mathcal{F}$ and set $\hat{\mu}^{(k)} = \arg\min_{\nu \in \mathcal{H}_k} D(\mu\|\nu)$. Then, the adjusted modeling error decomposes as

$$V(\mu;\hat{\mu}^{(K)}) = \sum_{k=1}^{K} V(\hat{\mu}^{(k-1)};\hat{\mu}^{(k)}) \tag{3.2}$$

where the incremental error adjustments may be evaluated as

$$V(\hat{\mu}^{(k)};\hat{\mu}^{(k+1)}) = (h(\hat{\mu}^{(k+1)}) - h(\hat{\mu}^{(k)})) - \delta(K(\hat{\mu}^{(k)}) - K(\hat{\mu}^{(k+1)})) \tag{3.3}$$

This suggests that we approach model thinning by incremental m-projections to lower-order exponential families. Each m-projection attempts to select the "next" embedded family so as to reduce V. The inductive thinning continues so long as we can identify further projections to lower-order families with $V < 0$. Thus, δ gives an upper bound threshold on the allowed information loss (under m-projection) per removed model parameter.

In the next section, we develop our technique for m-projection to a selected embedded exponential family. In Section 3.3, we specify an inductive procedure for selection of embedded families employing m-projection as a subroutine.

3.2 Information Projection

In this section we develop our m-projection procedure. Here, we are given a graphical model $\mu \in \mathcal{F}$ where $\mathcal{F}$ corresponds to the exponential family of (normalizable) Gibbs distributions with interaction potentials $(\phi_\Lambda(x_\Lambda;\theta_\Lambda) = \theta_\Lambda \cdot t_\Lambda(x_\Lambda), \forall \Lambda \in \mathcal{H}_\Gamma^\phi)$. We specify an embedded exponential family $\mathcal{H} \subseteq \mathcal{F}$ based on a reduced set of statistics $t_\mathcal{H}(x_\Gamma) \subseteq t(x_\Gamma)$ with corresponding exponential parameters $\theta_\mathcal{H}$ and moment parameters $\eta_\mathcal{H}$. Those statistics of the family $\mathcal{F}$ being omitted are denoted $t'_\mathcal{H}(x)$ with exponential and moment parameters denoted similarly. The embedded exponential family $\mathcal{H}$ is an e-flat submanifold of $\mathcal{F}$ specified by:

$$\mathcal{H} = \{\nu \in \mathcal{F} \mid \theta'_\mathcal{H}(\nu) = 0\} \tag{3.4}$$

We then wish to evaluate the m-projection of μ to this e-flat submanifold.

$$\hat{\mu} = \arg\min_{\nu\in\mathcal{H}} D(\mu\|\nu) \tag{3.5}$$

By Proposition 12, this minimization problem has a unique minimizer $\nu \in \mathcal{H}$ characterized by the necessary and sufficient condition that $\eta_\mathcal{H}(\nu) = \eta_\mathcal{H}(\mu)$. We denote these desired moments by $\eta^*_\mathcal{H} \equiv \eta_\mathcal{H}(\mu)$ which are obtained by inference of the given model μ. This m-projection is dual to a corresponding class of e-projections. We define the m-flat submanifold $\mathcal{H}'$ as the family of models in $\mathcal{F}$ satisfying the moment constraints $\eta^*_\mathcal{H}$.

$$\mathcal{H}' = \{\mu' \in \mathcal{F} \mid \eta_\mathcal{H}(\mu') = \eta^*_\mathcal{H}\} \tag{3.6}$$

Then, by Proposition 14, for any $\nu \in \mathcal{H}$ the e-projection of ν to the m-flat submanifold $\mathcal{H}'$,

$$\hat{\nu} = \arg \min_{\mu \in \mathcal{H}'} D(\mu \| \nu), \tag{3.7}$$

is also the m-projection of (any) $\mu \in \mathcal{H}'$ to the e-flat submanifold $\mathcal{H}$, i.e. $\hat{\nu} = \hat{\mu}$. This is due to the fact that the e-flat submanifold $\mathcal{H}$ and the m-flat submanifold $\mathcal{H}'$ are $\mathcal{I}$-orthogonal submanifolds intersecting at $\hat{\nu} = \hat{\mu}$. This duality is useful as we may then apply iterative scaling techniques (for e-projection) to calculate the desired m-projection. The fundamental idea is illustrated in Figure 3-1 and detailed in the following discussion.

3.2.1 Moment Matching

We now describe how iterative scaling may be performed to calculate the m-projection of μ to the e-flat submanifold $\mathcal{H}$, a hyperplane in exponential coordinates specified by $\theta'_{\mathcal{H}} = 0$ (see top panel of Figure 3-1). Here, we specify the approach using the iterative proportional fitting (IPF) procedure, discussed previously in Section 2.2.4, to perform moment matching.[2]

First, inference is performed with respect to the model μ to calculate the moments $\eta^*_{\mathcal{H}} = \eta_{\mathcal{H}}(\mu)$. This calculation is performed using exact recursive inference techniques such as described in Section 2.3. This determines the m-flat submanifold $\mathcal{H}'$ indicated by a straight solid line in moment coordinates in our illustration (see lower panel of Figure 3-1). Once these moments $\eta^*_{\mathcal{H}}$ are known, the original model μ is no longer required and may be discarded.[3] In principle, this inference calculation actually gives the m-projection $\hat{\mu}$ which is uniquely specified (in $\mathcal{H}$) by the moment coordinates $\eta_{\mathcal{H}}(\hat{\mu}) = \eta^*_{\mathcal{H}}$. However, what we desire is the explicit computation of the corresponding exponential coordinates $\theta^*_{\mathcal{H}} \equiv \theta_{\mathcal{H}}(\hat{\mu})$.

To recover $\theta^*_{\mathcal{H}}$ from $\eta^*_{\mathcal{H}}$, we may employ iterative proportional fitting within the embedded exponential family $\mathcal{H}$. To do so, we must first specify the exponential coordinates of some initial guess $\hat{\nu}^{(0)} \in \mathcal{H}$. Ideally, this should be chosen so that $\hat{\nu}^{(0)}$ is near the sought after m-projection $\hat{\mu}$. A variety of initialization methods might be recommended for different contexts.[4] However, in the context of model thinning, the submanifold $\mathcal{H}$ is itself presumably chosen to be near the given model μ so that the parameters $\theta'_{\mathcal{H}}(\mu)$ are small. In this case it may be reasonable simply to set $\theta_{\mathcal{H}}(\hat{\nu}^{(0)}) = \theta_{\mathcal{H}}(\mu)$ (but $\theta'_{\mathcal{H}}(\hat{\nu}^{(0)}) = 0$) which gives a nearby point in exponential coordinates to $\theta(\mu)$. This is the initialization method shown in our illustration (see

[2] Later, in Section 3.2.2, we specify our own accelerated version of iterative scaling which we call *loopy iterative scaling* (LIS). We then use LIS in place of IPF for m-projection.

[3] In a computer implementation of this procedure we may overwrite μ so that m-projection is performed "in place" by modification of the given model. We need only store those active moment characteristics of the model corresponding to statistics of the thinned family.

[4] One general and robust method is to m-project to an even lower-order embedded exponential family of $\mathcal{H}$ which has simple structure allowing the m-projection to be calculated directly. For instance, we could m-project to the family of fully-factored distributions (fully-disconnected graphical models). The m-projection is then given by the product of marginal distributions. We could also m-project to families of Markov chains or trees using Minka's m-projection method [96].

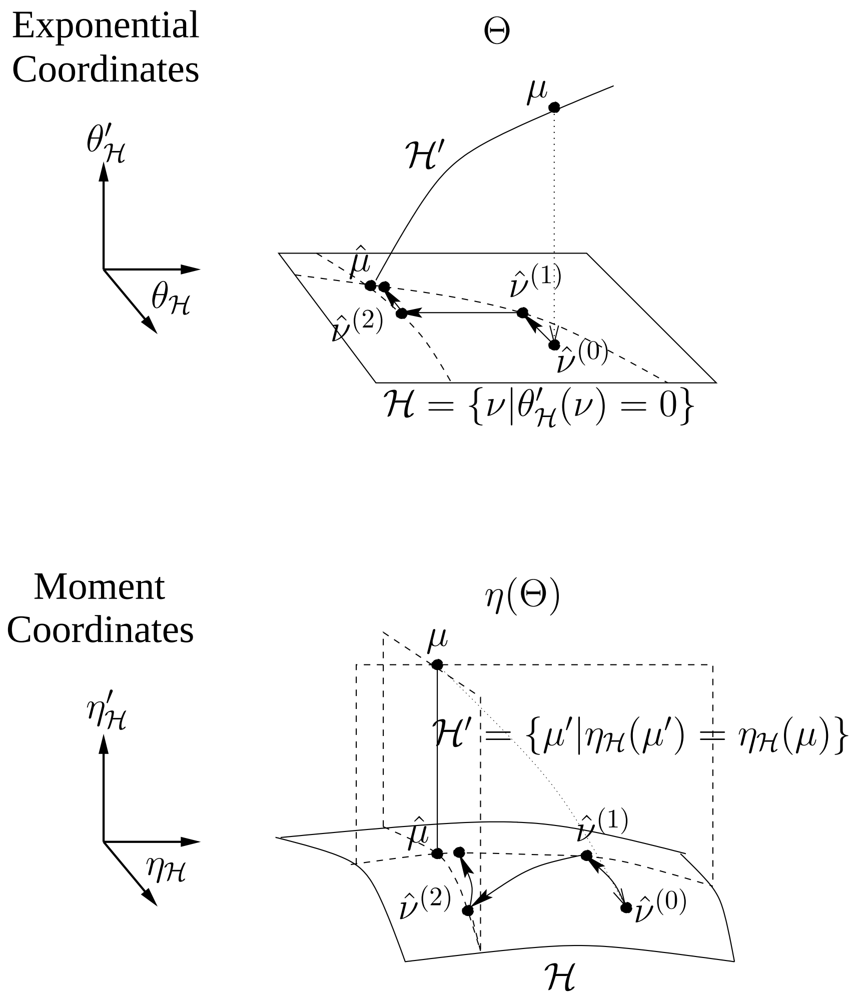


Figure 3-1: M-Projection by moment matching depicted in both exponential coordinates (top) and moment coordinates (bottom). Given μ, we wish to minimize $D(\mu\|\nu)$ over $\nu \in \mathcal{H}$ (solve for the m-projection of μ to the e-flat submanifold $\mathcal{H}$). This is given by the intersection $\mathcal{H} \cap \mathcal{H}' = \{\hat{\mu}\}$ where $\mathcal{H}'$ is the $\mathcal{I}$-orthogonal m-geodesic containing μ. To obtain $\theta(\hat{\mu})$, we first provide an initial guess $\hat{\nu}^{(0)} \in \mathcal{H}$. This seeds a moment-matching procedure which solves for $\nu \in \mathcal{H}$ such that $\eta_{\mathcal{H}}(\nu) = \eta_{\mathcal{H}}(\mu)$. The method shown is the IPF version of iterative scaling which generates a sequence of alternating e-projections $\hat{\nu}^{(k)}$ to the two m-flat submanifolds (shown at bottom) with intersection $\mathcal{H}'$. Each m-flat submanifold satisfies a subset of the moments we are trying to match. Hence, the sequence $\hat{\nu}^{(k)}$ converges to the e-projection $\hat{\nu}$ of $\hat{\nu}^{(0)}$ to $\mathcal{H}'$. Since each e-projection stays in $\mathcal{H}$ the e-projection $\hat{\nu}$ is also the m-projection of μ to $\mathcal{H}$, i.e. $\hat{\nu} = \hat{\mu}$. Note, to obtain a good approximation for μ in $\mathcal{H}$ it is not strictly necessary that we match moments exactly. Rather, we may terminate iterative scaling once $\hat{\nu}^{(k)}$ is within some KL-divergence from each of the m-flat submanifolds.

top panel of Figure 3-1). In any case, once an initial guess $\hat{\nu}^{(0)}$ is given, we may then e-project this initial starting point to $\mathcal{H}'$ by iterative scaling as described in Section 2.2.4. As illustrated in Figure 3-1, this generates a sequence $\hat{\nu}^{(k)}$ of e-projections (solid straight lines in top figure), each imposing a subset of the active moment constraints. This sequence converges to $\hat{\nu}$, the e-projection of $\hat{\nu}^{(0)}$ to $\mathcal{H}'$ satisfying all active moment constraints $\eta^*_{\mathcal{H}}$. In view of duality, this also gives $\hat{\mu}$, the desired m-projection of μ to $\mathcal{H}$. Finally, we remark that $\hat{\mu}$ is also the maximum entropy model subject to active moment constraints. That is, $\hat{\mu} = \arg\max_{\mu \in \mathcal{H}'} h[\mu]$.

We now specify this general moment matching procedure for the information form of the GMRF. Here, the graphical model μ corresponds to the information model $x_\Gamma \sim \mathcal{N}^{-1}(h, J)$. This is Markov with respect to the adjacency graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ having edges $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ for all $\gamma, \lambda \in \Gamma$ such that $J_{\gamma,\lambda} \neq 0$ (see Proposition 6). We then consider $\mathcal{F}$ as the family of GMRFs which are Markov with respect $\mathbf{G}_\Gamma$ such that $J_{\gamma,\lambda} = 0$ for all $\{\gamma, \lambda\} \notin \mathcal{E}_\Gamma$. We may pose model thinning as m-projection to the subfamily $\mathcal{H}$ which are Markov with respect to an embedded graph $\mathbf{G}'_\Gamma = (\Gamma, \mathcal{E}'_\Gamma)$ (such that $\mathcal{E}'_\Gamma \subseteq \mathcal{E}_\Gamma$) imposing further sparsity upon the interaction matrix J. As was shown in Section 2.1.5, the information parameters (h, J) correspond to exponential parameters θ in the description of the Gaussian density as an exponential family. Hence, imposing Markov structure by setting off-diagonal entries of J to zero also corresponds to selection of an embedded exponential family based on a reduced set of statistics.

We specify the following procedure to evaluate the m-projection to the embedded exponential family by iterative proportional fitting. This moment-matching procedure is structured according to a collection of cliques $\mathcal{C} \subseteq \mathcal{C}(\mathbf{G}'_\Gamma)$ which cover the thinned interaction graph $\mathbf{G}'_\Gamma = (\Gamma, \mathcal{E}'_\Gamma)$, so that each vertex $\gamma \in \Gamma$ and each edge $\{\gamma, \lambda\} \in \mathcal{E}'_\Gamma$ is contained in some $\Lambda \in \mathcal{C}$.

M-Projection/Moment-Matching in GMRFs:

- *Input.* Graphical model $\mu = (h, J)$, graph $\mathbf{G}'_\Gamma = (\Gamma, \mathcal{E}'_\Gamma)$, cliques $\mathcal{C} \subseteq \mathcal{C}(\mathbf{G}'_\Gamma)$ covering $\mathbf{G}'_\Gamma$, moment-matching tolerance $\epsilon > 0$.
- *Inference.* Calculate $\eta^*_\Lambda = (\hat{x}_\Lambda, P_\Lambda)$ for all $\Lambda \in \mathcal{C}$. Let $(\hat{h}^*_\Lambda, \hat{J}^*_\Lambda) = (P_\Lambda^{-1}\hat{x}_\Lambda, P_\Lambda^{-1})$.
- *Initialization.* Set $J_{\gamma,\lambda} = 0$ for all $\{\gamma, \lambda\} \notin \mathcal{E}'_\Gamma$.
- *Iterative Scaling.* Until convergence, do the following:
  - *Inference.* Calculate $\eta_\Lambda = (\hat{x}_\Lambda, P_\Lambda)$ for all $\Lambda \in \mathcal{C}$. Let $(\hat{h}_\Lambda, \hat{J}_\Lambda) = (P_\Lambda^{-1}\hat{x}_\Lambda, P_\Lambda^{-1})$.
  - *Test for convergence.* Calculate $d_\Lambda = D(\eta^*_\Lambda \| \eta_\Lambda)$, the KL-divergence between Gaussian distributions with moments η^*_Λ and η_Λ, for $\Lambda \in \mathcal{C}$ and set $\hat{d} = \max_\Lambda d_\Lambda$. If $\hat{d} < \epsilon$, then terminate iterative scaling loop.
  - *IPF Update.* Pick $\Lambda \in \mathcal{C}$, set $h_\Lambda \leftarrow h_\Lambda + (\hat{h}^*_\Lambda - \hat{h}_\Lambda)$ and $J_\Lambda \leftarrow J_\Lambda + (\hat{J}^*_\Lambda - \hat{J}_\Lambda)$.
- *Output.* Thinned model $\hat{\mu} = (h, J)$ giving m-projection of input μ to family of GMRFs $\mathcal{H}$ Markov w.r.t. $\mathbf{G}'_\Gamma$.

We also remark that alternative iterative scaling procedures (such as generalized iterative scaling [38] and improved iterative scaling [106] discussed in Section 2.2.4) may be used in place of IPF to possibly accelerate convergence. The advantage of such alternatives is that for every execution of the IS loop (each requiring a global inference computation), all parameters of the model are updated (whereas in IPF, only one clique is updated per inference computation). We also offer our own alternative to IPF described next.

3.2.2 Loopy Iterative Scaling

We now develop our modeling approach which may be seen as a hybrid method combining iterative scaling techniques for loopy graphs (discussed in the previous section) with exact methods for m-projection to families of Markov trees (such as proposed by Minka [96]). Essentially, we extend the method for m-projection to Markov trees to yield an iterative refinement procedure in loopy graphs. We relate this approach to the Bethe and Kikuchi approximations (to be discussed) employed in some variational *inference* methods such as loopy belief propagation (Yedidia [135], Yedidia et al [136]), tree reparameterization (Wainwright [129]) and the cluster variation method (Kappen and Wiegerinck [80]).[5] We develop an iterative refinement procedure, inspired by but distinct from Bethe and Kikuchi approximation, which we call *loopy*

[5] However, we must emphasize that our goal here is *modeling* rather than *inference*. Consequently, we consider these Bethe and Kikuchi approximations in somewhat of a different light than the reader may be accustomed to.

iterative scaling (LIS). This *modeling* approach may be considered as the modeling analog of the loopy belief propagation *inference* approach.

We first define what we mean by Bethe and Kikuchi approximation (in the context of modeling), and then develop our iterative refinement procedure, based on a generalized notion of Kikuchi approximation which we call the *relative Kikuchi approximation.* The leads to an iterative modeling procedure which then closely resembles iterative scaling methods such as discussed in the preceding section.

Bethe Approximation. We define the *Bethe approximation*[6] for a probability distribution $\mu(x_\Gamma)$ as a graphical model constructed from a collection of singleton and pairwise marginal distributions of μ on the vertices and edges of a graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$. Given the marginal distributions $(\mu_\Lambda(x_\Lambda), \Lambda \in \{\{\gamma\} | \gamma \in \Gamma\} \cup \mathcal{E}_\Gamma)$[7], let the probability distribution μ_{Bethe} be defined by

$$\mu_{\text{Bethe}}(x_\Gamma) = \frac{1}{Z} \prod_{\gamma \in \Gamma} \psi_{\text{Bethe}}(x_\gamma) \prod_{\{\gamma,\lambda\} \in \mathcal{E}_\Gamma} \psi_{\text{Bethe}}(x_\gamma, x_\lambda) \tag{3.8}$$

with compatibility functions:

$$\psi_{\text{Bethe}}(x_\gamma) = \mu(x_\gamma) \tag{3.9}$$

$$\psi_{\text{Bethe}}(x_\gamma, x_\lambda) = \frac{\mu(x_\gamma, x_\lambda)}{\mu(x_\gamma)\mu(x_\lambda)} \tag{3.10}$$

and where Z is the normalization constant. Note that this is a Gibbs distribution $\mu_{\text{Bethe}} = \frac{1}{Z} \exp \sum_\Lambda \phi_{\text{Bethe}}(x_\Lambda)$ with interaction potentials defined as

$$\phi_{\text{Bethe}}(x_\gamma) = \log \mu(x_\gamma) \tag{3.11}$$

$$\phi_{\text{Bethe}}(x_\gamma, x_\lambda) = \log \mu(x_\gamma, x_\lambda) - \phi_{\text{Bethe}}(x_\gamma) - \phi_{\text{Bethe}}(x_\lambda) \tag{3.12}$$

for each site $\gamma \in \Gamma$ and each edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$. This probability distribution factors with respect to the graph $\mathbf{G}_\Gamma$ so that $(\mathrm{x}_\Gamma \sim \mu_{\text{Bethe}}, \mathbf{G}_\Gamma)$ defines a MRF. This defines the Bethe approximation and gives an explicit recursive procedure for constructing this approximation from the specified marginals.

Now let us consider why μ_{Bethe} might be considered a reasonable approximation for μ within the family $\mathcal{F}(\mathbf{G}_\Gamma)$ of pairwise MRFs defined on $\mathbf{G}_\Gamma$.[8] Recall that, ultimately,

[6] See also discussion and references in Yedidia [135].

[7] Note that, in the context of modeling, we presume these marginal distributions are *given* which is the reverse situation to inference where we wish to calculate (or somehow estimate) these marginals from a given graphical model. In the context of model thinning, which is our real interest here, these marginals must be computed by recursive inference of the model $\mu(x)$.

[8] That is, the family of Gibbs random fields which factor according to the hypergraph $\mathbf{H}_\Gamma = (\Gamma, \mathcal{E}_\Gamma \cup \{\{\gamma\} | \gamma \in \Gamma\})$ so that all potentials are either singleton effects or pairwise interactions. This family is Markov with respect to $\mathbf{G}_\Gamma$ but may not contain every MRF on $\mathbf{G}_\Gamma$ because of the restriction to pairwise interactions.

our goal is to determine the *m-projection* of μ to this family.

$$\hat{\mu} = \arg \min_{\nu \in \mathcal{F}(\mathbf{G}_\Gamma)} D(\mu \| \nu) \tag{3.13}$$

This m-projection is also the maximum-entropy distribution with marginal distributions $\mu(x_\gamma)$ for each site $\gamma \in \Gamma$ and $\mu(x_\gamma, x_\lambda)$ for each edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$. The maximum-entropy distribution lies in the family $\mathcal{F}(\mathbf{G}_\Gamma)$ and is uniquely determined within that family by those singleton and edgewise marginal distributions. That is, if we can find $\nu \in \mathcal{F}(\mathbf{G}_\Gamma)$ such that $\nu(x_\gamma) = \mu(x_\gamma)$ for all $\gamma \in \Gamma$ and $\nu(x_\gamma, x_\lambda) = \mu(x_\gamma, x_\lambda)$ for all $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ then we have found the m-projection, i.e. $\nu = \hat{\mu}$.

In Trees, Bethe Approximation = M-Projection. To motivate this perspective, suppose that the graph $\mathbf{G}_\Gamma$ is in fact a tree. Then, the Bethe approximation corresponds precisely to the canonical factorization sought in the "refactorization" view of belief propagation where local potentials are directly related to local marginal distributions.[9] Then, in tree graphs $\mathbf{G}_\Gamma$, the singleton and edgewise marginal distributions of the Bethe approximation agree *exactly* with those of μ.[10] Hence, Bethe approximation actually gives an *exact* method for m-projection to families of Markov trees.[11]

In General, Bethe Approximation $\neq$ M-Projection. More generally, in loopy graphs $\mathbf{G}_\Gamma$, the marginals of the Bethe approximation will *not* exactly agree with those of μ. Indeed, this is unfortunately the case even if μ actually is itself a member of the family $\mathcal{F}(\mathbf{G}_\Gamma)$. Nevertheless, if the "loopiness" of the model is nearly negligible, so that any additional interactions induced by variable elimination are weak, then we expect the marginals of the Bethe approximation to agree approximately with those of μ. This is shown in that the Bethe approximation is a fixed point of both loopy belief propagation and tree reparameterization where the "pseudo-marginals" computed under these loopy inference methods match the true marginals of μ (Wainwright [129]). In this regard, the Bethe approximation μ_{Bethe} based on $\mathbf{G}_\Gamma$ may be understood as an *approximation* for the desired m-projection $\hat{\mu}$.

However, it is still our objective to develop an iterative extension of this Bethe approximation which *will* allow us to find the desired m-projection in loopy graphs (at least to a desired level of precision). But first, before we develop this iterative method, we consider a more general form of Bethe approximation which allows higher-order marginal distributions (involving more than two sites of the field) to be incorporated

[9] This "refactorization" viewpoint was reviewed in Section 2.3 and is a guiding principle in tree reparameterization (Wainwright [129]) but also arises in analysis of junction tree inference procedures (Shenoy and Shafer [122], Dawid [40]).

[10] This occurs because, in the decimation approach to inference, eliminating any leaf node of the tree produces a message identical to one. That is, $\int \psi_{\text{Bethe}}(x_\gamma, x_\lambda)\psi_{\text{Bethe}}(x_\lambda) dx_\lambda = \int \mu(x_\lambda | x_\gamma) dx_\lambda = 1$ so that the potential ψ_γ is *not* modified by elimination of vertex λ. Repeating this leaf-elimination procedure until only a single site $\gamma \in \Gamma$ remains shows that $\mu_{\text{Bethe}}(x_\gamma) = \psi_{\text{Bethe}}(x_\gamma) = \mu(x_\gamma)$. Similarly, eliminating all but two adjacent sites $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ shows that $\mu_{\text{Bethe}}(x_\gamma, x_\lambda) = \psi_{\text{Bethe}}(x_\gamma, x_\lambda)\psi_{\text{Bethe}}(x_\gamma)\psi_{\text{Bethe}}(x_\lambda) = \mu(x_\gamma, x_\lambda)$.

[11] This method for m-projection to families of Markov trees is equivalent to the method proposed by Minka [96] although he considers a causal (directed) factorization on the Markov tree rather than the symmetric Bethe approximation discussed here.

into the approximation in a consistent manner. We interpret these approximations as truncated versions of an exact representation of μ obtained by a Möbius inversion procedure to be discussed.[12]

As a preliminary to this discussion, let us say that a hypergraph $\mathbf{H}_\Gamma = (\Gamma, \mathcal{H}_\Gamma)$ is *intersection complete* if intersections of hyperedges are also hyperedges. That is, for every pair of hyperedges $\Lambda_1, \Lambda_2 \in \mathcal{H}_\Gamma$ the intersection $\Lambda_{1,2} = \Lambda_1 \cap \Lambda_2$ is contained in the collection $\mathcal{H}_\Gamma$. Also, we will say that $\mathbf{H}_\Gamma$ is a *hypertree* if it is the clique hypergraph of some chordal graph. The maximal hyperedges of a hypertree may be linked together so as to form a junction tree, an acyclic clique graph satisfying the running-intersection property (Section 2.3.2).

Kikuchi Approximation. We define the *Kikuchi approximation*[13] for a probability distribution $\mu(x_\Gamma)$ as a graphical model constructed from the marginal distributions of μ on the hyperedges of an intersection complete hypergraph $\mathbf{H}_\Gamma = (\Gamma, \mathcal{H}_\Gamma)$. Given the marginal distributions $(\mu_\Lambda(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$, let the probability distribution $\mu_{\text{Kikuchi}}(x_\Gamma)$ be defined by

$$\mu_{\text{Kikuchi}}(x_\Gamma) = \frac{1}{Z} \prod_{\Lambda \in \mathcal{H}_\Gamma} \psi_{\text{Kikuchi}}(x_\Lambda) \tag{3.14}$$

with compatibility functions defined by

$$\psi_{\text{Kikuchi}}(x_\Lambda) = \frac{\mu(x_\Lambda)}{\prod_{\Lambda' \subsetneq \Lambda} \psi_{\text{Kikuchi}}(x_{\Lambda'})} \tag{3.15}$$

for each $\Lambda \in \mathcal{H}_\Gamma$ and with normalization constant Z. That is, for *minimal hyperedges* (not a superset of another hyperedge) we set $\psi_{\text{Kikuchi}}(x_\Lambda)$ to the marginal $\mu(x_\Lambda)$ but for *non-minimal hyperedges* (a proper superset of another hyperedge) we set $\psi_{\text{Kikuchi}}(x_\Lambda)$ to $\mu(x_\Lambda)$ divided by the product of all lower-order compatibility functions defined within Λ. Note that this is a Gibbs distribution $\mu_{\text{Kikuchi}}(x_\Gamma) = \frac{1}{Z} \exp \sum_\Lambda \phi_{\text{Kikuchi}}(x_\Lambda)$ with interaction potentials recursively defined by

$$\phi_{\text{Kikuchi}}(x_\Lambda) = \log \mu(x_\Lambda) - \sum_{\Lambda' \subsetneq \Lambda} \phi_{\text{Kikuchi}}(x_{\Lambda'}) \tag{3.16}$$

for each $\Lambda \in \mathcal{H}_\Gamma$. The probability distribution μ_{Kikuchi} factors with respect to the hypergraph $\mathbf{H}_\Gamma$ so that $\mathrm{x}_\Gamma \sim \mu_{\text{Kikuchi}}$ is Markov with respect to the interaction graph $\mathbf{G}_\Gamma = \text{adj } \mathbf{H}_\Gamma$. This defines the Kikuchi approximation and also gives an explicit recursive procedure for constructing the approximation. We also note that Bethe approximation is a special case of Kikuchi approximation.

Now, let us consider why this might be considered as a good approximation for the m-projection of μ to the family of graphical models $\mathcal{F}(\mathbf{H}_\Gamma)$ with interaction hypergraph $\mathbf{H}_\Gamma$. The desired m-projection minimizes the KL-divergence $D(\mu\|\nu)$ over

[12] The author credits this interpretation of Kikuchi approximations to Martin Wainwright (based on informal discussions).

[13] See also discussion and references in Yedidia [135].

$\nu \in \mathcal{F}(\mathbf{H}_\Gamma)$ and is the maximum-entropy distribution with marginal distributions $(\mu(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$.

Again, there are some special cases where the Kikuchi approximation is equivalent to m-projection. In particular, if the hypergraph $\mathbf{H}_\Gamma$ is a hypertree then, by a similar argument as given for tree-structured Bethe approximations, the marginal distributions of the Kikuchi approximation will exactly agree with the desired marginals and the Kikuchi approximation is equivalent to m-projection. However, we would also like to develop m-projection techniques for more general circumstances in which $\mathbf{H}_\Gamma$ is *not* a hypertree. Then, the Kikuchi approximation is, at best, an approximation for the m-projection (the marginal distributions of the Kikuchi approximation need *not* exactly agree with the desired marginal distributions, even when μ is a member of the family $\mathcal{F}(\mathbf{H}_\Gamma)$).[14] Hence, we would like to develop an iterative extension of Kikuchi approximation which will allow us to find the desired m-projection. As a guide for developing our iterative refinement procedure, we consider the following alternative interpretation for how Kikuchi approximations are constructed.

Alternative Interpretation of Kikuchi Approximations. The Kikuchi approximation may also be understood as a *truncated Möbius inversion* of $\mu(x)$ based on the log-marginals of μ. To see this, consider the collection of functions defined by

$$U_\Lambda(x_\Lambda) = \log \mu(x_\Lambda) \tag{3.17}$$

for each $\Lambda \subseteq \Gamma$. Based on these functions, we may construct a potential specification for μ with potentials defined by

$$V_\Lambda(x_\Lambda) = \sum_{\Lambda' \subseteq \Lambda} (-1)^{|\Lambda \setminus \Lambda'|} U_{\Lambda'} \tag{3.18}$$

for each $\Lambda \subseteq \Gamma$. By the Möbius inversion lemma, $U_\Lambda = \sum_{\Lambda' \subseteq \Lambda} V_{\Lambda'}$. This shows that $V_\Lambda = U_\Lambda - \sum_{\Lambda' \subsetneq \Lambda} V_{\Lambda'}$ which is analogous to the recursive definition of the Kikuchi potentials ϕ_{Kikuchi} given in (3.16). Also, $U_\Gamma = \log \mu = \sum_{\Lambda \subseteq \Gamma} V_\Lambda$ so that we obtain an *exact* representation for μ as a Gibbs distribution with potential specification $V = (V_\Lambda(x_\Lambda), \Lambda \subseteq \Gamma)$.

$$\mu(x_\Gamma) = \exp \sum_{\Lambda \subseteq \Gamma} V(x_\Lambda) \tag{3.19}$$

However, computing V_Λ for every $\Lambda \subseteq \Gamma$ is intractable and does not (in general) give a sparse graphical model (i.e. higher-order potentials do not necessarily vanish). But suppose that most of the variation in $U(x_\Gamma) = \log \mu(x_\Gamma)$ is nevertheless captured by the lower-order interactions in V. Then, we might consider approximation of μ by

[14] Although, in a sense, we might expect the marginals of the Kikuchi approximation to at least *approximately* agree with the desired marginals. We suggest this because the Kikuchi approximation is a fixed point of certain generalized versions of loopy belief propagation structured according to the hypergraph $\mathbf{H}_\Gamma$. For instance, if we consider any embedded hypertree of $\mathbf{H}_\Gamma$, a subset of the hyperedges which form a hypertree, and then run belief propagation on just this embedded hypertree, then the pseudo-marginals produced by this inference are precisely the desired marginal distributions. Then, insofar as such approximate inference methods are accurate, then the Kikuchi approximation should provide a good approximation for the m-projection.

truncation of the potential specification V and normalizing the resulting approximation for μ. This is the basic idea underlying Kikuchi approximations. Note also that the Bethe approximation is a special case of Kikuchi approximation. To be precise, the Bethe approximation based on graph $\mathbf{G}_\Gamma$ is equivalent to the Kikuchi approximation based on hyperedges $\mathcal{H}_\Gamma = \mathcal{E}_\Gamma \cup \{\{\gamma\}|\gamma \in \Gamma\}$. Hence, the Bethe approximation may also be understood as a truncated Möbius inversion representation of μ.

We should also remark that the Kikuchi approximation may be specified in an equivalent *non-recursive form* (Kappen and Wiegerinck [80]) expressed directly in terms of the marginal distributions $(\mu(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$:

$$\begin{aligned} \mu_{\text{Kikuchi}}(x_\Lambda) &= \frac{1}{Z} \prod_{\Lambda \in \mathcal{H}_\Gamma} \mu^{c_\Lambda}(x_\Lambda) && (3.20) \\ &= \frac{1}{Z} \exp\left\{ \sum_{\Lambda \in \mathcal{H}_\Gamma} c_\Lambda \log \mu(x_\Lambda) \right\} && (3.21) \end{aligned}$$

The coefficients $(c_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$ have integer values and may be computed recursively as follows. If Λ is a maximal hyperedge (not a proper subset of another hyperedge), then $c_\Lambda = 1$. Otherwise, compute c_Λ recursively so as to satisfy the condition

$$\sum_{\Lambda \supseteq \Lambda'} c_\Lambda = 1 \tag{3.22}$$

for each (non-maximal) hyperedge $\Lambda' \in \mathcal{H}_\Gamma$. That is, once c_Λ is known for every hyperedge Λ such that $\Lambda \supsetneq \Lambda'$, we may compute $c_{\Lambda'}$ by the formula:

$$c_{\Lambda'} = 1 - \sum_{\Lambda \supsetneq \Lambda'} c_\Lambda. \tag{3.23}$$

This recursion then inductively determines all coefficients $(c_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$ specifying the Kikuchi approximation. Note that these coefficients are determined by the structure of the hypergraph $\mathbf{H}_\Gamma$ and may be precomputed independent of μ. For instance, in the Bethe approximation, where $\mathcal{H}_\Gamma = \mathcal{E}_\Gamma \cup \{\{\gamma\}|\gamma \in \Gamma\}$, we have $c_{\{\gamma,\lambda\}} = 1$ for each edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ and $c_\gamma = 1 - \deg \gamma$ for each site γ.[15]

Based on these considerations, we now develop a generalized form of Kikuchi approximation which leads to an iterative refinement procedure for the computation of m-projections.

Relative Kikuchi Approximation. We define the *relative Kikuchi approximation* for a probability distribution $\mu(x)$ constructed from a probability distribution $\nu(x)$ and an intersection complete hypergraph $\mathbf{H}_\Gamma$ as follows. Given $\nu(x)$ and the marginal

[15] Recall that deg γ denotes the *degree* of vertex γ, the number of vertices adjacent to γ in $\mathbf{G}_\Gamma = \text{adj } \mathbf{H}_\Gamma$.

distributions $(\mu(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$, we construct the relative Kikuchi approximation by

$$\hat{\nu}(x_\Gamma) = \frac{1}{Z}\nu(x)\exp\left\{\sum_{\Lambda\in\mathcal{H}_\Gamma} \Delta\phi(x_\Gamma)\right\} \tag{3.24}$$

with *relative potentials*

$$\Delta\phi(x_\Lambda) = \log\frac{\mu(x_\Lambda)}{\nu(x_\Lambda)} - \sum_{\Lambda'\subsetneq\Lambda} \Delta\phi(x_{\Lambda'}). \tag{3.25}$$

defined for each hyperedge $\Lambda \subset \mathcal{H}_\Gamma$ and with normalization constant Z. The idea here is that relative Kikuchi approximation is a procedure for updating (i.e. refining) an available approximation ν with the intent of adjusting the marginals of the initial approximation ν so as to better agree with the desired marginals as in μ.

Interpretation. This may also be understood as a *truncated Möbius inversion* representation of μ but constructed relative to ν. To see this, consider the collection of functions defined by

$$U_\Lambda(x_\Lambda) = \log\frac{\mu(x_\Lambda)}{\nu(x_\Lambda)} \tag{3.26}$$

for each $\Lambda \subseteq \Gamma$. Based on these functions, define a second collection of functions given by

$$V_\Lambda(x_\Lambda) = \sum_{\Lambda'\subseteq\Lambda} (-1)^{|\Lambda\backslash\Lambda'|} U_{\Lambda'}(x_{\Lambda'}) \tag{3.27}$$

Then, by the Möbius inversion lemma, we obtain an *exact* representation for μ relative to ν.

$$\mu(x_\Gamma) = \nu(x_\Gamma)\exp\sum_{\Lambda\subseteq\Gamma} V_\Lambda(x_\Lambda) \tag{3.28}$$

However, this is intractable since we have defined relative potentials V_Λ for every $\Lambda \subseteq \Gamma$. But let us conjecture that most of the variation in $U(x_\Gamma) = \log\mu(x_\gamma) - \log\nu(x_\gamma)$ is captured by lower-order potentials of the collection $V = (V_\Lambda, \Lambda \subseteq \Gamma)$. Then, we might consider approximation of μ by truncation of the potential specification V and normalization of the resulting approximation for μ. This is the idea underlying the proposed class of relative Kikuchi approximations.

The relative Kikuchi approximation can also be written in a *non-recursive form* as

$$\hat{\nu}(x_\Gamma) = \frac{1}{Z}\nu(x)\exp\left\{\sum_{\Lambda\in\mathcal{H}_\Gamma} c_\Lambda \log\frac{\mu(x_\Lambda)}{\nu(x_\Lambda)}\right\} \tag{3.29}$$

where the coefficients c_Λ are again determined by $\mathbf{H}_\Gamma$ as described previously for the Kikuchi approximation. It is then apparent that

$$\hat{\nu}(x_\Gamma) \propto \nu(x_\Gamma) \times \frac{\mu_{\text{Kikuchi}}(x_\Gamma)}{\nu_{\text{Kikuchi}}(x_\Gamma)} \tag{3.30}$$

which has the form of a structured IPF update. This may be viewed as "fusing" local IPF updates (i.e. the factor arising in the update formula of the IPF modeling procedure discussed in the previous section) according to the combination rule:

$$\hat{\nu}(x_\Gamma) \propto \nu(x_\Gamma) \times \prod_{\Lambda \in \mathcal{H}_\Gamma} \left(\frac{\mu(x_\Lambda)}{\nu(x_\Lambda)} \right)^{c_\Lambda} \tag{3.31}$$

This may be interpreted as first updating each maximal hyperedge Λ by the IPF update $\mu(x_\Lambda)/\nu(x_\Lambda)$. But then, whenever IPF updates overlap, having intersection Λ', we divide (or multiply) by the local IPF update $\mu(x_{\Lambda'})/\nu(x_{\Lambda'})$ an appropriate number of times $c_{\Lambda'}$ so as to correct for any overcounted (undercounted) subfields. Continuing this correction procedure until all hyperedges have been accounted for then reconstructs the relative Kikuchi approximation.

Viewing this relative Kikuchi approximation $\hat{\nu}$ as refining ν to give an improved approximation for μ, this suggests the following iterative refinement procedure which is apparently closely related to iterative scaling techniques.

Iterative Refinement. Finally, we specify an iterative procedure to calculate the m-projection of $\mu(x_\Gamma)$ to the family of MRFs which respect the graph $\mathbf{G}_\Gamma$. Let $\mathbf{H}_\Gamma = (\Gamma, \mathcal{H}_\Gamma)$ be an intersection complete hypergraph such that (i) adj $\mathbf{H}_\Gamma = \mathbf{G}_\Gamma$ and (ii) $\mathcal{C}^*(\mathbf{G}_\Gamma) \subset \mathcal{H}_\Gamma$. Condition (i) insures that the family $\mathcal{F}(\mathbf{H}_\Gamma)$ respects the graph $\mathbf{G}_\Gamma$. Condition (ii) is necessary to insure that the family $\mathcal{F}(\mathbf{H}_\Gamma)$ contains *all* Gibbs random fields which respect $\mathbf{G}_\Gamma$.[16] Calculate the coefficients $(c_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$. Also, perform recursive inference for the model $\mu(x_\Gamma)$ computing the collection of marginal distributions $(\mu(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$. Now, we would like to determine $\hat{\mu} \in \mathcal{F}(\mathbf{H}_\Gamma)$ satisfying the marginal constraints $\hat{\mu}(x_\Lambda) = \mu(x_\Lambda)$ for all $\Lambda \in \mathcal{H}_\Gamma$. Based on the idea of relative Kikuchi approximation, we propose the following procedure for solving these marginal constraints within the family $\mathcal{F}(\mathbf{H}_\Gamma)$.

Suppose that we have some initial guess $\hat{\nu}^{(0)} \in \mathcal{F}(\mathbf{H}_\Gamma)$ for an approximation of μ. This could be initialized in a variety of ways. For instance, we could set this to the product of marginal distributions:

$$\hat{\nu}^{(0)}(x_\Gamma) = \prod_{\gamma \in \Gamma} \mu(x_\gamma) \tag{3.32}$$

More generally, we could initialize $\hat{\nu}^{(0)}$ to any Kikuchi approximation which respects the graph $\mathbf{G}_\Gamma$. Also, when thinning exponential families, we could initialize $\hat{\nu}^{(0)}$ by setting some exponential parameters of μ to zero so as to respect the graph $\mathbf{G}_\Gamma$.

In any case, given this starting point $\hat{\nu}^{(0)}$, we then generate a sequence of relative Kikuchi approximations where each approximation is constructed from the preceding

[16] However, if the desired m-projection is known to lie in a lower-order family (e.g. the family of pairwise MRFs) then this latter condition could probably be relaxed. Yet, it may still be desirable to keep these higher-order hyperedges in order to allow more accurate Kikuchi approximation possibly improving the stability and convergence properties of the following procedure.

approximation and attempts to provide an improved approximation for μ:

$$\hat{\nu}^{(k+1)}(x_\Gamma) = \frac{1}{Z^{(k)}}\hat{\nu}^{(k)} \exp\left\{\sum_{\Lambda\in\mathcal{H}_\Gamma} c_\Lambda \log \frac{\mu(x_\Lambda)}{\hat{\nu}^{(k)}(x_\Lambda)}\right\} \tag{3.33}$$

Each relative Kikuchi approximation is structured according to the hypergraph $\mathbf{H}_\Gamma$ and requires computation of the collection of marginal distributions $(\hat{\nu}^{(k)}(x_\Lambda), \Lambda \in \mathcal{H}_\Gamma)$ of the preceding iterate. Since this iteration is seeded by $\hat{\nu}^{(0)} \in \mathcal{F}(\mathbf{H}_\Gamma)$ and each relative Kikuchi update also factors with respect to $\mathbf{H}_\Gamma$, by induction, all later iterates remain in this family and are hence Markov with respect to $\mathbf{G}_\Gamma$. We call this iterative refinement procedure *loopy iterative scaling* (LIS).

Note the striking resemblance between our LIS approach and related iterative scaling methods based on e-projections. Essentially, our approach is analogous to iterative scaling but where we replace e-projections by relative Kikuchi approximations. The intent of relative Kikuchi approximations is the same as for e-projection in the iterative scaling approach – to impose marginal constraints. Also, the Kikuchi update formula for $\hat{\nu}^{(k+1)}$ actually has the same form as an e-projection imposing marginal constraints, i.e. an exponential model based on the distribution $\hat{\nu}^{(k)}$ and statistics $t^{(k+1)}(x_\Lambda) = \log \mu(x_\Lambda) - \log \hat{\nu}^{(k)}(x_\Lambda)$ where the coefficients c_Λ are interpreted as exponential parameters. This might point the way to some further refinements of our method. Finally, we remark on the possibility of employing a different hypergraph at each iteration of the refinement procedure. That is, rather than specifying one hypergraph which both respects and covers the graph $\mathbf{G}_\Gamma$ we might instead specify a collection of hypergraphs which individually respect and collectively cover the graph $\mathbf{G}_\Gamma$. Then, by iterating over these hypergraphs and performing relative Kikuchi approximations structured according to each hypergraph, we obtain a more general update scheme which includes both the proposed LIS method and extant iterative scaling methods. For instance, the IPF technique is recovered by selecting hypergraphs to correspond to individual cliques of $\mathbf{G}_\Gamma$. This also suggests further novel possibilities such as performing relative Kikuchi updates based on embedded trees. This, apparently, would correspond to the modeling analog of tree reparameterization (Wainwright [129]). However, we focus on just the single-hypergraph LIS approach in this thesis.

In closing, we specify an implementation of this LIS approach for the information form of GMRFs:

Loopy Iterative Scaling for GMRFs:

- *Input.* Graphical model $\nu = (h, J)$, hypergraph $\mathbf{H}_\Gamma$, coefficients $(c_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$, tolerance ϵ, prescribed moments $\eta^*_\Lambda = (\hat{x}^*_\Lambda, P^*_\Lambda)$ for all $\Lambda \in \mathcal{H}_\Gamma$.
- *Initialization.* Let $(\hat{h}^*_\Lambda, \hat{J}^*_\Lambda) = ((P^*_\Lambda)^{-1}\hat{x}^*_\Lambda, (P^*_\Lambda)^{-1})$ for all $\Lambda \in \mathcal{H}_\Gamma$.
- *Loop.* Until convergence, do the following.
  - *Inference.* Calculate $\eta_\Lambda = (\hat{x}_\Lambda, P_\Lambda)$ for all $\Lambda \in \mathcal{H}_\Gamma$. Let $(\hat{h}_\Lambda, \hat{J}_\Lambda) = (P_\Lambda^{-1}\hat{x}_\Lambda, P_\Lambda^{-1})$.
  - *Test for convergence.* Let $d_\Lambda = D(\eta^*_\Lambda \| \eta_\Lambda)$ and $\hat{d} = \max_\Lambda d_\Lambda$. If $\hat{d} < \epsilon$, then terminate iterative scaling loop.
  - *Update.* Set $h \leftarrow h + \sum_{\Lambda \in \mathcal{H}_\Gamma} c_\Lambda(\hat{h}^*_\Lambda - \hat{h}_\Lambda)$ and $J \leftarrow J + \sum_{\Lambda \in \mathcal{H}_\Gamma} c_\Lambda(\hat{J}^*_\Lambda - \hat{J}_\Lambda)$ (zero pad local updates in taking the sum).
- *Output.* modified (h, J) giving e-projection of input model to m-flat submanifold specified by prescribed moments $(\eta^*_\Lambda, \Lambda \in \mathcal{H}_\Gamma)$.

This loopy iterative scaling procedure may be used in place of IPF in the moment-matching approach to m-projection. Experiments performed in Section 3.4 indicate the utility of this method. In the next section, we develop our incremental model thinning procedure where LIS may be used for the m-projection subroutine. We also remark that some possible extensions of this m-projection method and also some other promising alternatives are discussed as recommendations for further research in Chapter 5.

3.3 Model Selection

In this section we develop our approach to model thinning. As in the previous section, we are again given a graphical model μ which we wish to approximate by a more compact yet faithful model ν. But now we are free to select which statistics of the model μ are retained and which are neglected. Selection of the statistics $t_\mathcal{H}(x)$ corresponds to selection of an exponential family $\mathcal{H}$ containing ν. This may also be posed as selection of the graphical structure of the model ν. We employ the information criterion $V(\mu; \nu)$ both to guide our selection of $\mathcal{H}$ and to determine $\nu \in \mathcal{H}$. Subject to $\nu \in \mathcal{H}$, minimization of $V(\mu; \nu)$ reduces to the m-projection problem addressed in the previous section. We now consider selection of the family $\mathcal{H}$ so as to (approximately) minimize $V(\mu; \mathcal{H}) \equiv \min_{\nu \in \mathcal{H}} V(\mu; \nu)$.

3.3.1 Inductive Approach

We now consider an *inductive* approach for selection of the embedded family $\mathcal{H} \subseteq \mathcal{F}$. This is a *double-loop procedure* where the *outer-loop* selects a sequence of nested exponential families and the *inner-loop* performs m-projection to each selected family by iterative moment matching. We denote the sequence of m-projections produced in the outer loop by $\hat{\mu}^{(k)}$ where $k = 0, 1, 2, \dots$ is a counter incremented in the outer loop. For iteration k of the outer loop, the inner loop generates a sequence $\hat{\nu}^{(k,i)}$ of models converging towards $\hat{\mu}^{(k+1)}$.

This procedure is initialized by $(\mathcal{H}_0, \hat{\mu}^{(0)}) \equiv (\mathcal{F}, \mu)$, where $\mu \in \mathcal{F}$ is the model we wish to thin, and then generates a sequence $(\mathcal{H}_k, \hat{\mu}^{(k)})$ of nested exponential families $\mathcal{H}_0 \supset \mathcal{H}_1 \supset \dots \supset \mathcal{H}_k$ and associated m-projections $\hat{\mu}^{(k)} \equiv \arg\min_{\nu \in \mathcal{H}_k} D(\mu \| \nu)$. The outer-loop attempts to select the next embedded exponential family $\mathcal{H}_{k+1}$ which comes nearest to the preceding m-projection $\hat{\mu}^{(k)}$. The inner-loop then calculates the m-projection of $\hat{\mu}^{(k)}$ to $\mathcal{H}_{k+1}$ employing previously discussed iterative moment-matching procedures. For instance, the IPF method illustrated previously in Figure 3-1 could be used. However, we prefer our LIS approach which provides an accelerated approach to moment matching. By Proposition 16, this $\hat{\mu}^{(k)}$ is also the m-projection of μ to $\mathcal{H}_k$. The advantage of this inductive approach is that the iterate $\hat{\mu}^{(k)}$ is available to help select the next embedded family $\mathcal{H}_{k+1}$. This thinning procedure continues until we can no longer identify an embedded family $\mathcal{H}_{k+1}$ such that $V(\hat{\mu}^{(k)}; \mathcal{H}_{k+1}) < 0$ so that the cumulative information criterion $V(\mu; \hat{\mu}^{k+1})$ is decreased by m-projection to $\mathcal{H}_{k+1}$. This is equivalent to requiring that $D(\hat{\mu}^{(k)} || \hat{\mu}^{(k+1)}) = h[\hat{\mu}^{(k+1))}] - h[\hat{\mu}^{(k)}] < \delta(K(\mathcal{H}_k) - K(\mathcal{H}_{k+1}))$, or that the information loss per removed model parameter does not exceed the threshold δ. This may be seen as a greedy suboptimal procedure for selection of $\mathcal{H}$ to (approximately) minimize $V(\mu; \mathcal{H})$. This general approach is outlined below.

Inductive Model Thinning:

- *Input.* Model $\mu \in \mathcal{F}$, information threshold δ, moment-matching tolerance ϵ (much smaller than δ).

- *Outer Loop.* Set $k = 0$, $(\hat{\mu}^{(0)}, \mathcal{H}_0) = (\mu, \mathcal{F})$. Do the following until termination is indicated.

  - *Select Embedded Family.* Select $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ so as to (at least approximately) minimize $D(\hat{\mu}^{(k)} \| \mathcal{H}_{k+1})$. If $D(\hat{\mu}^{(k)} \| \mathcal{H}_{k+1}) / (K(\mathcal{H}_k) - K(\mathcal{H}_{k+1})) > \delta$ then terminate thinning. Else, calculate moment coordinates η^*_{k+1} of $\hat{\mu}^{(k)}$ in $\mathcal{H}_{k+1}$.
  - *Inner Loop.* Initialize guess $\nu^{(k,0)} \in \mathcal{H}_{k+1}$ and then perform iterative moment-matching within family $\mathcal{H}_{k+1}$ generating sequence $\hat{\nu}^{(k,i)}$ approaching $\hat{\mu}^{(k+1)}$. Terminate moment-matching, setting $\hat{\mu}^{(k+1)} = \hat{\nu}^{(k,i)}$, when marginal distributions of $\hat{\nu}^{(k,i)}$ agree with those specified by η^*_{k+1} to within tolerance ϵ.
  - *Repeat.* Set $k \leftarrow k + 1$ and continue thinning.

- *Output.* Thinned model $\hat{\mu}^{(k)} \in \mathcal{H}_k$.

This approach to model thinning is essentially the inverse of the model building procedures of Pietra et al [106] and of Bach and Jordan [6]. While those two approaches employ e-projections to introduce additional statistics into the model (maximizing the information gain), we instead take the reverse approach of inductively deleting statistics by m-projections (minimizing the information loss).

Graphical Formulation. To implement this approach for exponential family graphical models, we consider selection of embedded exponential families $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ within the framework of thinning the associated adjacency graph. Let $\mathbf{H}_\Gamma^{(k)}$ denote the hypergraph describing the interaction structure of the family $\mathcal{H}_k$. The associated adjacency graph $\mathbf{G}_\Gamma^{(k)} = \text{adj } \mathbf{H}_\Gamma^{(k)}$ gives the Markov structure of the family $\mathcal{H}_k$. We consider embedded exponential families specified by deletion of hyperedges from this interaction hypergraph. In the case of families where all interactions are pairwise (such as for GMRFs) this is equivalent to pruning edges from the adjacency graph thus imposing further Markov restrictions upon the family. We will focus on this latter viewpoint for GMRFs, but recommend the former strategy (pruning hyperedges) more generally.

Hence, we specify an embedded graph $\mathbf{G}_\Gamma^{(k+1)} = (\Gamma, \mathcal{E}_\Gamma^{(k+1)})$ such that $\mathcal{E}_\Gamma^{(k+1)} \subseteq \mathcal{E}_\Gamma^{(k)}$. This is viewed as pruning or cutting the edges $\mathcal{K} = \mathcal{E}_\Gamma^{(k)} \setminus \mathcal{E}_\Gamma^{(k+1)}$. This specifies an embedded exponential family $\mathcal{H}_{k+1} \subseteq \mathcal{H}_k$ based on those statistics of family $\mathcal{H}_k$ which only couple sites which are adjacent in $\mathbf{G}_\Gamma^{(k+1)}$ (neglecting any statistics associated with pruned edges). We let $\theta_\mathcal{K}$ denote the exponential parameters scaling those neglected statistics. The embedded family is then specified as an e-flat submanifold $\mathcal{H}_{k+1} = \{\nu \in \mathcal{H}_k | \theta_\mathcal{K}(\nu) = 0\}$. Equivalently, this is the submanifold of $\mathcal{H}_k$ which is Markov with respect to $\mathbf{G}_\Gamma^{(k+1)}$. For the information representation (h, J) of GMRFs, this corresponds to setting to zero those off-diagonal entries of the interaction matrix J corresponding to pruned edges. That is, we set $J_{\gamma,\lambda} = 0$ for all $\{\gamma, \lambda\} \in \mathcal{K}$. Under this edge-pruning approach to model thinning, we would like to predict which edges of the graphical model $\hat{\mu}^{(k)}$ may be pruned while keeping $D(\hat{\mu}^{(k)} \| \hat{\mu}^{(k+1)})$ as small as possible.

3.3.2 Lower-Bound Estimate of Information Loss

Here we develop a lower bound estimate of the information loss due to pruning one edge from a graphical model by m-projection. Mutual information, conditional mutual information and conditional KL-divergence play a role here and are now defined.

Definition 7 *The* mutual information *between two random variables* x *and* y *with probability distribution $p(x, y)$ is defined as*

$$I_p(\mathrm{x}; \mathrm{y}) = E_p \left\{ \log \frac{p(\mathrm{x}, \mathrm{y})}{p(\mathrm{x})p(\mathrm{y})} \right\} \tag{3.34}$$

where $p(x)$ and $p(y)$ are marginal distributions of $p(x, y)$.

This is just the KL-divergence $D(p(\mathrm{x}, \mathrm{y}) \| q(\mathrm{x}, \mathrm{y}))$ of the factored approximation $q(x, y) = p(x)p(y)$ relative to the true $p(x, y)$. This indicates the following variational interpretation of mutual information.

Lemma 1 (Mutual Information) *Let $p(x, y)$ be the probability distribution for random variables* x *and* y*. Let $Q_{\mathrm{x}\perp\mathrm{y}}$ be the family of all factored probability distributions with respect to* x *and* y*.*

$$Q_{\mathrm{x}\perp\mathrm{y}} = \{q(x, y) = q(x)q(y)\} \tag{3.35}$$

Then, the minimum KL-divergence $D(p\|q)$ over $q \in Q_{\mathrm{x}\perp\mathrm{y}}$ is the mutual information $I_p(\mathrm{x}; \mathrm{y})$.

$$I_p(\mathrm{x}; \mathrm{y}) = \min_{q \in Q_{\mathrm{x}\perp\mathrm{y}}} D(p\|q) \tag{3.36}$$

Furthermore, the minimum is uniquely obtained by $q(x, y) = p(x)p(y)$ where $p(x)$ and $p(y)$ are the marginals of p.

Proof. A simple computation, for $q \in Q_{\mathrm{x}\perp\mathrm{y}}$, decomposes the KL-divergence as:

$$\begin{aligned}
D(p\|q) &= E_p\left\{\log \frac{p(\mathrm{x}, \mathrm{y})}{q(\mathrm{x}, \mathrm{y})}\right\} && (3.37)\\
&= E_p\left\{\log \frac{p(\mathrm{x}, \mathrm{y})}{q(\mathrm{x})q(\mathrm{y})}\right\} && (3.38)\\
&= E_p\left\{\log \frac{p(\mathrm{x}, \mathrm{y})}{p(\mathrm{x})p(\mathrm{y})} + \log \frac{p(\mathrm{x})p(\mathrm{y})}{q(\mathrm{x})q(\mathrm{y})}\right\} && (3.39)\\
&= E_p\left\{\log \frac{p(\mathrm{x}, \mathrm{y})}{p(\mathrm{x})p(\mathrm{y})} + \log \frac{p(\mathrm{x})}{q(\mathrm{x})} + \log \frac{p(\mathrm{y})}{q(\mathrm{y})}\right\} && (3.40)\\
&= I_p(\mathrm{x}; \mathrm{y}) + D(p(\mathrm{x})\|q(\mathrm{x})) + D(p(\mathrm{y})\|q(\mathrm{y})) && (3.41)\\
&\geq I_p(\mathrm{x}; \mathrm{y}) && (3.42)
\end{aligned}$$

The inequality follows from the non-negativity of KL-divergence. Equality occurs if and only if both $p(x) = q(x)$ and $p(y) = q(y)$ so that $q(x, y) = p(x)p(y)$. □

The family $Q_{\mathrm{x}\perp\mathrm{y}}$ corresponds to the hypothesis that x and y are independent. This suggests that mutual information might play a useful role in characterizing the KL-divergence induced under m-projections imposing Markov structure (conditional independencies). This idea is refined by considering the following averaged version of mutual information.

Definition 8 *The* conditional mutual information (CMI) $I_p(\mathrm{x}; \mathrm{y}|\mathrm{z})$ *between* x *and* y *given* z *under probability distribution p is defined as*

$$I_p(\mathrm{x}; \mathrm{y}|\mathrm{z}) = E_p\left\{\log \frac{p(\mathrm{x}, \mathrm{y}|\mathrm{z})}{p(\mathrm{x}|\mathrm{z})p(\mathrm{y}|\mathrm{z})}\right\} \tag{3.43}$$

where $p(x, y|z)$, $p(x|z)$ and $p(y|z)$ are conditional distributions of $p(x, y, z)$.

This should be distinguished from the *specific* conditional mutual information as a function of z (the mutual information under the conditional distribution $p(x, y|z)$ for a specific value of z). If we denote this latter quantity by $I(z) = I_p(\mathrm{x}; \mathrm{y}|z)$ then

$I_p(\mathrm{x};\mathrm{y}|\mathrm{z}) = \int p(z)I(z)dz$. Note that the specific CMI is a KL-divergence but CMI is not. However, CMI is related to the following averaged version of KL-divergence:

Definition 9 *The* conditional Kullback-Leibler divergence (CKL) *is defined as*

$$D(p(\mathrm{x};\mathrm{y}|\mathrm{z})\|q(\mathrm{x};\mathrm{y}|\mathrm{z})) = E_p\left\{\log\frac{p(\mathrm{x},\mathrm{y}|\mathrm{z})}{q(\mathrm{x},\mathrm{y}|\mathrm{z})}\right\} \tag{3.44}$$

where $p(x,y|z)$ and $q(x,y|z)$ are conditional probability distributions and the expectation is with respect to $p(x,y,z) = p(x,y|z)p(z)$.

This should be distinguished from the *specific* CKL-divergence,

$$D(p(\mathrm{x},\mathrm{y}|\mathrm{z}=z)\|q(\mathrm{x},\mathrm{y}|\mathrm{z}=z)) = E_p\left\{\log\frac{p(\mathrm{x},\mathrm{y}|z)}{p(\mathrm{x}|z)p(\mathrm{y}|z)}\right\}.$$

This is the KL-divergence between $p(x,y|z)$ and $q(x,y|z)$ evaluated as a function of z. Denoting this function by $d(z) = D(p(\mathrm{x},\mathrm{y}|z)\|q(\mathrm{x},\mathrm{y}|z))$, we have $D(p(\mathrm{x},\mathrm{y}|\mathrm{z})\|q(\mathrm{x},\mathrm{y}|\mathrm{z})) = \int p(z)d(z)dz$.

CMI is just the CKL-divergence,

$$I_p(\mathrm{x};\mathrm{y}|\mathrm{z}) = D(p(\mathrm{x};\mathrm{y}|\mathrm{z})\|q(\mathrm{x};\mathrm{y}|\mathrm{z})) \tag{3.45}$$

where $q(x,y|z) = p(x|z)p(y|z)$. We now give the following lemma extending the variational interpretation of mutual information:

Lemma 2 (Conditional Mutual Information) *Let $p(x,y,z)$ be the probability distribution for random variables* x, y *and* z. *Let $Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}}$ be the family of probability distributions on* $(\mathrm{x},\mathrm{y},\mathrm{z})$ *defined as:*

$$Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}} = \{q(x,y,z) = q(x|z)q(y|z)q(z)\} \tag{3.46}$$

Then, the minimum KL-divergence $D(p\|q)$ over $q \in Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}}$ is the conditional mutual information $I_p(\mathrm{x};\mathrm{y}|\mathrm{z})$.

$$I_p(\mathrm{x};\mathrm{y}|\mathrm{z}) = \min_{q\in Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}}} D(p\|q) \tag{3.47}$$

Furthermore, the minimum is uniquely obtained by $q(x,y,z) = p(x|z)p(y|z)p(z)$.

Proof. We decompose the KL-divergence $D(p\|q)$ for $q \in Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}}$ as:

$$\begin{aligned}
D(p\|q) &= E_p\left\{\log\frac{p(\mathrm{x},\mathrm{y},\mathrm{z})}{q(\mathrm{x},\mathrm{y},\mathrm{z})}\right\} && (3.48)\\
&= E_p\left\{\log\frac{p(\mathrm{x},\mathrm{y}|\mathrm{z})p(\mathrm{z})}{q(\mathrm{x}|\mathrm{z})q(\mathrm{y}|\mathrm{z})q(\mathrm{z})}\right\} && (3.49)\\
&= E_p\left\{\log\frac{p(\mathrm{x},\mathrm{y}|\mathrm{z})}{p(\mathrm{x}|\mathrm{z})p(\mathrm{y}|\mathrm{z})} + \log\frac{p(\mathrm{x}|\mathrm{z})}{q(\mathrm{x}|\mathrm{z})} + \log\frac{p(\mathrm{y}|\mathrm{z})}{q(\mathrm{y}|\mathrm{z})} + \log\frac{p(\mathrm{z})}{q(\mathrm{z})}\right\} && (3.50)\\
&= I_p(\mathrm{x};\mathrm{y}|\mathrm{z}) + D(p(\mathrm{x}|\mathrm{z})\|q(\mathrm{x}|\mathrm{z}))
\end{aligned}$$

$$
\begin{aligned}
& +D(p(\mathrm{y}|\mathrm{z})\|q(\mathrm{y}|\mathrm{z}))+D(p(\mathrm{z})\|q(\mathrm{z})) && (3.51)\\
\geq\ & I_p(\mathrm{x};\mathrm{y}|\mathrm{z}) && (3.52)
\end{aligned}
$$

The inequality follows from non-negativity of KL and CKL (which is an expected KL-divergence). Equality occurs if and only if $q(x|z) = p(x|z)$, $q(y|z) = p(y|z)$ and $q(z) = p(z)$ such that $q(x, y, z) = p(x|z)p(y|z)p(z)$.□

The family $Q_{\mathrm{x}\perp\mathrm{y}|\mathrm{z}}$ corresponds to the hypothesis that x and y are conditionally independent given z. The KL-divergence induced by imposing this conditional independency upon an arbitrary model p (by m-projection) is the conditional mutual information between x and y assuming z under model p. This is closely related to our problem of pruning an edge from a graphical model by m-projection. We show this by the following proposition.

Proposition 17 (Lower-bound for Edge-Pruning) *Let $\mathcal{F}$ be an exponential family of graphical models which are Markov with respect to $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$. Let $\mathcal{H}_{\setminus\{\gamma,\lambda\}} \subseteq \mathcal{F}$ be the embedded exponential family which is Markov w.r.t. the embedded graph $\mathbf{G}'_\Gamma = (\Gamma, \mathcal{E}'_\Gamma = \mathcal{E}_\Gamma \setminus \{\gamma, \lambda\})$ with edge $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$ removed. Then, for $\mu \in \mathcal{F}$, the minimum KL-divergence $D(\mu\|\nu)$ over $\nu \in \mathcal{H}_{\setminus\{\gamma,\lambda\}}$ is bounded below by the conditional mutual information between states x_γ and x_λ given the state of the boundary $\mathrm{x}_{\partial\{\gamma,\lambda\}}$.*

$$I_\mu(\mathrm{x}_\gamma; \mathrm{x}_\lambda|\mathrm{x}_{\partial\{\gamma,\lambda\}}) \leq \min_{\nu\in\mathcal{H}_{\setminus\{\gamma,\lambda\}}} D(\mu\|\nu) \tag{3.53}$$

Furthermore, the lower bound is met with equality when $\Lambda = \{\gamma, \lambda\} \cup \partial\{\gamma, \lambda\}$ is a clique in $\mathbf{G}_\Gamma$.

Proof. Let $\mathcal{H} = \mathcal{H}_{\setminus\{\gamma,\lambda\}}$ and $\mathcal{Q} = Q_{\mathrm{x}_\gamma\perp\mathrm{x}_\lambda|\mathrm{x}_{\partial\{\gamma,\lambda\}}}$. The family $\mathcal{H}$ imposes a set of conditional independencies (one for each edge not contained in $\mathcal{E}_\Gamma$) and, hence, corresponds to a restriction of $\mathcal{Q}$ imposing just the single conditional independency associated with the pruned edge. Then,

$$\mathcal{H} \subseteq \mathcal{Q} \Rightarrow \min_{\nu\in\mathcal{H}} D(\mu\|\nu) \geq \min_{\nu\in\mathcal{Q}} D(\mu\|\nu) \tag{3.54}$$

By Lemma 2, the right hand side equals $I_\mu(\mathrm{x}_\gamma; \mathrm{x}_\lambda|\mathrm{x}_{\partial\{\gamma,\lambda\}})$ which proves (3.53). When Λ is a clique, $\mathcal{H}$ doesn't impose any stronger Markov restriction than $\mathcal{Q}$ such that equality occurs. □

In general, computation of the CMI $I(\mathrm{x}_\gamma, \mathrm{x}_\lambda|\mathrm{x}_{\partial\{\gamma,\lambda\}})$ requires inference of the marginal distribution of the neighborhood Λ. Provided the model μ has low tree-width, this should provide a tractable approach for estimating the importance of each edge of μ.

For GMRFs, this computation is greatly simplified since the specific CMI $I(\mathrm{x}; \mathrm{y}|\mathrm{z} = z)$ actually does not vary with z so that the (averaged) CMI may be evaluated as $I(\mathrm{x}; \mathrm{y}|\mathrm{z} = 0)$. Recalling that the partial potential specification $\phi^\Lambda = (h_\Lambda, J_\Lambda)$ of the information representation gives the conditional distribution $p(x_\Lambda|0)$ we see that the CMI between x_γ and x_λ given $x_{\partial\{\gamma,\lambda\}} = 0$ is determined by the local specification $\phi^{\{\gamma,\lambda\}} = (h_{\gamma,\lambda}, J_{\gamma,\lambda})$.

Let $\Lambda = \{\gamma, \lambda\}$ and set $(\hat{x}_{\Lambda|0}, P_{\Lambda|0}) \equiv (J_\Lambda^{-1} h_\Lambda, J_\Lambda^{-1})$. This gives the conditional distribution assuming zero boundary conditions $(\mathrm{x}_\Lambda | x_{\partial\Lambda} = 0) \sim \mathcal{N}(\hat{x}_{\Lambda|0}, P_{\Lambda|0})$. Partitioning $P_{\Lambda|0}$ as $((P_{\Lambda|0})_{\gamma,\gamma}(P_{\Lambda|0})_{\gamma,\lambda}; (P_{\Lambda|0})_{\lambda,\gamma}(P_{\Lambda|0})_{\lambda,\lambda})$, the mutual information $I(x_\gamma; x_\lambda|0)$ associated with this conditional distribution is computed from the conditional covariance $P_{\Lambda|0}$ (see Cover and Thomas [31]) as

$$I(\mathrm{x}_\gamma; \mathrm{x}_\lambda|0) = -\frac{1}{2} \log \left(1 - \frac{\det{(P_{\Lambda|0})_{\gamma,\lambda}}}{\sqrt{\det{(P_{\Lambda|0})_{\gamma,\gamma}} \det{(P_{\Lambda|0})_{\lambda,\lambda}}}} \right) \tag{3.55}$$

This may be computed from the partial canonical correlations $\{\rho_i\}$ between x_γ and x_λ under the conditional covariance $P_{\Lambda|0}$. These are given by the singular values of the matrix $(P_{\Lambda|0})_{\gamma,\gamma}^{-1/2}(P_{\Lambda|0})_{\gamma,\lambda}(P_{\Lambda|0})_{\lambda,\lambda}^{-1/2}$. The mutual information may then be computed by

$$I(\mathrm{x}_\gamma; \mathrm{x}_\lambda|0) = -\frac{1}{2} \sum_i \log(1 - \rho_i^2) \tag{3.56}$$

Also, as shown in Sudderth [125], these coefficients may be computed directly from J_Λ (omitting the inverse) as the singular values of the matrix $-J_{\gamma,\gamma}^{-1/2} J_{\gamma,\lambda} J_{\lambda,\lambda}^{-1/2}$.

In practice we find that this lower-bound estimate given by CMI comes quite close to the actual KL-divergence incurred by m-projection even when the neighborhood of the pruned edge is not complete (which would force equality). We might expect this since $\mu \in \mathcal{F}$ already satisfies all other Markov restrictions imposed by $\mathcal{H}$ such that the m-projection to $\mathcal{H}$ differs little from the m-projection to $\mathcal{Q}$ imposing just $\mathrm{x}_\gamma \perp \mathrm{x}_\lambda | \mathrm{x}_{\partial\{\gamma,\lambda\}}$. Furthermore, when pruning multiple "weak" edges (where the CMI is small for each edge), we find that the sum of CMI values associated to each edge gives a good estimate of the KL-divergence incurred by pruning all edges via a single m-projection. We expect this by the Pythagorean decomposition and by conjecturing that pruning weak edges should not substantially modify the parameters of other edges such that the other CMI values remain reasonably stable under the incremental m-projection approach.

Based on these observations we propose the following approach for pruning edges from a graphical model.

3.3.3 Batch Edge-Pruning Procedure

We now specify a "batch" algorithm for pruning edges from a graphical model. This is an incremental thinning procedure implemented by a sequence of nested m-projections. This generates a sequence of graphical models $\hat{\mu}^{(k)}$ with progressively thinned interaction graphs $\mathbf{G}_\Gamma^{(k)} = (\Gamma, \mathcal{E}_\Gamma^{(k)})$.

At step k, the CMI values associated with each edge $E \in \mathcal{E}_\Gamma^{(k)}$ of the graphical model are evaluated to estimate the KL-divergence which would be incurred by pruning just that edge from the model. We estimate the relative importance of each edge by $\delta_E = I_E / K_E$, the CMI I_E normalized by K_E, the number of model parameters/statistics removed by pruning edge E. The N weakest edges $\mathcal{K} = \{E_k\}_{k=1}^N$ (having the lowest δ_E values) are then selected for removal where $N = |\mathcal{K}|$ is made as large as possible while still satisfying

$$\delta_{\mathcal{K}} \equiv \frac{\sum_{E \in \mathcal{K}} I_E}{\sum_{E \in \mathcal{K}} K_E} < \left(\frac{1 + \log |\mathcal{K}|}{|\mathcal{K}|} \right) \delta \tag{3.57}$$

Note that $\sum I_E$ estimates the KL-divergence incurred by pruning all edges simultaneously and $\sum K_E$ is the total order reduction. Hence, $\delta_{\mathcal{K}}$ estimates the KL-divergence per removed model parameter which is compared to the precision parameter δ of our information criterion $V(\hat{\mu}^{(k)}; \hat{\mu}^{(k+1)})$. In making this comparison we scale δ by $(1 + \log N)/N$ in order to force the batch pruning procedure to be more conservative when pruning many edges at once.[17] This set of edges is then pruned from the graphical model by m-projection of $\hat{\mu}^{(k)}$ to the family $\mathcal{H}^{(k+1)} \subseteq \mathcal{F}$ which is Markov with respect to $\mathbf{G}_\Gamma^{(k+1)} = (\Gamma, \mathcal{E}_\Gamma^{(k+1)})$ where $\mathcal{E}_\Gamma^{(k+1)} = \mathcal{E}_\Gamma^{(k)} \setminus \mathcal{K}$. This procedure terminates when $\delta_E > \delta$ for all remaining edges.

We find that this leads to a procedure which, at first, prunes large batches of very weak edges (where δ_E is much smaller than the precision parameter δ). The size of the "batches" selected on later iterations rapidly decreases until only near-threshold edges remain. These are then pruned one at a time by always selecting the next weakest edge. This allows the effect of pruning weaker edges (by m-projection) to be observed before deciding whether stronger edges (with δ_E near δ) are negligible and should also be removed. Note that each m-projection adjusts the remaining parameters so that the CMI values I_E and apparent strengths δ_E of remaining edges are also adjusted. Hence, some edges with δ_E initially less than δ might actually be retained if earlier m-projections reinforce those interactions. Likewise, some edges with δ_E initially greater than δ might eventually be pruned if earlier m-projections weaken those interactions. We outline this thinning procedure below.

[17] This strategy for selecting the batch size, arrived at by trial and error, is somewhat arbitrary but seems to work well over a variety of test cases.

Batch Edge Pruning:

- *Input.* graphical model μ, family $\mathcal{F}$ containing μ, interaction graph $\mathbf{G}_\Gamma$, threshold δ, tolerance ϵ.
- *Induction Loop.* Set $k = 0$, $\mathcal{H}^{(0)} = \mathcal{F}$, $\hat{\mu}^{(0)} = \mu$ and $\mathbf{G}_\Gamma^{(0)} = \mathbf{G}_\Gamma$. Do the following until termination is indicated in selection step:
  - *Evaluate Lower-Bounds.* For each edge $E = \{\gamma, \lambda\} \in \mathcal{E}_\Gamma^{(k)}$, calculate $I_E = I_{\hat{\mu}^{(k)}}(\mathrm{x}_\gamma; \mathrm{x}_\lambda; \mathrm{x}_{\partial E})$ (in general this requires inference of $\hat{\mu}^{(k)}$, but not for GMRFs).
  - *Select Weak Edges.* Select maximal subset $\mathcal{K} \subseteq \mathcal{E}_\Gamma^{(k)}$ s.t. $\delta_{\mathcal{K}} = \sum I_E / \sum K_E < \delta(1 + \log|\mathcal{K}|)/|\mathcal{K}|$. If $|\mathcal{K}| = 0$, terminate induction loop.
  - *Thin Interaction Graph.* Set $\mathcal{E}_\Gamma^{(k+1)} = \mathcal{E}_\Gamma^{(k)} \setminus \mathcal{K}$ and $\mathbf{G}_\Gamma^{(k+1)} = (\Gamma, \mathcal{E}_\Gamma^{(k+1)})$. Let $\mathcal{H}^{(k+1)} \subseteq \mathcal{H}^{(k)}$ denote subfamily Markov w.r.t. $\mathbf{G}_\Gamma^{(k+1)}$.
  - *M-Project.* Set $\hat{\mu}^{(k+1)} = \arg\min_{\nu \in \mathcal{H}^{(k+1)}} D(\hat{\mu}^{(k)} \| \nu)$ evaluated by loopy iterative scaling version of moment-matching procedure with convergence tolerance ϵ.
  - *Iterate.* Set $k \leftarrow k + 1$ and repeat induction loop.
- *Output.* Thinned graphical model $\hat{\mu}^{(k)} \in \mathcal{H}^{(k)}$ Markov w.r.t. $\mathbf{G}_\Gamma^{(k)}$.

3.4 Simulations

In this section we perform simulations to demonstrate previously discussed methods for moment matching (Section 3.2) and model thinning (Section 3.3) in Gauss-Markov random fields. We first describe four Gaussian test cases which are the basis for subsequent experiments. Then we demonstrate the moment matching approach (Section 3.2) for m-projection to a specified family of *thinned* GMRFs (specified by a thinned interaction graph) and compare the performance of the standard IPF method (Section 3.2.1) to our LIS method (Section 3.2.2). Next, we look at the more general model thinning approach (Section 3.3) where we also select the graphical structure of the thinned model so as to (approximately) minimize our information criterion (Section 3.1) for a specified value of the precision parameter δ. Finally, we introduce the idea of cavity modeling (to be developed further in Chapter 4) and show how our modeling thinning technique, in combination with variable elimination, provides a robust approach to cavity modeling.

3.4.1 Test Cases

We now describe our test cases. To perform controlled model thinning experiments, we construct examples which are "full" Gaussian distributions (where the information matrix $J = P^{-1}$ has many non-zero entries) yet are nevertheless near a lower-order family of GMRFs (such that many entries of the information matrix are near zero). Then, it is natural to consider approximation of the given Gaussian distribution by a lower-order GMRF (by forcing selected entries of the information matrix to zero).

Towards this end, we first posit a *generative model* which, in fact, does have a sparse information matrix. Then we corrupt this model to produce our test case as follows. First, we perform Monte-Carlo simulation to generate a collection of independent samples of this generative model. That is, we simulate N independent, identically distributed samples $x^1, \ldots, x^N \sim g$ where g is the generative model. Calculation of the sample mean and covariance

$$\tilde{x} = \frac{1}{N}\sum_{k=1}^{N} x^k \tag{3.58}$$

$$\tilde{P} = \frac{1}{N}\sum_{k=1}^{N}(x^k - \tilde{x})(x^k - \tilde{x})' \tag{3.59}$$

then provides our *test model* $\mathrm{x} \sim \mathcal{N}(\tilde{x}, \tilde{P})$ which we attempt to thin in later experiments.[18] This test model may be regarded as a noisy version of the generative model. For large sample size N, the test model will be near the generative model with high probability. Yet, due to the finite sample size, the test model will tend to exhibit spurious interactions and will not respect the Markov structure of the underlying generative model. That is, the information matrix of the test model $\tilde{P}^{-1}$ will not be sparse but many elements of this matrix will be near zero. This test model then provides a natural candidate for model thinning[19].

In each of the following four test cases, we specify a generative model which is a zero-mean GMRF having some desired interaction graph with uniform interactions on all edges of the graph. We only consider GMRFs with scalar-valued states. For each generative model, a corresponding test model has been recorded based on $N = 1000$ samples of the generative model. Illustrations of each test model are provided. To indicate the relative strength of interactions in the test model we display the (fully connected) interaction graph but where the apparent intensity of each edge $\{\gamma, \lambda\}$ is set according to the *partial correlation coefficient*[20]

$$\rho_{\gamma,\lambda} \equiv \frac{\mathrm{cov}(\mathrm{x}_\gamma, \mathrm{x}_\lambda | \mathrm{x}_{\partial\{\gamma,\lambda\}})}{\sqrt{\mathrm{var}(\mathrm{x}_\gamma | \mathrm{x}_{\partial\{\gamma,\lambda\}})\mathrm{var}(\mathrm{x}_\lambda | \mathrm{x}_{\partial\{\gamma,\lambda\}})}} \tag{3.60}$$

[18] In fact, this test model is the maximum-likelihood distribution in the full Gaussian family (not imposing any conditional independencies) based on the randomly generated samples $x^1, \ldots, x^N$. This may also be regarded as the m-projection of the empirical distribution $p^*(x) = \frac{1}{N}\sum_k \delta(x - x^k)$ to the full Gaussian family.

[19] In this scenario, model thinning by m-projection to lower-order families of GMRFs also corresponds to maximum-likelihood estimation within the lower-order family.

[20] The partial correlation coefficient is also related to the conditional mutual information $I_{\gamma,\lambda} \equiv I(\mathrm{x}_\gamma; \mathrm{x}_\lambda | \mathrm{x}_{\partial\{\gamma,\lambda\}})$ (3.43,3.56). For a GMRF with scalar-valued states at each site, the conditional mutual information is $I_{\gamma,\lambda} = -\frac{1}{2}\log(1 - \rho^2_{\gamma,\lambda})$ so that the square of the partial correlation coefficient is given by $\rho^2_{\gamma,\lambda} = 1 - \exp\{-2I_{\gamma,\lambda}\}$.

$$= -\frac{J_{\gamma,\lambda}}{\sqrt{J_{\gamma,\gamma}J_{\lambda,\lambda}}} \tag{3.61}$$

which has absolute value less than one. We scale the apparent intensity of each edge in proportion to $\sqrt{|\rho_{\gamma,\lambda}|}$ such that weak interactions, with $\rho_{\gamma,\lambda}$ near zero, appear to fade into the background (the square-root is introduced to adjust the contrast of this edge-rendering approach such that weaker edges are still apparent).

Test Case 1. This test case was generated by a GMRF with 16 sites arranged on a circle, as shown in Figure 3-2 (top left), with interactions between nearest neighbors on the circle. That is, each site is coupled to two other sites, one to either side of that site. The information matrix of the generative model has ones along the main diagonal, -0.4 along the diagonals corresponding to adjacent sites and zeros elsewhere. The (non-zero) partial correlation coefficients in the generative model are $\rho = 0.4$. A randomly generated test case based on this generative model has been recorded and is shown in Figure 3-2. Embedded among spurious interactions, the Markov structure of the underlying generative model is still apparent (top right). The sample mean and covariance specifying the test model are also shown (bottom left and bottom right).

Test Case 2. This test case was also generated by a GMRF with 16 sites arranged on a circle but where we now let each site interact with four neighbors, two to either side of the site as shown in Figure 3-3 (top left). The information matrix of the generative model has ones along the main diagonal, -0.2 at those locations corresponding to interactions, and zeros elsewhere. The (non-zero) partial correlation coefficients in the generative model are $\rho = 0.2$. The interactions, mean and covariance of the randomly generated test model are shown in Figure 3-3. Note that, in this case it is more difficult to pick out the true Markov structure of the underlying generative model.

Test Case 3. This test case was generated by a GMRF with 25 sites arranged on a 5×5 2D grid, as shown in Figure 3-4 (top left), with horizontal and vertical interactions between nearest neighbors in the grid. The information matrix of the generative model has ones along the main diagonal, -0.2 at those locations corresponding to interactions, and zeros elsewhere. The (non-zero) partial correlation coefficients in the generative model are $\rho = 0.2$. The interactions, mean and covariance of the randomly generated test model are also shown in Figure 3-4.

Test Case 4. This test case was also generated by a 5×5 GMRF but where, in addition to horizontal and vertical interactions, we also have diagonal interactions as shown in Figure 3-5 (top left). The information matrix matrix has ones along the main diagonal, -0.15 at those locations corresponding to interactions, and zeros elsewhere. The (non-zero) partial correlation coefficients in the generative model are

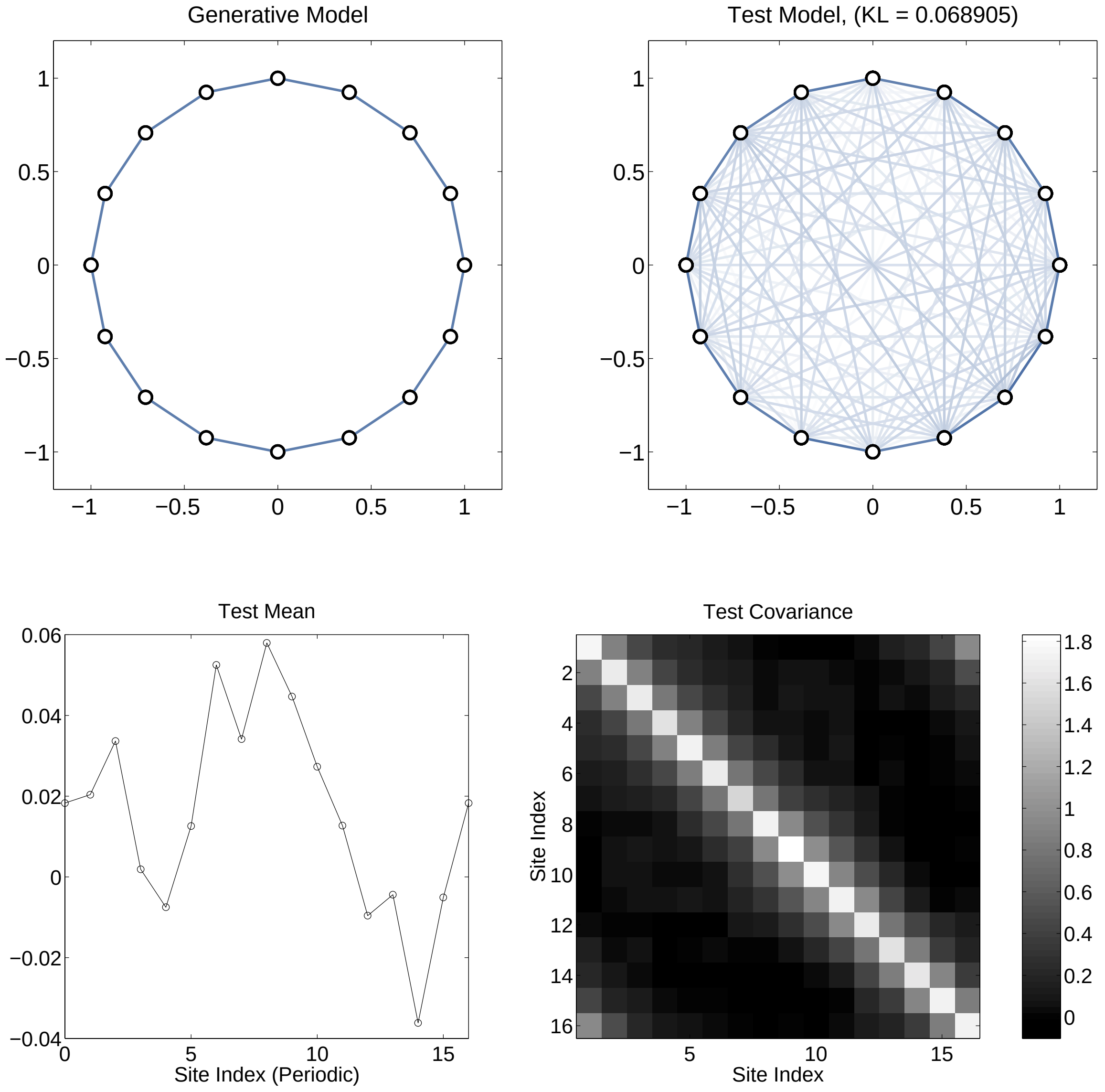


Figure 3-2: Test Case 1: the interaction graph of the underlying generative model (top left), the (fully-connected) interaction graph of the sample-averaged test model (top right), the sample means (bottom left), and the sample covariance (bottom right).

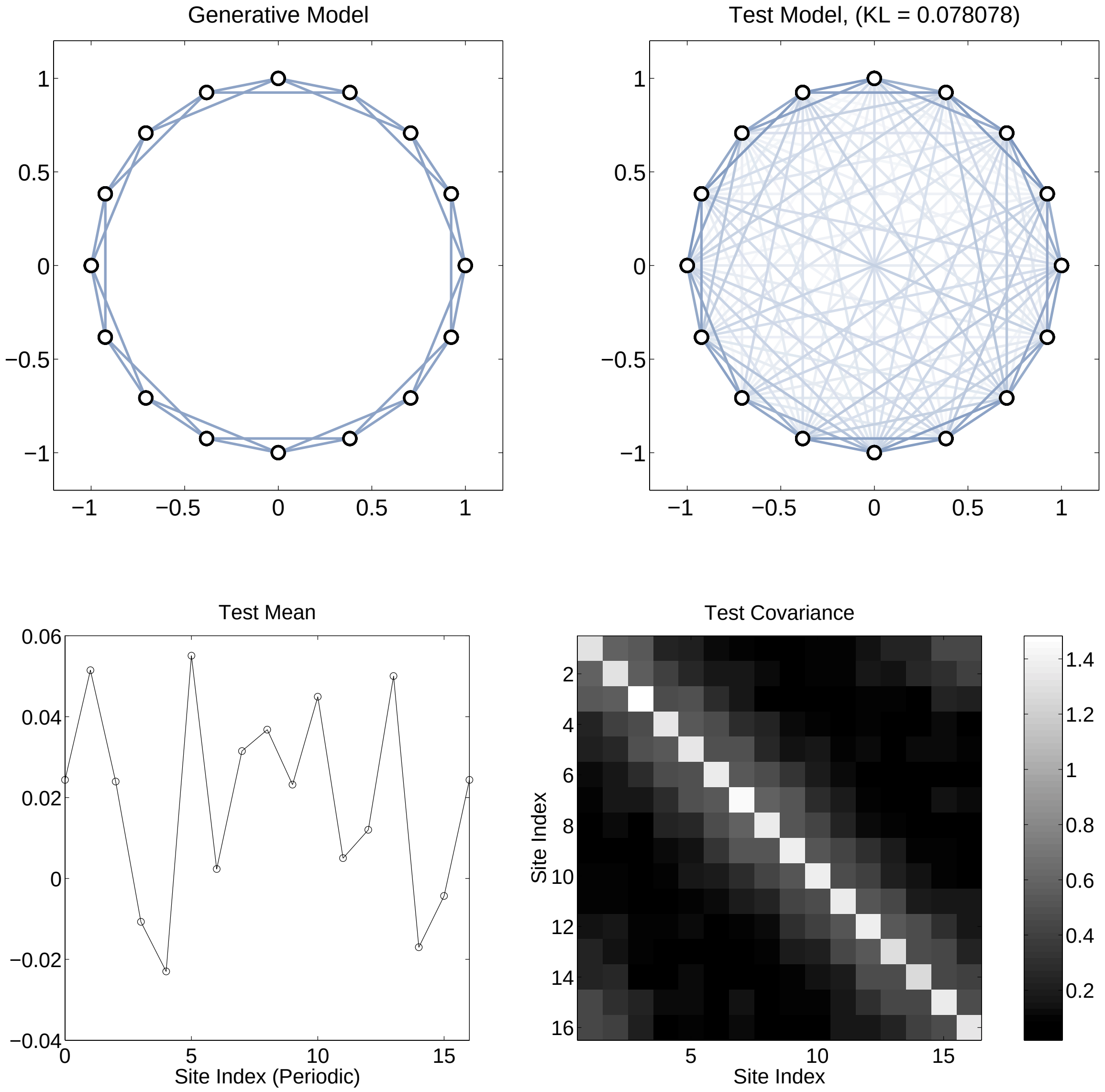


Figure 3-3: Test Case 2: the interaction graph of the underlying generative model (top left), the (fully-connected) interaction graph of the sample-averaged test model (top right), the sample means (bottom left), and the sample covariance (bottom right).

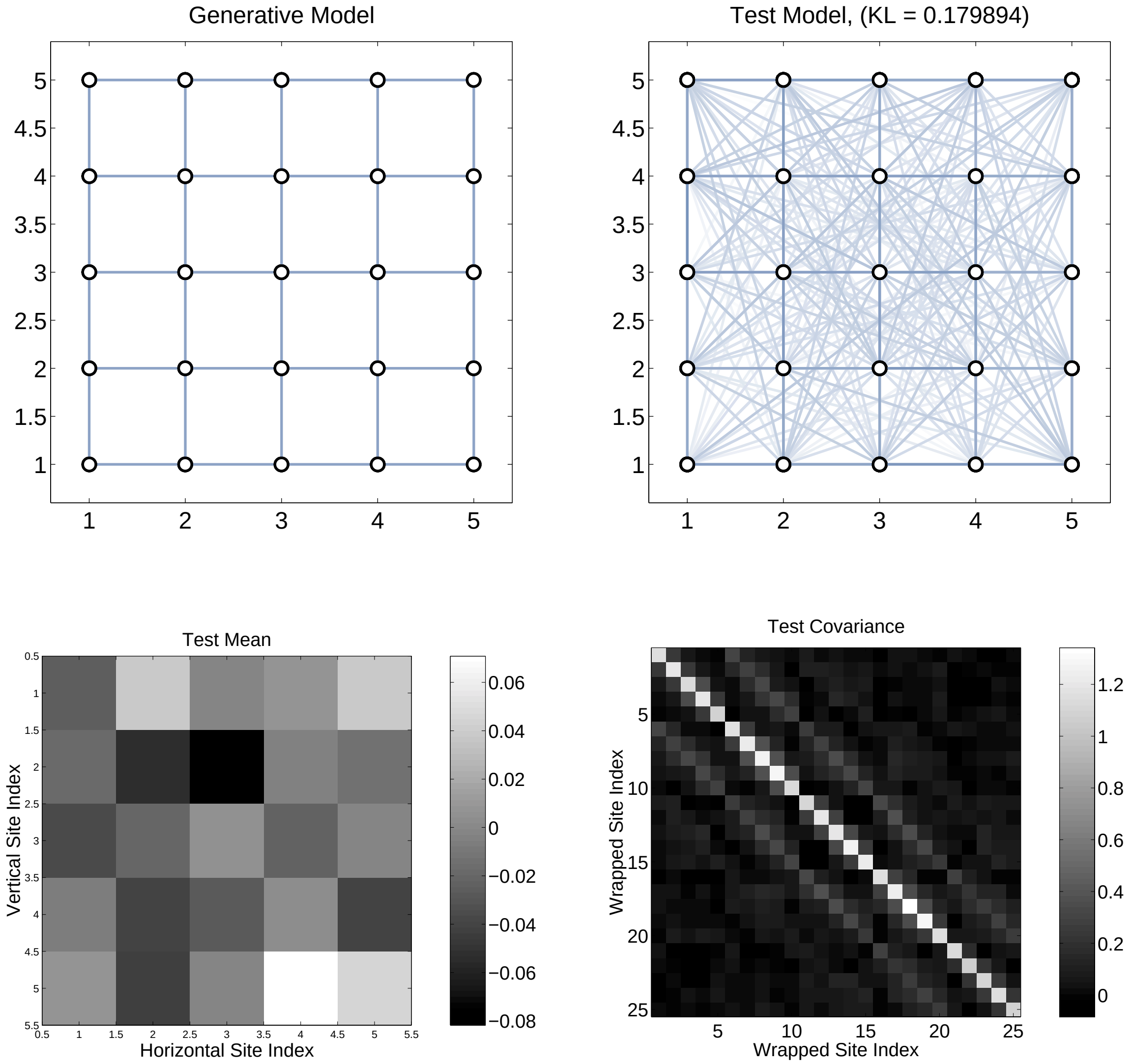


Figure 3-4: Test Case 3: the interaction graph of the underlying generative model (top left), the (fully-connected) interaction graph of the sample-averaged test model (top right), the sample means (bottom left), and the sample covariance (bottom right).

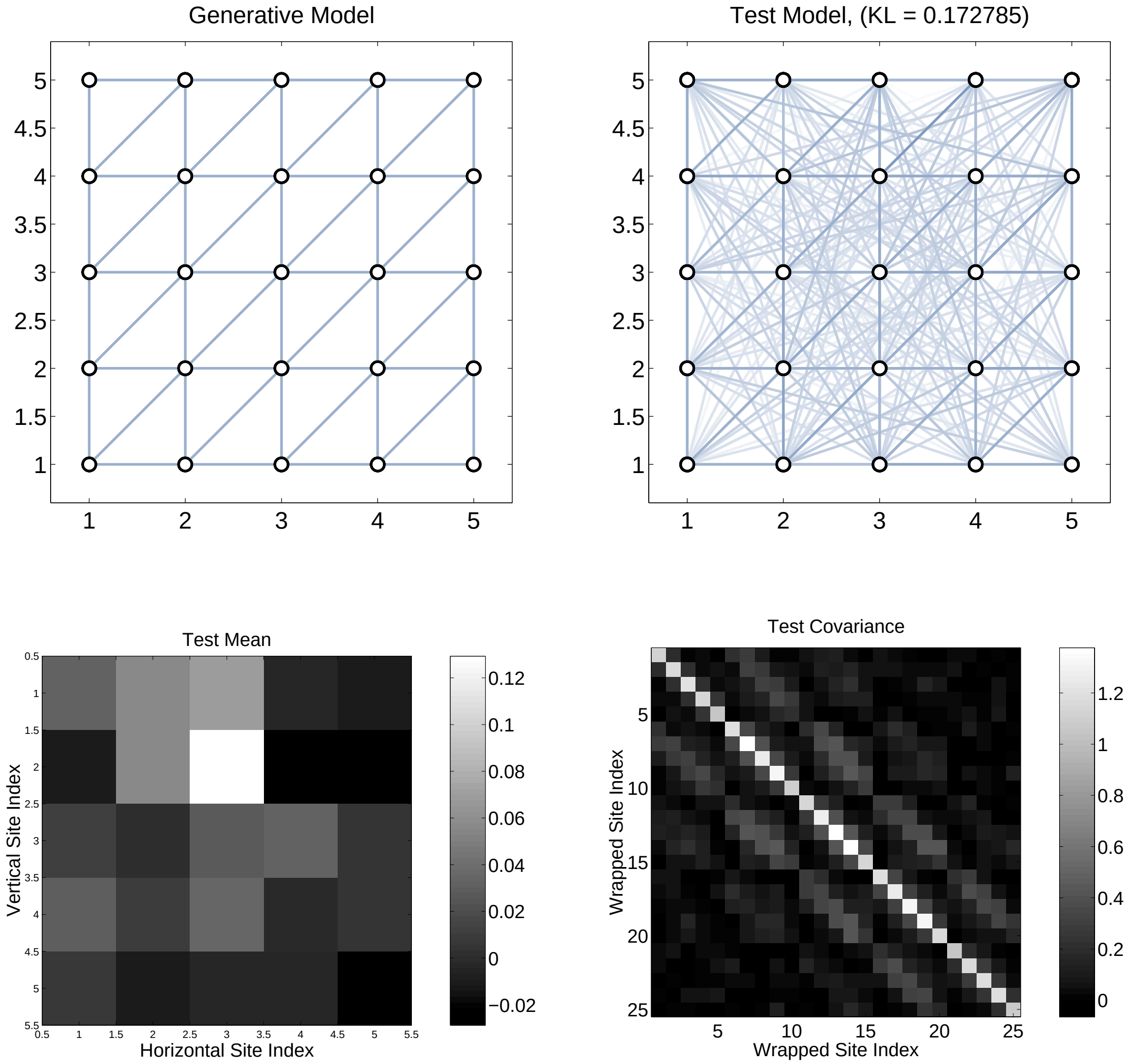


Figure 3-5: Test Case 4: the interaction graph of the underlying generative model (top left), the (fully-connected) interaction graph of the sample-averaged test model (top right), the sample means (bottom left), and the sample covariance (bottom right).

$\rho = 0.15$. The interactions, mean and covariance of the randomly generated test model are shown in Figure 3-5.

Note that, due to sampling noise, all test models have randomly varying means and non-stationary covariance structure.

3.4.2 Moment Matching Experiments

We now demonstrate the moment-matching approach to m-projection and compare the performance of our LIS moment-matching procedure to that of the standard IPF moment-matching procedure.[21] In all four of our test cases, a natural problem to consider is m-projection of the given test model to the family of GMRFs which respect the interaction graph of the underlying generative distribution.

In Section 3.2.1 we considered how m-projection to an exponential family may be posed as moment matching and gave a procedure for m-projection to families of GMRFs employing the IPF moment-matching technique. Recall that in IPF we match moments by iterating over the cliques of the thinned graphical model updating local clique parameters so as to match moments on that clique. While only a local subset of parameters are updated at each iteration, a global inference calculation is nevertheless required in order to calculate this update. In Section 3.2.2 we presented our LIS moment matching approach, a more aggressive alternative to IPF. This approach also performs a global inference calculation, but then updates all parameters simultaneously at each iteration. Essentially, this is accomplished by summing the IPF updates. But when cliques overlap we "correct" for this by subtracting off (or adding) the appropriately scaled IPF update on the intersection such that each clique is "counted" just once. However, with our more aggressive approach, we can no longer formally guarantee convergence of the method and some empirical investigation is warranted. Also, this allows comparison to the standard IPF method.

We now show the result of applying both methods, IPF and our LIS method, to perform the indicated m-projection in each of our four test cases. In the IPF approach, we iterate over just the maximal cliques of the interaction graph. In Test Case 1 and 3, these are just the edges of the graph. In Test Case 2 and 4, each maximal clique contains 3 sites. In the LIS approach, we construct our updates with respect to maximal cliques but also must consider intersections of maximal cliques, intersections of intersections and so forth. Hence, in all of our examples, this means that the LIS approach calculates IPF updates for *all* cliques of the graph (not just the maximal cliques) and fuses these updates in the manner described in Section 3.2.2.

In both approaches, we measure the discrepancy between the current moments (of the graphical model being adjusted) and the desired values of these moments (taken from the test model) as follows. For each maximal clique Λ, we calculate the (normalized) marginal KL-divergence $d_\Lambda = D(p^*(\mathrm{x}_\Lambda)\|p(\mathrm{x}_\Lambda))/K_\Lambda$ where K_Λ is the number of model parameters associated with clique Λ, $p(\mathrm{x}_\Lambda)$ is the current marginal

[21] We initially intended to also compare with the GIS technique, but actually found that convergence of IPF was more rapid than in GIS in all four of our test cases. Hence, only IPF is shown here. We have not implemented the IIS technique but should like to also make this comparison in the future.

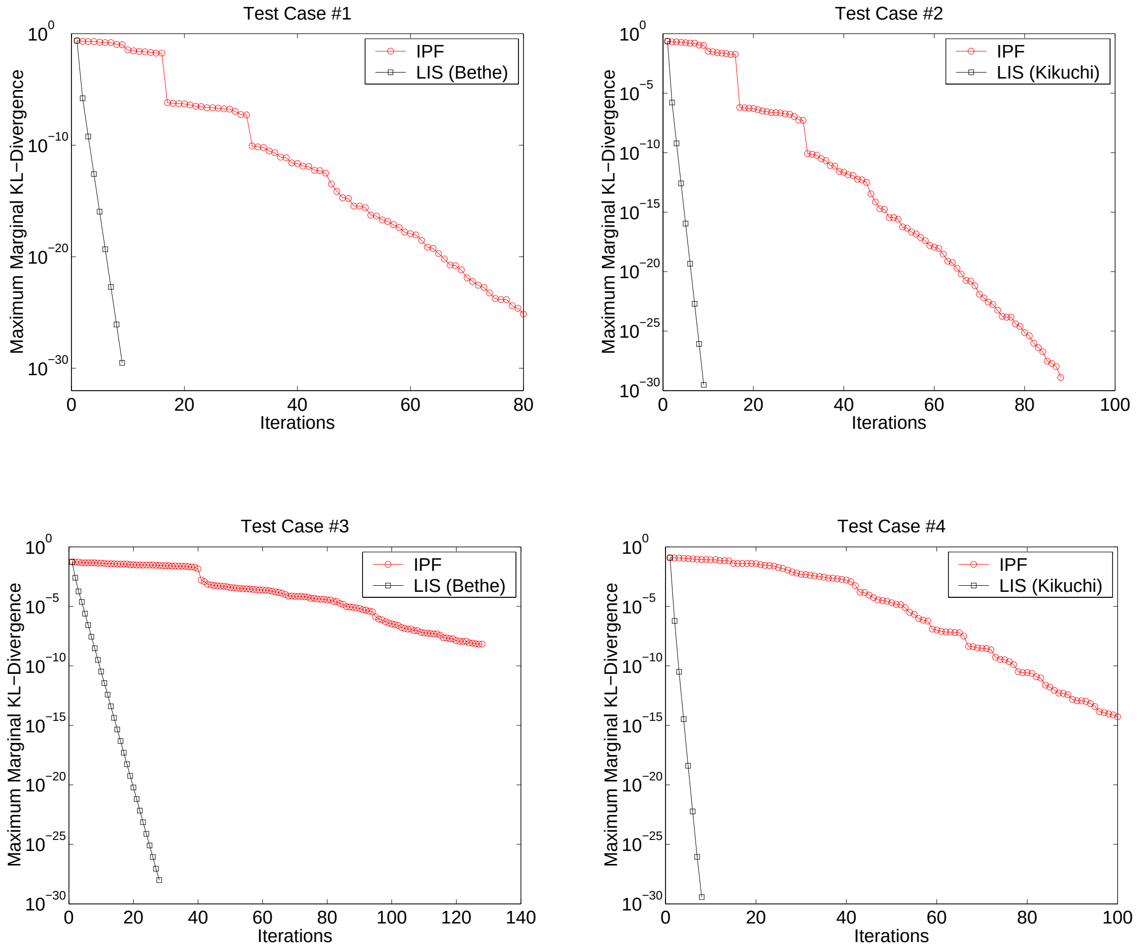


Figure 3-6: Plots showing convergence of iterative moment matching procedures. Both standard IPF (iterative proportional fitting) and our LIS (loopy iterative scaling) are shown. LIS is based on either the Bethe approach (Test Case 1 and 3) or the Kikuchi approach (Test Case 2 and 4).

distribution of the thinned model and $p^*(\mathrm{x}_\Lambda)$ is the marginal distribution of the test model. We then take $\hat{d} = \max_\Lambda d_\Lambda$ as our measure of discrepancy. In the IPF approach, the clique with the largest d_Λ value is updated at each iteration. In both approaches, moment matching continues until all d_Λ values are less than a specified tolerance[22] $\epsilon = 10^{-28}$. In Figure 3-1 we show the convergence of both methods in all four test cases by plotting the maximum (normalized) marginal KL-divergence $\hat{d}$ after each iteration of either IPF or LIS.

Note that, in all four test cases, the LIS method not only converges, but converges much more quickly (requiring fewer iterations) than the IPF method. In fact, roughly speaking, one iteration of LIS appears to give a comparable error reduction as one

[22] This matches moments nearly to machine precision. In practice, such a small tolerance is probably of no benefit. Larger tolerances of $10^{-6} - 10^{-12}$ would probably suffice.

entire pass of iterative proportional fitting (by a "pass" we mean however many iterations of IPF that are required to visit each maximal clique). Yet the computation per iteration of both methods is comparable[23]. We also note that the convergence of LIS appears to be linear in our semi-log plots, so that the maximum marginal KL-divergence vanishes exponentially. The rate at which errors dissipate depends upon the example.

To explain the excellent performance of LIS, we suggest the following interpretation. In Test Case 1, we note that the interaction graph has the form of a single "long loop" such that the "loopy" character of this model is perhaps negligible. Hence, performing one iteration of loopy belief propagation, which is exact after only one iteration in non-loopy graphs (trees and chains), is almost exact for such long loops and converges very quickly in this case. A similar interpretation holds for Test Case 2. In this example, there are actually many shorter loops which presumably are *not* negligible. However, by clustering maximal cliques in the Kikuchi approach, these shorter loops are then embedded within those clusters where LIS computes exact IPF updates. The only "loopiness" not explicitly captured in this computation is again the global "long loop" structure of the interaction graph. Hence, the Kikuchi LIS updates are very nearly exact and again give rapid convergence. We are surprised, however, to find that LIS still seems to do quite well even in Test Case 3 and 4.[24] Since these examples both have many shorter loops which are *not* embedded in maximal cliques, it is less clear that we should expect this good performance. At this time, we cannot adequetly explain the good performance of LIS in such cases and think that further analysis is warranted.

We remark that, in all the experiments we have performed thus far (in addition to the examples given here), we have found that, so long as LIS is performed with respect to the *maximal cliques* (as in these experiments), the LIS approach appears to be stable and converges quickly. This is the case even when the approach is seeded with the fully factored approximation where all interactions are initially neglected. It is also possible to perform the edge-based Bethe approach even when the graphical model has higher-order cliques than edges (such as in Test Case 2 and 4). However, we cannot recommend this latter approach as we find that convergence is typically slower and sometimes, when there exist strongly-coupled higher-order cliques, the method may actually become unstable. Hence, we recommend that LIS always be performed with respect to maximal cliques (although weakly-coupled maximal cliques could perhaps be excluded without ill effect). This, however, limits the applicability of the method to graphical models where the maximal cliques are small. Yet, this is not really so great a restriction as these are precisely the types of graphical models we wish to consider in the context of model thinning. Besides, the requisite moment calculation already restricts the class of tractable models to those having thin interaction graphs (where the interaction graph can be triangulated while keeping maximal cliques small). In these cases at least, LIS appears to provide a tractable and reliable approach to moment matching.

[23] The computation per iteration in LIS is roughly twice that of IPF in these test cases.

[24] Although, the rate of convergence is somewhat slower in Test Case 3.

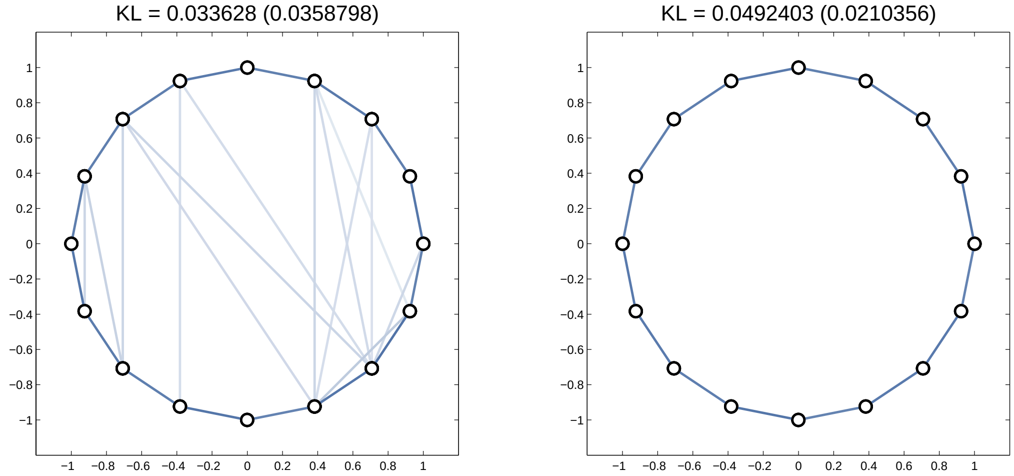


Figure 3-7: Model thinning in Test Case 1. Starting from the fully-connected test model (shown at top-right in Figure 3-2), our batch edge-pruning procedure thins the model in two steps. The first m-projection (left) prunes the majority of weak interactions in the model, the second m-projection (right) prunes all but the "true" interactions present in the generative model (top-left Figure 3-2). Further edge-pruning m-projections with information loss less than δ are not possible.

3.4.3 Model Thinning Experiments

We now turn our attention to the more general model thinning approach where we adaptively select the graphical structure of the thinned model (rather than presupposing some interaction graph as we did in the preceding moment matching experiments).

In these experiments, we first convert all four test models into the information form $(h, J) = (\tilde{P}^{-1}\tilde{x}, \tilde{P}^{-1})$. We then perform the model thinning procedure, described in Section 3.3, to incrementally thin this (initially fully connected) graphical model by a series of m-projections to a sequence of nested families of GMRFs. Each m-projection is performed using the LIS moment-matching technique (but IPF could be used instead). Lower-order families of GMRFs are specified by selecting a set of edges, corresponding to weak interactions, to prune from the graphical model. Deleting these edges from the graphical model by m-projection corresponds to releasing a set of moment constraints and maximizing entropy subject to a reduced set of moment constraints. This m-projection is actually performed by moment matching within the lower-order family of GMRFs which are Markov with respect to the thinned interaction graph. This moment matching procedure is seeded simply by setting to zero those entries of the information matrix corresponding to pruned edges. This however, perturbs the remaining moments of the model which are then enforced by iterative moment matching.

We demonstrate this model thinning approach in our four test cases by displaying the interaction graphs of each m-projection to a lower order family. Since edges are selected for pruning based on the conditional mutual information, $I_{\gamma,\lambda} = -\frac{1}{2}\log(1 - \rho^2_{\gamma,\lambda})$, we again indicate the strength of interaction by rendering edges with apparent

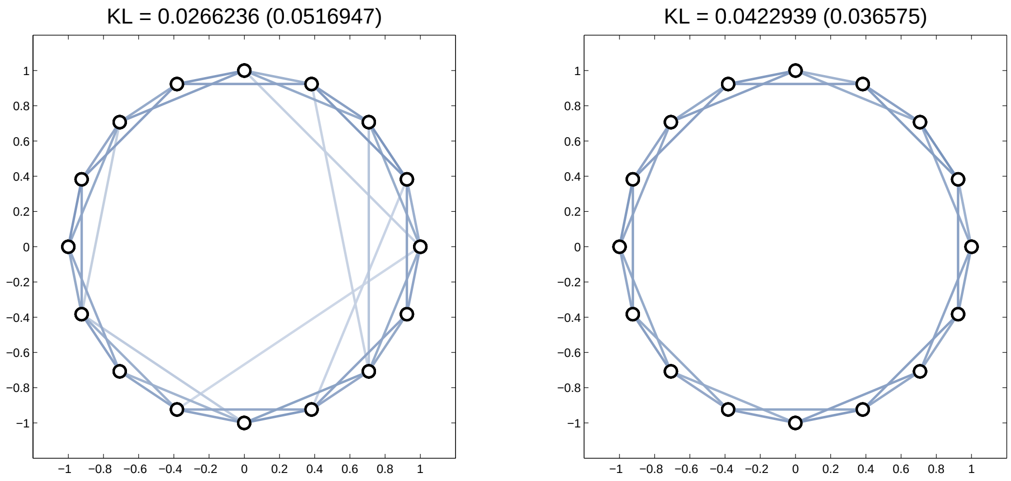
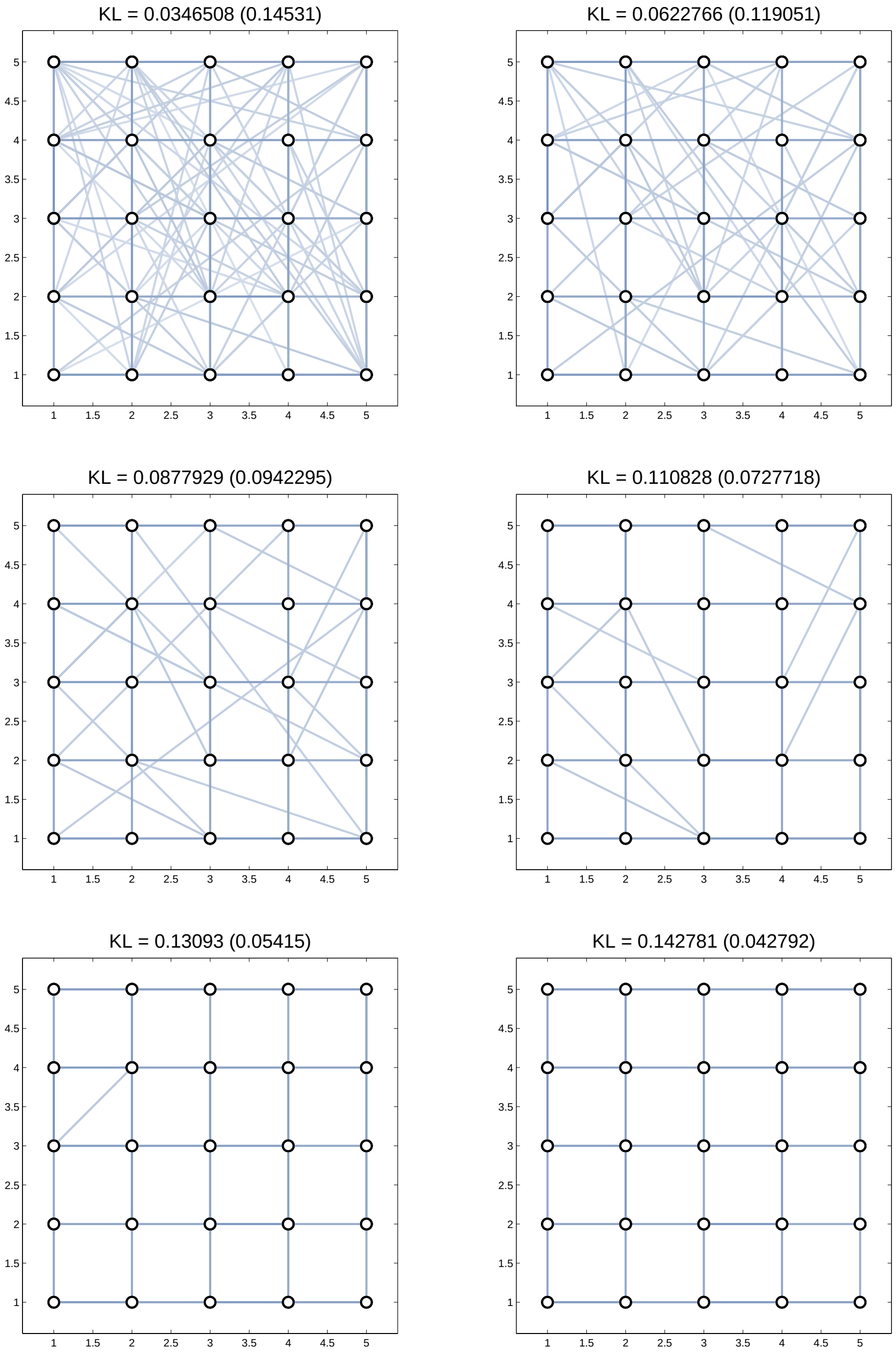


Figure 3-8: Model thinning in Test Case 2. Starting from the fully-connected test model (top-right in Figure 3-3), our batch edge-pruning procedure again thins the model in two steps. The first m-projection prunes the majority of weak interactions in the model, the second m-projection prunes all but the "true" interactions present in the generative model (top-left Figure 3-3).

intensity proportional to $\sqrt{|\rho_{\gamma,\lambda}|}$. The sequence of thinned models (m-projections) are shown in Figures 3-7, 3-8, 3-9, and 3-10. Thinning in all four test cases is performed for precision parameter $\delta = 0.05$ and moment matching tolerance $\epsilon = 10^{-10}$. Also shown, for each m-projection, is the KL-divergence $D(\mu \| \hat{\mu}^{(k)})$ of the thinned model $\hat{\mu}^{(k)}$ relative to the original test model μ (also shown, in parentheses, is the KL-divergence $D(g \| \hat{\mu}^{(k)})$ of the thinned model relative to the generative model).

Note that, in three out four of our test cases, the actual graphical structure of the generative distribution is recovered. In Test Case 4, one edge present in the generative model was omitted in the final thinned model. This is, of course, a function of the test case and such close correspondence need not always occur.[25] We have also observed that, during the moment matching procedure, the "intensity" of spurious interactions tend to fade while the strength of true interactions are reinforced. This suggests the advantage of our incremental approach to edge pruning over the simpler approach where all edges to be pruned are selected by a single initial threshold comparison. It is also interesting to note that while the KL-divergence of the thinned model relative to the test model is increasing (as it must), the KL-divergence relative to the generative model is actually decreasing. We should expect to see this trend in view of the connection to the AIC which favors lower-order models in order to reduce the expected KL-divergence relative to the generative model by virtue of having fewer

[25] In particular, decreasing N in these experiments makes it more difficult to distinguish "true" interactions (present in the generative model) from spurious interactions (arising due to the finite sample size). Then, model thinning is more apt to produce an interaction graph differing from that of the generative distribution (either missing some true interactions or including some spurious interactions).


KL = 0.0346508 (0.14531)

KL = 0.0622766 (0.119051)

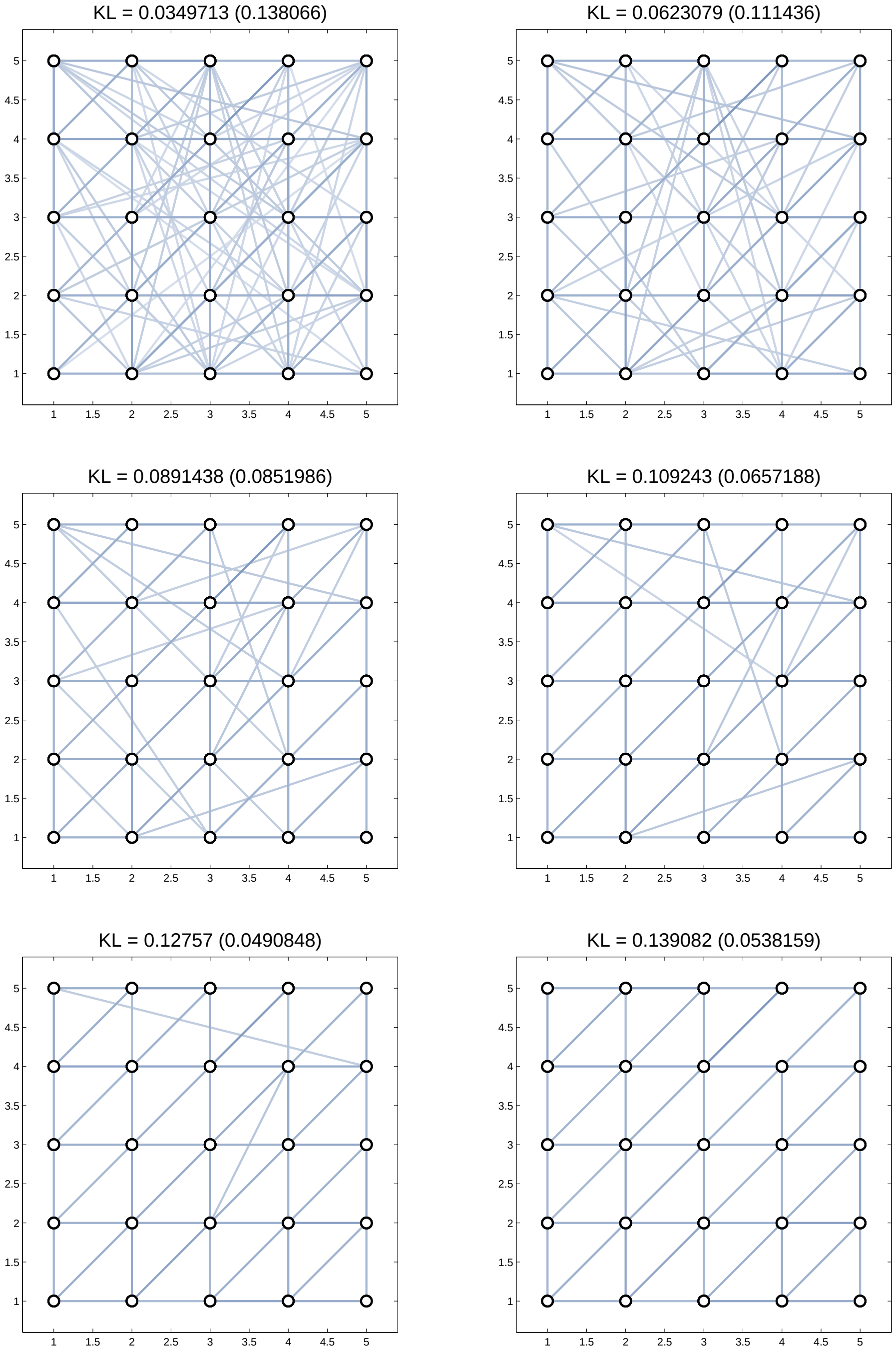
KL = 0.0877929 (0.0942295)

KL = 0.110828 (0.0727718)

KL = 0.13093 (0.05415)

KL = 0.142781 (0.042792)

Figure 3-9: Model thinning in Test Case 3. Starting from the fully-connected test model (top-right Figure 3-4), six m-projections were required to thin the model. Note that earlier m-projections prune larger batches of very weak interactions and later m-projections prune smaller batches of more significant interactions. We again recover the interaction graph of the generative model (top-left Figure 3-4).



KL = 0.0349713 (0.138066)

KL = 0.0623079 (0.111436)

KL = 0.0891438 (0.0851986)

KL = 0.109243 (0.0657188)

KL = 0.12757 (0.0490848)

KL = 0.139082 (0.0538159)

Figure 3-10: Model thinning in Test Case 4. Starting from the fully-connected test-model (top-right Figure 3-5), six m-projections were required to thin the model. In this case, we almost recover that same interaction graph as in the generative model (top-left Figure 3-5) but there is one “mistake” in that a single edge present in the generative model is missing in the thinned model.

parameters to estimate. This shows the advantage of some model thinning in the presence of model uncertainty (that is when the available test model is not exact but rather is a noisy observation of the unknown generative model). This effect is most pronounced when the generative model does in fact have sparse Markov structure. Yet, as we will see in the following subsection, model thinning can also prove beneficial in this regard even when the generative model is actually fully connected but where many of the these interactions are weak.

3.4.4 Prelude to Cavity Modeling

In this final experiment, we introduce a preliminary notion of *cavity modeling* and show how our model thinning technique proves useful in this context. Consider the 5×5 GMRF we constructed for Test Case 3. Suppose, rather then attempting to design a thinned approximation for the entire 5×5 field, we only wish to select a thinned model for the surface of this 5×5 GMRF (for the moment, let us just consider the surface to consist of those 16 site in the outer most "square" going around the perimeter of the grid). We call such an approximation a *cavity model* to suggest that we are approximating the cavity left after removing those vertices in the interior (not in the surface). We may construct such a cavity model by combining model thinning and variable elimination in various ways. We perform several experiments which indicate the robustness and flexibility of this approach.

First, for the sake of comparison, let us perform variable elimination with respect to the generative model. This procedure is shown in Figure 3-11. Let us call this the *truth cavity model*. Note that the fill edges arising in variable elimination produces a fully-connected cavity model. Yet, many of those fill edges correspond to weak interactions having partial correlation coefficients (and conditional mutual information values) near zero. This suggests we might do nearly as well with a thinned cavity model for the surface of this GMRF. Let us consider two paths for constructing such a thinned cavity model starting from our (noisy) test model displayed previously in Figure 3-4. In the following, all model thinning is performed adaptively with precision parameter $\delta = 0.005$ and moment-matching tolerance $\epsilon = 10^{-12}$.

Eliminate-Thin. In the first approach, we perform variable elimination of the (fully connected) test model as shown in Figure 3-12. Because the test model is fully connected, we eliminate all nodes at once which requires significantly more computation than in the incremental variable elimination approach shown for the sparse generative model. Let us call this the *test cavity model*. Next, we thin this test cavity model producing the *eliminate-thin test model* also shown in Figure 3-12.

First note that variable elimination reduces the KL-divergence relative to the generative model. That is, the KL-divergence of the test cavity model (relative to the truth cavity model) is smaller than the KL-divergence of the original test model (relative to the generative model).[26] Second, observe that while the KL-divergence

[26] In fact, this is always the case due to the chain rule for KL-divergence, $D(p(x,y)\|q(x,y)) = D(p(x|y)\|q(x|y)) + D(p(y)\|q(y))$, so that $D(p(y)\|q(y)) \leq D(p(x,y)\|q(x,y))$. That is, marginaliza-

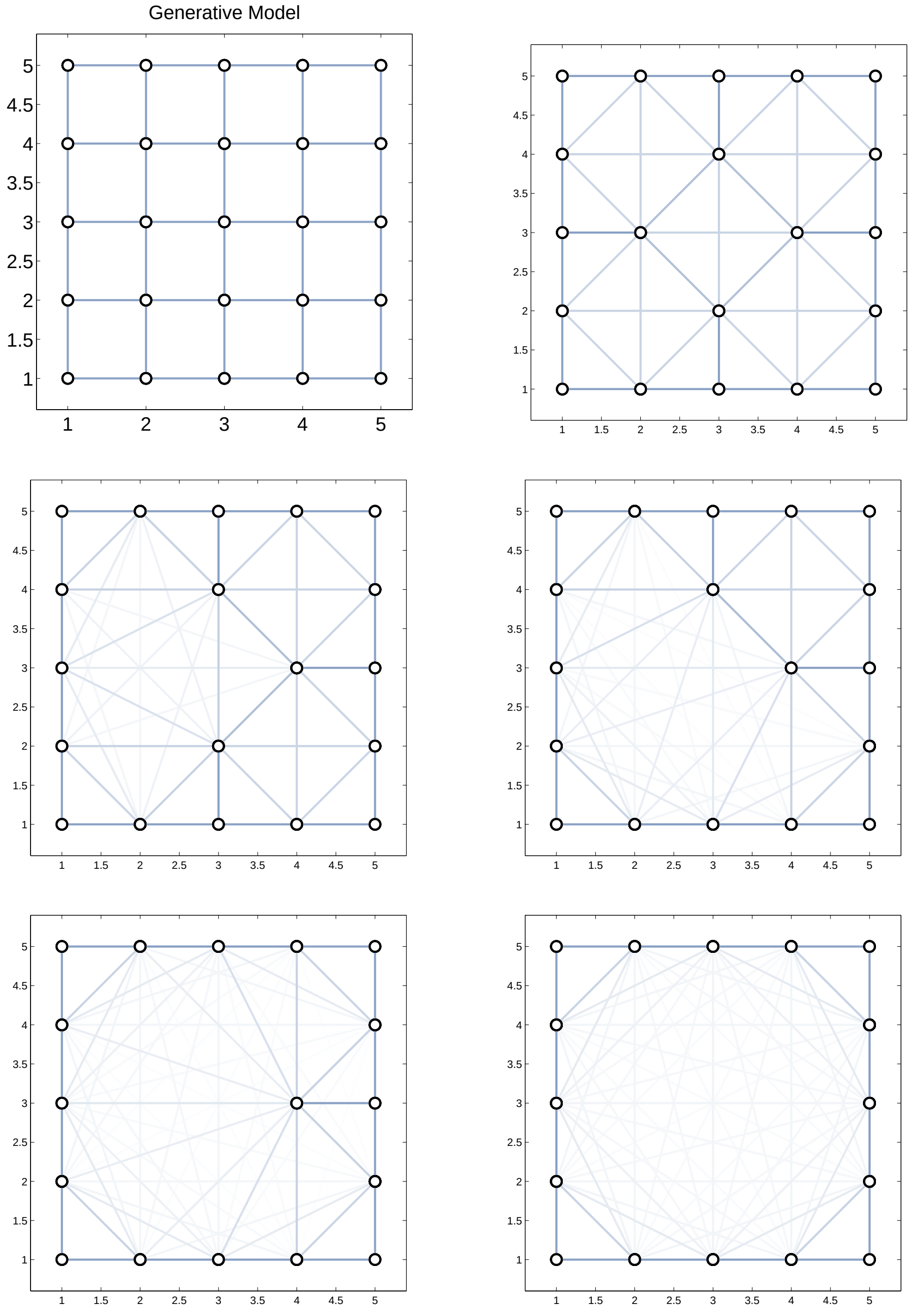


Figure 3-11: Variable elimination of *generative model* (top left) produces the fully connected *true cavity model* (bottom right).

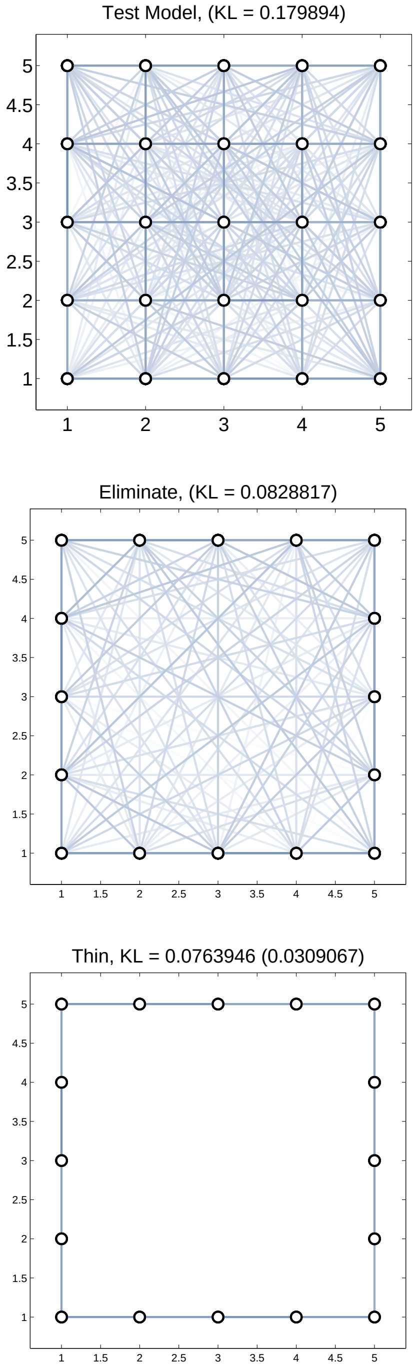


Figure 3-12: Variable elimination of the *test model* (top) produces the fully connected *test cavity model* (middle). Model thinning then yields the *thinned test cavity model* (bottom). At bottom, the KL-divergence of the thinned test cavity model relative to the (fully connected) test cavity model is shown. KL-divergences relative to the generative model are also shown in parentheses in all three figures.

incurred by thinning is significant (comparable to the KL-divergence of the test cavity model relative to the truth cavity model), the actual KL-divergence from the truth cavity model is again actually decreased by model thinning. This is the case even though the truth cavity model is fully connected. This may seem surprising, but we might expect to see this trend in view of the Akaike interpretation of our model thinning metric (as attempting to minimize the expected KL-divergence from the unknown generative model by projection to nearby lower-order families). These two effects together speak favorably for the robustness of our method when the given test model itself contains some noise or modeling errors. The main disadvantage of this approach is the expense of variable elimination and also of inference for the fully connected test cavity model (required by moment matching). This lead us to the question of whether an additional model thinning step could be introduced without ill effect, while at the same time reducing overall computational load. The following experiment explores this idea.

Thin-Eliminate-Thin. As a second approach for obtaining a thinned cavity model, let us consider what happens if we first thin the (full) test model and then repeat the above eliminate-and-thin procedure. That is, we start with the *thinned test model* obtained as in the model thinning experiments of the previous section. Then, we perform variable elimination of this thinned test model producing the *thin-eliminate test model*. This closely resembles variable elimination of the generative model and requires less computation then variable elimination in the full test model. Finally, we thin the thin-eliminate cavity model producing the *thin-eliminate-thin test model*. This procedure is illustrated in Figure 3-13.

Again, variable elimination reduces the actual KL-divergence relative to the generative model. In this case, however, the model thinning step does actually increase the KL-divergence relative to the generative model, but this increase is not very large (less than the KL-divergence of the preceding thin-eliminate test model relative to the truth cavity model). This occurs because the thinned test model is such a good fit to the actual generative model in this example.

Note also that the final KL-divergences of both methods, eliminate-thin and thin-eliminate-thin, relative to the truth cavity model are the same (0.031). In fact, the KL-divergence of the thin-eliminate-thin model relative to the eliminate-thin model is only about 9.14×10^{-13} (near the moment matching tolerance $\epsilon = 10^{-12}$ used in these experiments). This is no accident. Since both variable elimination and model thinning preserve moments of the model, these two approaches would actually have given *exactly* the same cavity model were our moment matching procedure exact. This holds more generally as long as both methods select the same set of final statistics (edges) for the thinned cavity model and none of these correspond to statistics (edges) which were thinned from the model at some point and then later reintroduced by variable elimination. In this regard, it seems that we may introduce as many intermediate model thinning steps as we like (to control the computational complexity of variable elimination) and this has little or no effect on the final cavity model.

tion reduces the KL-divergence between the generative model p and the test model q.

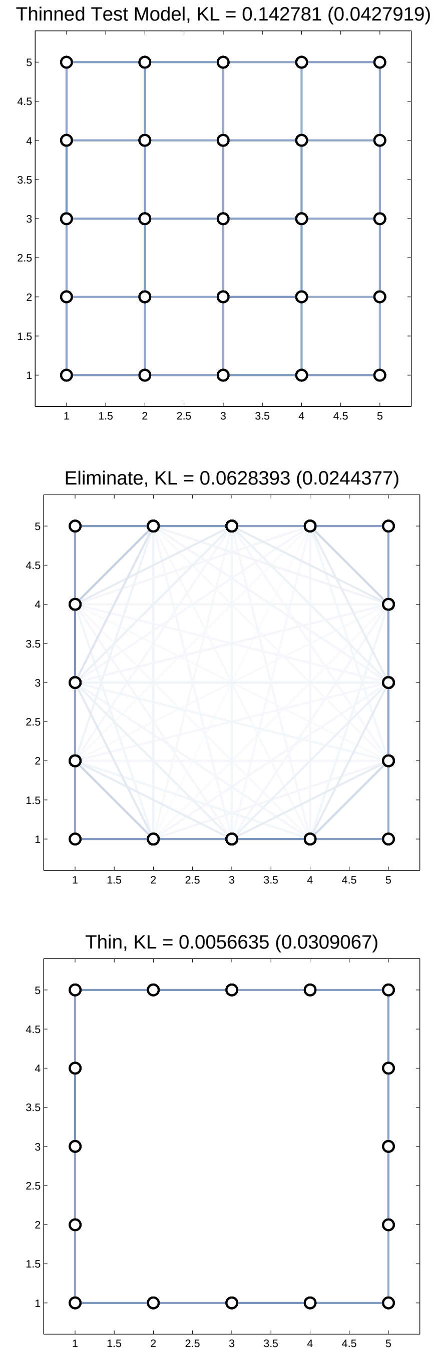

Thinned Test Model, KL = 0.142781 (0.0427919)
Eliminate, KL = 0.0628393 (0.0244377)
Thin, KL = 0.0056635 (0.0309067)
1
1.5
2
2.5
3
3.5
4
4.5
5

Figure 3-13: We begin with the *thinned test model* (top) duplicated from a previous model-thinning experiment (bottom right Figure 3-9). Variable elimination produces the fully connected *thin-eliminate test model* (middle). Further model thinning yields the *thin-eliminate-thin test model* (bottom). Respectively, the indicated KL-divergences are relative to the: test model (generative model), test cavity model (truth cavity model), thin-eliminate truth model (truth cavity model).

Finally, we note that we have also performed model thinning for the truth cavity model and found this produces a *thin-eliminate truth model* having the same "square" interaction graph as in our two test cavity models and has KL-divergence of 0.007 relative to the (fully connected) truth cavity model. In comparison, our test cavity models both have KL-divergence 0.031 from the (fully connected) truth cavity model (about five times larger than the minimal achievable KL-divergence for this graphical structure). Hence, in this test case at least, our test cavity models appear to come reasonably close to giving as good a thinned approximation as possible even though these are constructed from the noisy test model. Also, the KL-divergence of both test cavity models relative to the thin-eliminate truth model is 0.024 which is less than the KL-divergence 0.031 relative to the truth cavity model. It seems that performing model thinning for both test and truth models tends to reduce KL-divergence much as in variable elimination.

There are several lessons to be learned from these experiments. First, variable elimination always reduces KL-divergence such that differences between our test model and the true (unknown) generative model are reduced by variable elimination. Second, while model thinning increases the KL-divergence relative to the test model, the actual KL-divergence (relative to the unknown generative distribution) is often decreased (or increases only slightly in comparison to the observed KL-divergence). This effect is even more pronounced if we also thin the generative model. While this need not always occur, this seems to be a result of our information criterion which attempts to choose the model order (number of edges) so as to minimize the expected KL-divergence from the (uncertain) generative model. Finally, we may employ any combination of variable elimination and model thinning and we tend to obtain essentially the same cavity model. For all of these reasons, it would appear that the method of combining variable elimination with adaptive model thinning provides a robust approach to cavity modeling.

This, fundamentally, is the idea underlying the recursive cavity modeling approach developed in the following chapter. However, our model thinning procedure requires an inference subroutine to calculate those moments being matched during each m-projection. This means that optimal model thinning is only tractable when inference is tractable. Hence, some method is needed to "bootstrap" our cavity modeling approach for such intractable graphical models. In the following chapter, we discuss a recursive approach to cavity modeling which employs a recursive "divide-and-conquer" method for constructing cavity models in otherwise intractable models. This, in turn, supports a tractable approach to approximate inference which we demonstrate for GMRFs.

Chapter 4

Recursive Cavity Modeling

This chapter presents the recursive cavity modeling (RCM) approach for tractable inference of GMRFs. The method combines the information-projection model thinning procedures of the previous chapter with a global divide-and-conquer recursive inference scheme such as in multiscale modeling or the junction tree approach. The main idea is to employ model thinning to select compact yet faithful graphical models for the "Markov blankets" arising in the variable elimination approach to inference. In view of the information-projection interpretation of the model thinning procedure, we expect this approach to give very reliable inference of marginal distributions. Some experiments are given at the end of the chapter which appear to support this expectation. While we focus mainly on Gaussian MRFs, the basic framework outlined here should apply more generally.

The first section details the basic two-pass RCM approach and discusses the interpretation of information projections in this context. The main procedures are outlined and diagrams are also given to illustrate these procedures and to provide some intuition about the role "Markov blanket" models play in this approach to inference. Section 4.2 then discusses two iterative extensions intended to refine the approximations made during two-pass RCM. Section 4.3 presents some simulations to motivate and clarify these methods and to examine the performance of RCM both in terms of reliability of inference and scalability of computation.

4.1 Two-Pass RCM

This section describes the two-pass RCM approach for approximate inference of the marginal distributions of a Gauss-Markov random field (GMRF). The input for this procedure is a graphical model for the GMRF in the information form $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$ as discussed in Section 2.1.5. This is represented by a hypergraph $\mathbf{H}_\Gamma^\phi = (\Gamma, \mathcal{H}_\Gamma^\phi)$ based on the sites of the field Γ. The hyperedges $\mathcal{H}_\Gamma^\phi$ indicate interaction potentials $\phi_\Lambda(x_\Lambda)$ for $\Lambda \in \mathcal{H}_\Gamma^\phi$. For GMRFs, all interactions are either singleton (involving just one site)

$$\phi_\gamma(x_\gamma) = -\frac{1}{2}x_\gamma' J_\gamma x_\gamma + h_\gamma' x_\gamma \qquad (4.1)$$

or pairwise (involving two sites)

$$\phi_{\{\gamma,\lambda\}}(x_\gamma, x_\lambda) = -x_\gamma J_{\gamma,\lambda} x_\lambda. \tag{4.2}$$

The Markov structure of the field is then specified by the pairwise interactions as $\mathbf{G}_\Gamma^\phi = \text{adj } \mathbf{H}_\Gamma^\phi$.

As shown in Section 2.1.5, these are the canonical potentials (Section 2.1.3) relative to the zero ground-state $x^* = 0$. The conditional distribution $p(x_\Lambda | \mathrm{x}_{\partial\Lambda} = 0)$ of each subfield $\Lambda \subset \Gamma$ assuming zero boundary conditions is then given by $p(x_\Lambda|0) \propto \exp \phi^\Lambda(x_\Lambda)$ where ϕ^Λ is the partial potential

$$\phi^\Lambda(x^\Lambda) = -\frac{1}{2} x'_\Lambda J_\Lambda x_\Lambda + h'_\Lambda x_\Lambda \tag{4.3}$$

obtained by summing all potentials defined within that subfield. These conditional subfield models play a fundamental role in the RCM inference method. These are also GMRFs with respect to the subgraph $\mathbf{G}_\Lambda^\phi = \text{adj } \mathbf{H}_\Lambda^\phi$ induced by Λ.[1] Hence, these conditional subfield models (assuming zero boundary conditions) are naturally regarded as embedded graphical models within the graphical representation of the entire GMRF.

Given our graphical model we then wish to estimate the marginal distributions $p(x_\gamma)$ for all $\gamma \in \Gamma$.[2] The following inference procedure produces approximations of the marginal information parameters $x_\gamma \sim \mathcal{N}^{-1}(\hat{h}_\gamma, \hat{J}_\gamma)$ from which the marginal moments may be approximated by $(\hat{x}_\gamma, P_\gamma) = (\hat{J}_\gamma^{-1} \hat{h}_\gamma, \hat{J}_\gamma^{-1})$. These marginal information parameters would be given by exact calculation as:

$$\begin{aligned} \hat{h}_\gamma &= h_\gamma - J_{\gamma,\backslash\gamma} J_{\backslash\gamma}^{-1} h_{\backslash\gamma} & (4.4) \\ \hat{J}_\gamma &= J_\gamma - J_{\gamma,\backslash\gamma} J_{\backslash\gamma}^{-1} J_{\backslash\gamma,\gamma} & (4.5) \end{aligned}$$

However, this approach becomes intractable for large GMRFs. More efficient recursive methods, based on Markov trees (Section 2.3), are available but even these recursive methods become intractable when large state-dimensions (corresponding to large separators in $\mathbf{G}_\Gamma^\phi$) are required to construct such a Markov tree. Hence, RCM performs these calculations both recursively (to exploit the Markov structure of the field) and approximately (to give a tractable method when exact recursive methods do not). Approximations are introduced by our model thinning method so as to control the computational complexity of the recursive inference approach. This may be seen as a thinned variable elimination approach where *conditional* information projections (assuming zero boundary conditions) are introduced to select compact yet

[1] Recall, from Section 2.1.1, that $\mathbf{H}_\Lambda^\phi$, the subhypergraph of $\mathbf{H}_\Gamma^\phi$ induced by Λ, is comprised of all hyperedges contained in Λ. For canonical specifications, the conditional interaction graph adj $\mathbf{H}_\Lambda^\phi$ is also the subgraph $\mathbf{G}_\Lambda^\phi$ of $\mathbf{G}_\Gamma^\phi \equiv \text{adj } \mathbf{H}_\Gamma^\phi$ induced by Λ.

[2] Extension of the following inference procedure to also calculate the pairwise marginal distributions $p(x_\gamma, x_\lambda)$ for $\{\gamma, \lambda\} \in \mathcal{H}_\Gamma^\phi$ is straight-forward but is omitted to simplify presentation. This extension may prove useful to support tractable (albeit approximate) model identification based on RCM.

faithful "Markov blanket models" by pruning many of the fill edges arising in variable elimination. In the sequel, these Markov blanket models are called either "cavity" or "blanket" models depending upon the context.

The basic procedure consists of three stages: (i) *nested dissection* of the graphical model producing a *dissection tree*, (ii) a *cavity modeling* procedure which is described as an upward recursion with respect to the dissection tree, and (iii) a *blanket modeling* procedure which is described as a downward recursion on the dissection tree. The inference/modeling procedures (ii) and (iii) employ the model thinning methods of Chapter 3 as a subroutine. These model thinning steps, requiring exact inference of embedded graphical models corresponding to conditional subfield models, are tractable insofar as RCM generates tractable cavity and blanket models.

4.1.1 Nested Dissection

The first step of RCM is to dissect the GMRF. We specify a simple method for performing spatial dissection, essentially as in Luettgen's approach [91], but emphasize that our subsequent RCM inference procedures could also be applied for other dissection methods.

Here, we assume that the sites of the field Γ are naturally associated to spatial locations $(z(\gamma) \in \mathcal{R}^s, \gamma \in \Gamma)$, where s is the dimension of the space in which sites are located[3], and that the interactions of the field are mainly between nearby sites. The field is then recursively partitioned by spatial dissection as indicated in Figure 4-1. This dissection procedure begins by partitioning the set of all sites $\Lambda_0 = \Gamma$ into two disjoint subsets $\Lambda_1, \Lambda_2 \subset \Lambda_0$ such that $\Lambda_0 = \Lambda_1 \cup \Lambda_2$ and $\Lambda_1 \cap \Lambda_2 = \emptyset$. We will call these partitions *dissection cells*. These dissection cells are then themselves recursively partitioned into subcells.

The partitioning at each level of dissection is determined by alternating bisection of the spatial coordinates. This is accomplished by forming s lists of the sites of the field for $k = 1, \ldots, s$ where list k is sorted by the k-th spatial coordinate z_k. Spatial dissection is then accomplished by iterating over the coordinates $k = 1, \ldots, s$ and splitting the k-th list at the median value. The sites of the field and the s sorted coordinate lists are partitioned accordingly. This spatial dissection is performed recursively until the field has been divided into subfields which are sufficiently small so as to be tractable by exact inference methods (for instance, subfields consisting of just one site). This procedure is most readily implemented for sites arranged on regular grids and then has $\mathcal{O}(|\Gamma|)$ complexity.[4]

We encode the nested structure of this dissection procedure as a tree data structure. Here, we favor a *directed tree* specification given by a pair $\mathbf{T} = (\mathcal{N}, \mathcal{A})$ where $\mathcal{N}$ is the set of *nodes* and $\mathcal{A}$ is the set of *arcs*. An arc $(\alpha, \beta) \in \mathcal{A}$ is an *ordered* pair of nodes $\alpha, \beta \in \mathcal{N}$. The arc "points" from node α to node β. We also say that α is the *child* of β (or equivalently that β is the *parent* of α). A tree is *singly-connected*

[3] All examples shown are for planar GMRFs (s=2), but our RCM method could also be applied for multidimensional GMRFs (for instance, for GMRFs arrayed in 3D).

[4] More generally, for irregular arrangements of sites, we must explicitly sort each list of sites, once for each spatial coordinate, requiring $\mathcal{O}(s|\Gamma| \log |\Gamma|)$ computation.

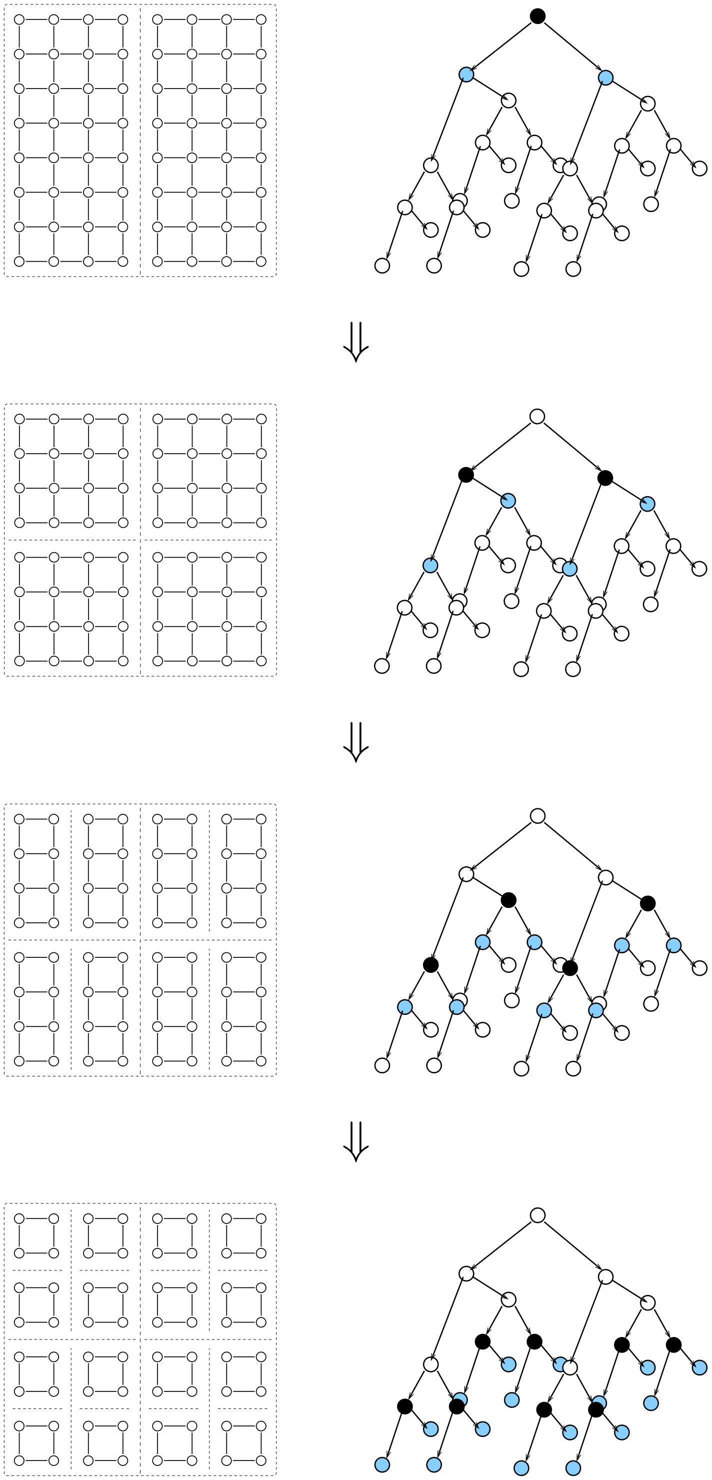


Figure 4-1: Illustration of nested dissection of an 8×8 square-grid graphical model. The dissection precedes top to bottom where the dissection cells are shown to the left and the corresponding nodes of the dissection tree are shown to the right. At each level of dissection, the darkened nodes indicate the cells being split and the shaded nodes indicate subcells. The RCM inference procedure is structured according to this dissection tree.

so that no two nodes may be linked together by two distinct sequences of arcs. A node with no parents is called a *root*. A node with no children is called a *leaf*. For a non-leaf node $\alpha \in \mathcal{N}$, we let α_i denote the i-th child of α (imposing some arbitrary ordering of the children). The trees that we consider have only one root and are connected so that any node of the tree may be reached from the root.

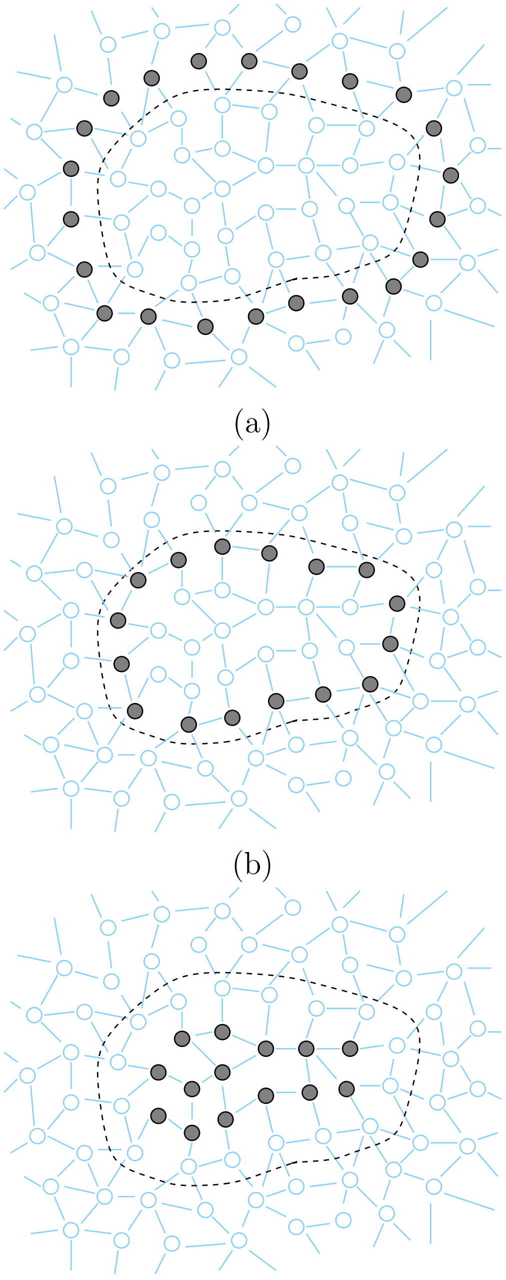
To encode the nested dissection of the field as a directed tree we let the nodes of the tree represent the cells generated by the dissection procedure. For node $\alpha \in \mathcal{N}$, we denote the associated dissection cell by $\Lambda_\alpha \subset \Gamma$. The arcs then indicate the nested structure of the dissection. For instance, if a dissection cell Λ_α is partitioned into subcells Λ_{α_1} and Λ_{α_2} then we include the arcs (α, α_1) and (α, α_2) in $\mathcal{A}$. The root of the tree α_0 corresponds to the set of all sites $\Lambda(\alpha_0) = \Gamma$. The leaf nodes of the tree correspond to the smallest dissection cells at the last level of dissection. We call this the *dissection tree*.

The subsequent RCM inference procedures are structured according to this dissection tree. We make use of the following terminology defined relative to the graph $\mathbf{G}_\Gamma^\phi$ describing the Markov structure of the field. The *blanket* of cell α is defined as $\Lambda_\alpha^b = \partial\Lambda_\alpha$. The *surface* of cell α is $\Lambda_\alpha^s = \partial\{\Gamma \setminus \Lambda_\alpha\}$. The *interior* of cell α is $\Lambda_\alpha^i = \Lambda_\alpha \setminus \Lambda_\alpha^s$. These are illustrated in Figure 4-2.

4.1.2 Cavity Modeling

Next, a recursive upward pass is executed with respect to the dissection tree. The aim of the upward pass is to construct a *cavity model* for each dissection cell. The cavity model at node α of the dissection tree is intended as a compact yet faithful graphical model for the surface of the dissection cell $\Lambda_\alpha^s \subset \Gamma$ sufficient (or nearly so) for inference outside of the cell. We denote this graphical model by $\tilde{\mu}_\alpha^s$. A recursive procedure for constructing cavity models is outlined below. The main idea is to construct cavity models from subcavity models (cavity models of subcells). The main subroutines are (a) variable elimination, (b) model thinning and (c) rejoining the (thinned) parts of the graphical model (reversing the dissection procedure).

Cavity Model Initialization. Cavity modeling begins at the leaves of the dissection tree. For each leaf α, we construct a cavity model for the corresponding dissection cell Λ_α as follows. First, the partial model of the subfield x_{Λ_α} is extracted. This is a graphical model μ_α consisting of the subhypergraph of $\mathbf{H}_\Gamma^\phi$ induced by $\Lambda_\alpha \subset \Gamma$ and the associated interaction potentials ϕ^{Λ_α}. In GMRFs, this is given by the subset of the information parameters $(h_{\Lambda_\alpha}, J_{\Lambda_\alpha})$. This partial model specifies the conditional distribution of the subfield assuming ground-state boundary conditions (set to zero in our information representation of GMRFs) outside of the dissection cell. The cavity model $\tilde{\mu}_\alpha^s$ is then constructed from the partial model μ_α in two steps: (i) variable elimination and (ii) model thinning. This initialization procedure is illustrated in Figure 4-3.

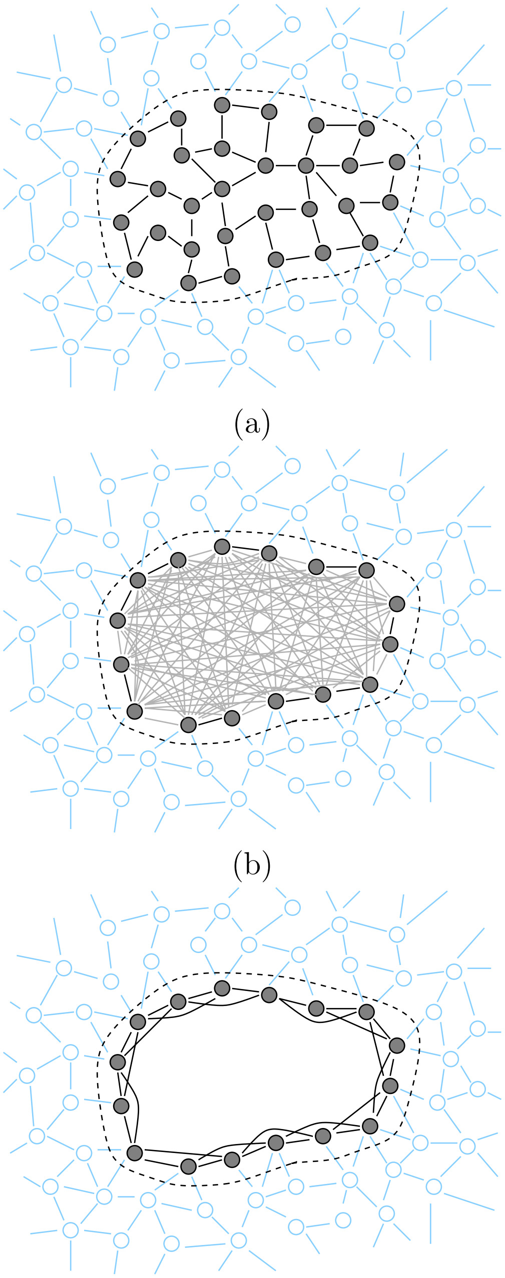


(a)

(b)

(c)

Figure 4-2: Diagrams illustrating the (a) blanket Λ^b, (b) surface Λ^s and (c) interior Λ^i of the subfield Λ enclosed by the dashed line.



(a)

(b)

(c)

Figure 4-3: Diagrams illustrating initialization of RCM cavity modeling: (a) conditional subfield model assuming zero boundary conditions (shown in bold) (b) interior nodes eliminated producing many interactions between sites in the surface of subfield (shown in grey) (c) result of model thinning procedure to prune weak interactions yielding desired "cavity model".

(1) Variable Elimination. First, we perform variable elimination for all sites in the interior of the cell (not in the surface). For Gaussian MRFs, this variable elimination corresponds to calculation of $\hat{\mu}^s_\alpha = (h^{\text{elim}}_{\Lambda^s}, J^{\text{elim}}_{\Lambda^s})$ from $\mu_\alpha = (h_\Lambda, J_\Lambda)$ as follows:

$$
\begin{aligned}
h^{\text{elim}}_{\Lambda^s} &= h_{\Lambda^s} - J_{\Lambda^s,\Lambda^i} J^{-1}_{\Lambda^i} h_{\Lambda^i} && (4.6)\\
J^{\text{elim}}_{\Lambda^s} &= J_{\Lambda^s} - J_{\Lambda^s,\Lambda^i} J^{-1}_{\Lambda^i} J_{\Lambda^i,\Lambda^s} && (4.7)
\end{aligned}
$$

We remark that this variable elimination may also be performed recursively by eliminating the sites sequentially one at a time. Depending upon the sparsity of the graphical model and the elimination order, this may reduce computation. In any case, variable elimination removes sites from the interior of the cell but creates additional interactions between sites in the surface of the cell. In our initialization of cavity models, elimination typically produces a fully-connected cavity model such as shown in Figure 4-3(b). The production of these additional fill edges is the main source of intractability of exact recursive inference methods.

(2) Model Thinning. Next, RCM performs model thinning so as to provide a tractable inference approach. This consists of the inductive thinning procedure described in Chapter 3 which adaptively prunes selected weak interactions from the graphical model by information projections. An AIC-like principle is employed to select the final embedded graphical model where our precision parameter δ controls the trade-off between the complexity and the accuracy of our thinned models.

During initialization, we apply this model thinning procedure for the (fully connected) cavity models $\hat{\mu}_\alpha = (h^{\text{elim}}_{\Lambda^s}, J^{\text{elim}}_{\Lambda^s})$ produced by variable elimination. Hence, the information projections performed by our model thinning subroutine are conditioned on zero boundary conditions. This allows for the m-projection to be performed in a tractable manner as inference is required only for embedded graphical models corresponding to conditional subfields. Applying model thinning for $\hat{\mu}_\alpha = (h^{\text{elim}}_{\Lambda^s}, J^{\text{elim}}_{\Lambda^s})$ gives our thinned cavity model $\tilde{\mu}^s_\alpha = (h^{\text{thin}}_{\Lambda^s}, J^{\text{thin}}_{\Lambda^s})$ with thinned interaction graph $\mathbf{G}^{\text{thin}}_{\Lambda^s_\alpha} = (\Lambda^s_\alpha, \mathcal{E}^{\text{thin}}_{\Lambda^s_\alpha})$. Such a thinned cavity model is illustrated in Figure 4-3(c).

We will not review all of the details of our model thinning approach here, but do wish to remind the reader of some key points. First, the model thinning m-projection imposes the constraint that the cavity model is Markov with respect to $\mathbf{G}^{\text{thin}}_{\Lambda^s_\alpha}$, or that $J^{\text{thin}}_{\Lambda^s_\alpha}$ is sparse such that $(J^{\text{thin}}_{\Lambda^s_\alpha})_{\gamma,\lambda} = 0$ for all pruned edges $\{\gamma, \lambda\} \notin \mathcal{E}^{\text{thin}}_{\Lambda^s_\alpha}$. Second, the m-projection of $\hat{\mu}_\alpha$ to this family of thinned graphical models, so as to minimize the KL-divergence $D(\hat{\mu}_\alpha \| \tilde{\mu}_\alpha)$ subject to those sparsity constraints, is determined by moment matching. That is we match conditional means (assuming $\mathrm{x}_{\Lambda^b_\alpha} = 0$) such that

$$
(J^{\text{thin}}_{\Lambda^s})^{-1} h^{\text{thin}}_{\Lambda^s} = (J^{\text{elim}}_{\Lambda^s})^{-1} h^{\text{elim}}_{\Lambda^s} \tag{4.8}
$$

and match a selected subset of conditional covariances such that

$$
(J^{\text{thin}}_{\Lambda^s})^{-1}_{\gamma,\lambda} = (J^{\text{elim}}_{\Lambda^s})^{-1}_{\gamma,\lambda} \tag{4.9}
$$

along the diagonal $\gamma = \lambda$ and for all edges $\{\gamma, \lambda\} \in \mathcal{E}^{\mathrm{thin}}_{\Lambda^s_\alpha}$ retained by our thinned cavity model. Together, these moment and sparsity conditions uniquely determine $\tilde{\mu}_\alpha = (h^{\mathrm{thin}}_{\Lambda^s}, J^{\mathrm{thin}}_{\Lambda^s})$. Finally, we recall that this also has the interpretation of maximizing entropy subject to just the moment constraints where the minimal KL-divergence is given by the entropy gain (information loss) $D(\hat{\mu}_\alpha \| \tilde{\mu}_\alpha) = h[\tilde{\mu}_\alpha] - h[\hat{\mu}_\alpha]$. In this regard, RCM may be understood as a "forgetful" inference procedure where only a subset of the moment characteristics are preserved during model thinning and we otherwise assume as little as possible about the cavity model. Our selection of which edges to prune essentially corresponds to selecting which moments constraints can be relaxed without too significantly perturbing the model (keeping the information loss per removed model parameter less than δ).

Region Merging. The cavity modeling procedure then proceeds up the dissection tree by merging cavity models of subcells as in Figure 4-4. Given two cavity models $\tilde{\mu}^s_{\alpha_1}$ and $\tilde{\mu}^s_{\alpha_2}$ for subcells α_1 and α_2 of cell α, we approximate the cavity model $\tilde{\mu}^s_\alpha$ as follows. First, the two subcavity models are *joined* by merging the potentials from both graphical models and also reinstating those interaction potentials between the subfields previously severed during the dissection procedure. This gives a graphical model for the conditional distribution of the subfield $\Lambda^s_{\alpha_1,\alpha_2} \equiv \Lambda^s_{\alpha_1} \cup \Lambda^s_{\alpha_2}$ assuming $\mathrm{x}_{\Lambda^b_\alpha} = 0$.

For GMRFs, this corresponds to coupling two cavity models

$$\begin{aligned} \mathrm{x}_{\Lambda^s_{\alpha_1}} &\sim \mathcal{N}^{-1}(h^{\mathrm{thin}}_{\Lambda^s_{\alpha_1}}, J^{\mathrm{thin}}_{\Lambda^s_{\alpha_1}}) \qquad (4.10) \\ \mathrm{x}_{\Lambda^s_{\alpha_2}} &\sim \mathcal{N}^{-1}(h^{\mathrm{thin}}_{\Lambda^s_{\alpha_2}}, J^{\mathrm{thin}}_{\Lambda^s_{\alpha_2}}) \end{aligned}$$

with interactions

$$\phi(x_{\Lambda^s_{\alpha_1}}, x_{\Lambda^s_{\alpha_2}}) = -\frac{1}{2} x'_{\Lambda^s_{\alpha_1}} K x_{\Lambda^s_{\alpha_2}}$$

where $K \equiv J_{\Lambda^s_{\alpha_1}, \Lambda^s_{\alpha_2}}$ are just those interactions between these subfields which were severed during the dissection procedure. This produces the joined cavity model $\mu^{\mathrm{join}}_\alpha$,

$$x_{\Lambda^s_{\alpha_1,\alpha_2}} \sim \mathcal{N}^{-1}(h^{\mathrm{join}}_{\Lambda^s_{\alpha_1,\alpha_2}}, J^{\mathrm{join}}_{\Lambda^s_{\alpha_1,\alpha_2}})$$

with information parameters

$$\begin{aligned} h^{\mathrm{join}}_{\Lambda^s_{\alpha_1,\alpha_2}} &= \begin{pmatrix} h^{\mathrm{thin}}_{\Lambda^s_{\alpha_1}} \\ h^{\mathrm{thin}}_{\Lambda^s_{\alpha_2}} \end{pmatrix} \qquad (4.11) \\ J^{\mathrm{join}}_{\Lambda^s_{\alpha_1,\alpha_2}} &= \begin{pmatrix} J^{\mathrm{thin}}_{\Lambda^s_{\alpha_1}} & K \\ K' & J^{\mathrm{thin}}_{\Lambda^s_{\alpha_2}} \end{pmatrix} \qquad (4.12) \end{aligned}$$

This initializes our cavity model at non-leaf nodes of the dissection tree.

Again, variable elimination is required to eliminate sites in the interior Λ^i_α. These are sites $\Lambda^s_{\alpha_1,\alpha_2} \setminus \Lambda^s_\alpha$ that are in the surface of one of the subcells α_1 or α_2 but are not in the surface of cell α. This produces $\hat{\mu}_\alpha = (h^{\mathrm{elim}}_{\Lambda^s_\alpha}, J^{\mathrm{elim}}_{\Lambda^s_\alpha})$ having additional interactions

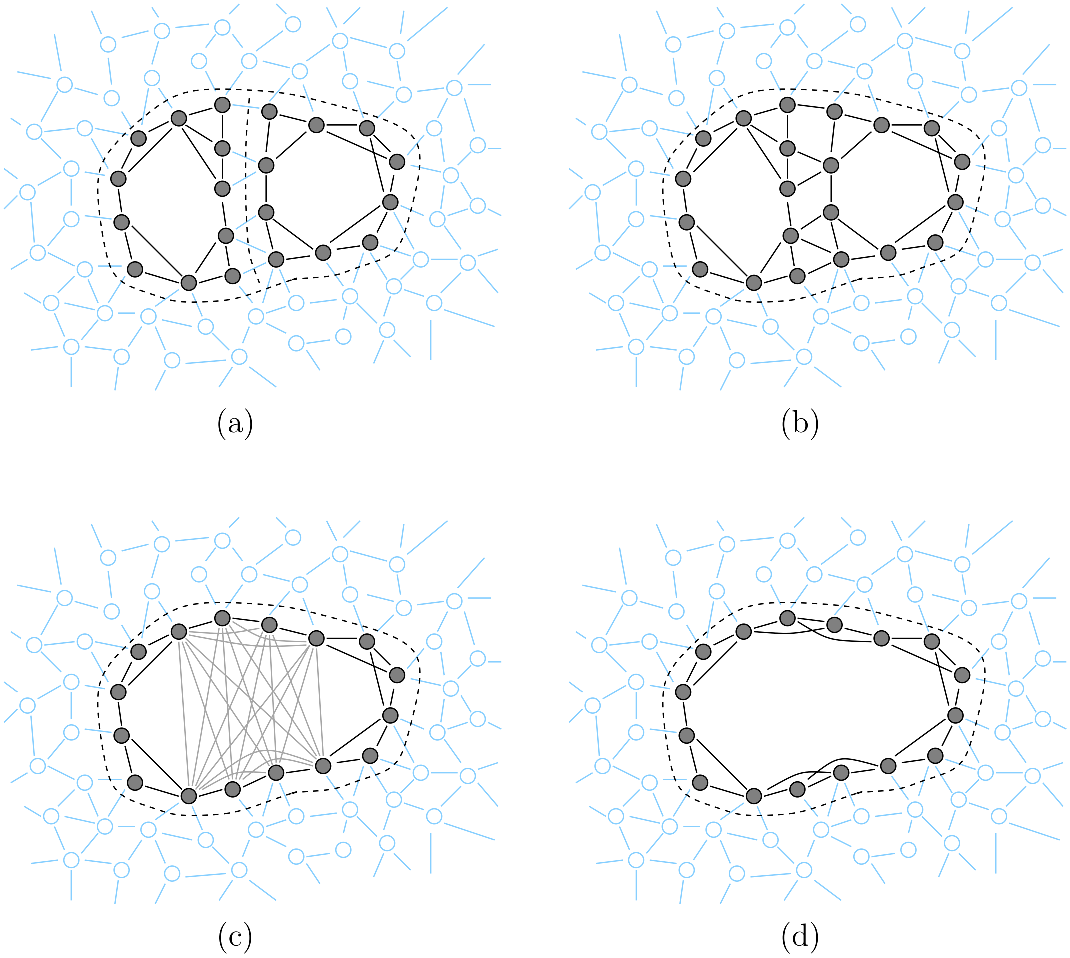

(a) (b) (c) (d)

Figure 4-4: Illustration of recursive method for constructing cavity models from subcavity models: (a) initialized by cavity models of subcells, (b) subcavity models are joined reinstating severed potentials, (c) latent variables (in the interior) are removed by variable elimination producing some additional interactions (but this does not cause the cavity model to become fully connected), (d) model thinning selects a compact yet faithful embedded graphical model, inductively pruning weak interactions from the model by information projection. This gives the cavity model which may in turn be used to construct larger cavity models.

due to fill but where many of these interactions tend to be "weak" interactions between distant sites on opposite sides of the "cavity" left by elimination. We would like to prune many of these weak interactions from our cavity model. Note that, due to earlier thinning of the subcavity models, the amount of fill is limited so that elimination does not produce an (intractable) fully-connected graphical model. Hence, recursive inference is tractable so that further model thinning is also tractable. Model thinning then gives our thinned cavity model $\tilde{\mu}_\alpha^s = (h_{\Lambda_\alpha^s}^{\mathrm{thin}}, J_{\Lambda_\alpha^s}^{\mathrm{thin}})$ at node α of the dissection tree. Thus, we have our tractable "bootstrap" method for constructing cavity models for larger dissection cells from cavity models of subcells.

Upward Cavity-Modeling Pass. This gives a recursive upward pass with respect to the dissection tree which begins at the leaves of the tree and then works up the tree building cavity models from subcavity models. The action of this upward pass may be viewed as reversing the dissection procedure but substituting thinned cavity models for subfields.

We define the following subroutines:

- $\mu_\Lambda = \mathrm{Extract}(\Lambda)$: form the partial model μ_Λ based on the induced subhypergraph $\mathbf{H}_\Lambda^\phi$ and associated potentials ϕ^Λ.

- $\mu_{1,2} = \mathrm{Join}(\mu_1,\mu_2)$: join two partial models μ_1 and μ_2 by reinstating the potentials between them severed during the dissection procedure.

- $\hat{\mu}_\Lambda = \mathrm{Elim}(\mu,\Lambda)$: perform variable elimination of μ eliminating all sites *except* those in Λ so as to produce a graphical model $\hat{\mu}_\Lambda$ for just the sites Λ.

- $\tilde{\mu} = \mathrm{Thin}(\mu)$: thin the graphical model μ using the inductive information projection techniques of Chapter 3.

With these subroutines, we specify the upward cavity modeling procedure as follows:

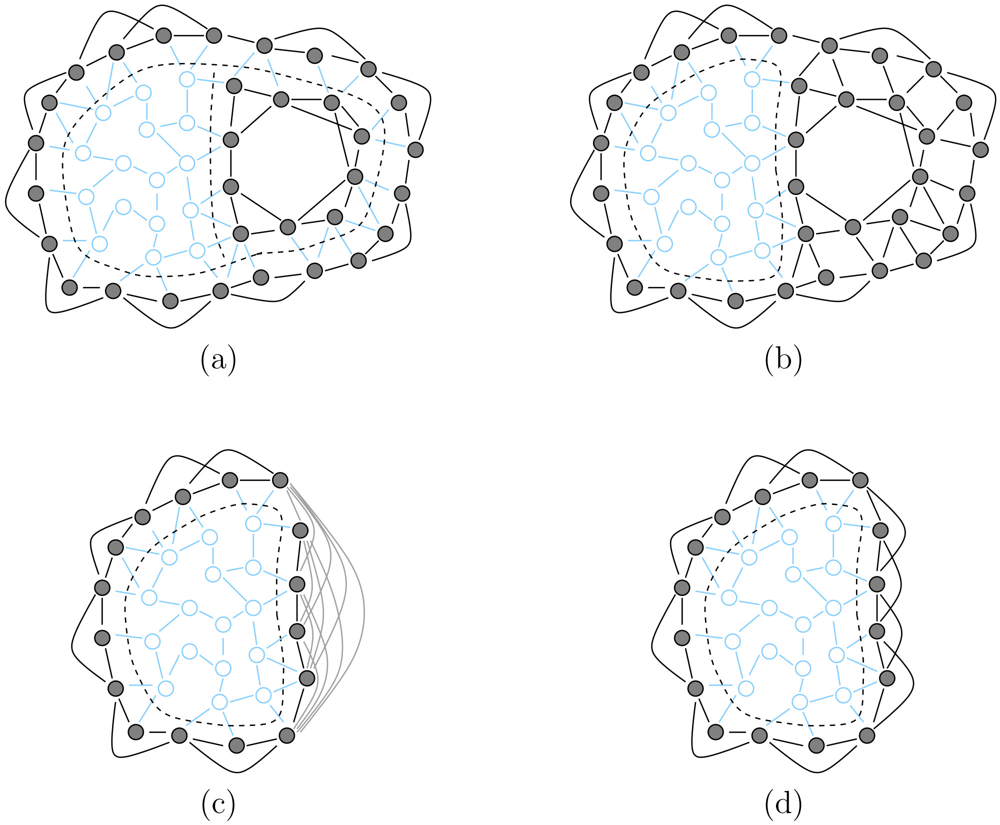
```
RCMUpwardPass($\alpha$)
    if ($\alpha$ is a leaf node)
        Extracts partial subfield model.
        $\mu_\alpha$ = Extract($\Lambda_\alpha$)
        Eliminate interior and thin.
        $\hat{\mu}^s_\alpha$ = Elim($\mu_\alpha$,$\Lambda^s_\alpha$)
        $\tilde{\mu}^s_\alpha$ = Thin($\hat{\mu}^s_\alpha$)
    else
        Recurse on subtrees.
        $\tilde{\mu}^s_{\alpha_1}$=RCMUpwardPass($\alpha_1$)
        $\tilde{\mu}^s_{\alpha_2}$=RCMUpwardPass($\alpha_2$)
        Join subcavity models along internal cut.
        $\mu^{s+}_\alpha$ = Join($\tilde{\mu}^s_{\alpha_1}$, $\tilde{\mu}^s_{\alpha_2}$)
        Eliminate and thin cavity model.
        $\hat{\mu}^s_\alpha$ = Elim($\mu^{s+}_\alpha$, $\Lambda^s_\alpha$)
        $\tilde{\mu}^s_\alpha$ = Thin($\hat{\mu}^s_\alpha$)
    end
    save $\tilde{\mu}^s_\alpha$ Store copy for reuse during downward pass.
    return $\tilde{\mu}^s_\alpha$
end
```

This procedure is invoked at the root node α_0 of the dissection tree, but processing begins at the leaves of the tree and propagates cavity models up the tree. A cavity model is stored at each node of the tree in preparation for the following downward procedure.

4.1.3 Blanket Modeling

Finally, a complementary recursive downward procedure is executed with respect to the dissection tree producing marginal models at the leaf nodes of the dissection tree. The downward pass operates by constructing *blanket models* for each dissection cell. The blanket model $\tilde{\mu}^b_\alpha$ is intended as a compact yet faithful graphical model for the blanket Λ^b_α of the subfield sufficient (or nearly so) for inference inside the subfield. The blanket models at the leaves of the dissection tree are then used to infer the desired marginal distributions. The main idea is to construct blanket models recursively from an adjacent cavity model (constructed by the preceeding upward pass) and an enclosing blanket model (the blanket model of the parent cell). This idea is illustrated in Figure 4-5.

In contrast to cavity models, blanket models for each subfield are constructed from the *exterior* of that subfield. Also, information projections of the model thinning procedure assume zero state *inside* the subfield. Essentially, blanket models are just cavity models for the complement of each subfield. Applying this method recursively results in the following recursive downward procedure with respect to the dissection



(a)

(b)

(c)

(d)

Figure 4-5: Illustration of procedure for constructing blanket models: (a) initialized by adjacent cavity model and enclosing blanket model, (b) models are joined reinstating severed potentials, (c) latent variables (not in the blanket) removed by variable elimination producing additional interactions, (d) model thinning selects a compact yet faithful embedded graphical model, inductively pruning weak interactions from the model by information projection. This gives the blanket model for the enclosed subfield which may now be inferred either directly or recursively.

tree.

Downward Blanket-Modeling Pass. This procedure uses the same subroutines as the earlier upward pass and operates on similar principles.

```
**RCMDownwardPass**($\alpha$, $\tilde{\mu}^b_\alpha$) *Takes blanket model as input.*
    **if** ($\alpha$ is a leaf node)
        *Output marginal model.*
        $\mu_\alpha$ = Extract($\Lambda_\alpha$)
        $\mu^{+b}_\alpha$ = Join($\mu_\alpha$, $\tilde{\mu}^b_\alpha$)
        $\hat{\mu}_\alpha$ = Elim($\mu^{+b}_\alpha$,$\Lambda_\alpha$)
        **save** $\hat{\mu}_\alpha$
    **else**
        *Restore cavity models.*
        **load** $\tilde{\mu}^s_{\alpha_1}$, $\tilde{\mu}^s_{\alpha_2}$
        *Initialize blanket models.*
        $\mu^{b+}_{\alpha_1}$ = Join($\tilde{\mu}^b_\alpha$, $\tilde{\mu}^s_{\alpha_2}$)
        $\mu^{b+}_{\alpha_2}$ = Join($\tilde{\mu}^b_\alpha$, $\tilde{\mu}^s_{\alpha_1}$)
        *Eliminate and thin.*
        $\hat{\mu}^b_{\alpha_1}$ = Elim($\mu^{b+}_{\alpha_1}$,$\Lambda^b_{\alpha_1}$)
        $\hat{\mu}^b_{\alpha_2}$ = Elim($\mu^{b+}_{\alpha_2}$,$\Lambda^b_{\alpha_2}$)
        $\tilde{\mu}^b_{\alpha_1}$ = Thin($\hat{\mu}^b_{\alpha_1}$)
        $\tilde{\mu}^b_{\alpha_2}$ = Thin($\hat{\mu}^b_{\alpha_2}$)
        *Recurse on subtrees.*
        RCMDownwardPass($\alpha_1$, $\tilde{\mu}^b_{\alpha_1}$)
        RCMDownwardPass($\alpha_2$, $\tilde{\mu}^b_{\alpha_2}$)
    **end**
**end**
```

This blanket-modeling downward procedure is invoked at the root node α_0 of the dissection tree (with an “empty” blanket model at the root node) and propagates blanket models down the tree. The recursion terminates at the leaves of the tree where the blanket models are joined with conditional subfield models to provide marginal models for those smallest dissection cells. Inference of marginal distributions at individual sites (or between adjacent sites) is then straight-forward by exact inference methods.

This completes specification of the two-pass RCM inference procedure. We illustrate this two-pass procedure in Figures 4-6 and 4-7 for our (fictitouos) 8×8 example with dissection tree as shown previously in Figure 4-1. Finally, we wish to point out that, if we were to omit the model thinning steps in both the upward and downward pass, then our approach reduces to a recursive variable elimination approach for *exact* computation of the marginal distributions of the GMRF. This latter exact approach, however, will become intractable for many larger GMRFs due to the construction of fully-connected cavity and blanket models throughout the computation.

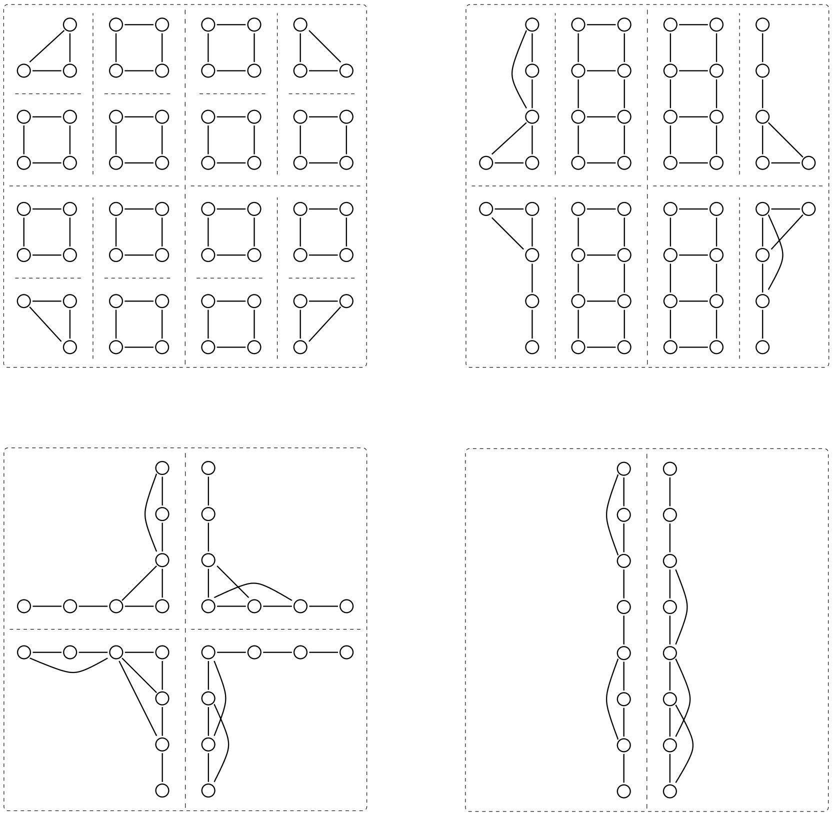


Figure 4-6: Illustration of RCM upward pass (showing cavity models at every level of dissection). This is a recursive procedure defined on the dissection tree (Figure 4-1) for building cavity models at each node of the dissection tree. Cavity models for larger dissection cells are built from cavity models of subcells.

In the next section we present two iterative extensions of this approach which attempt to refine the approximations we have introduced into our inference procedure. Later, in Section 4.3, we demonstrate RCM in some simulated examples.

4.2 Iterative RCM

In this section we give some extensions of the basic RCM inference framework described in the previous section. We describe two iterative methods based on the two-pass RCM procedure. The first, iterative renormalization, uses two-pass RCM as a subroutine but adjusts the potential specification between iterations. The second, iterative remodeling, modifies the two-pass algorithm to exploit previously constructed cavity and blanket models while selecting refined cavity and blanket models on later iterations.

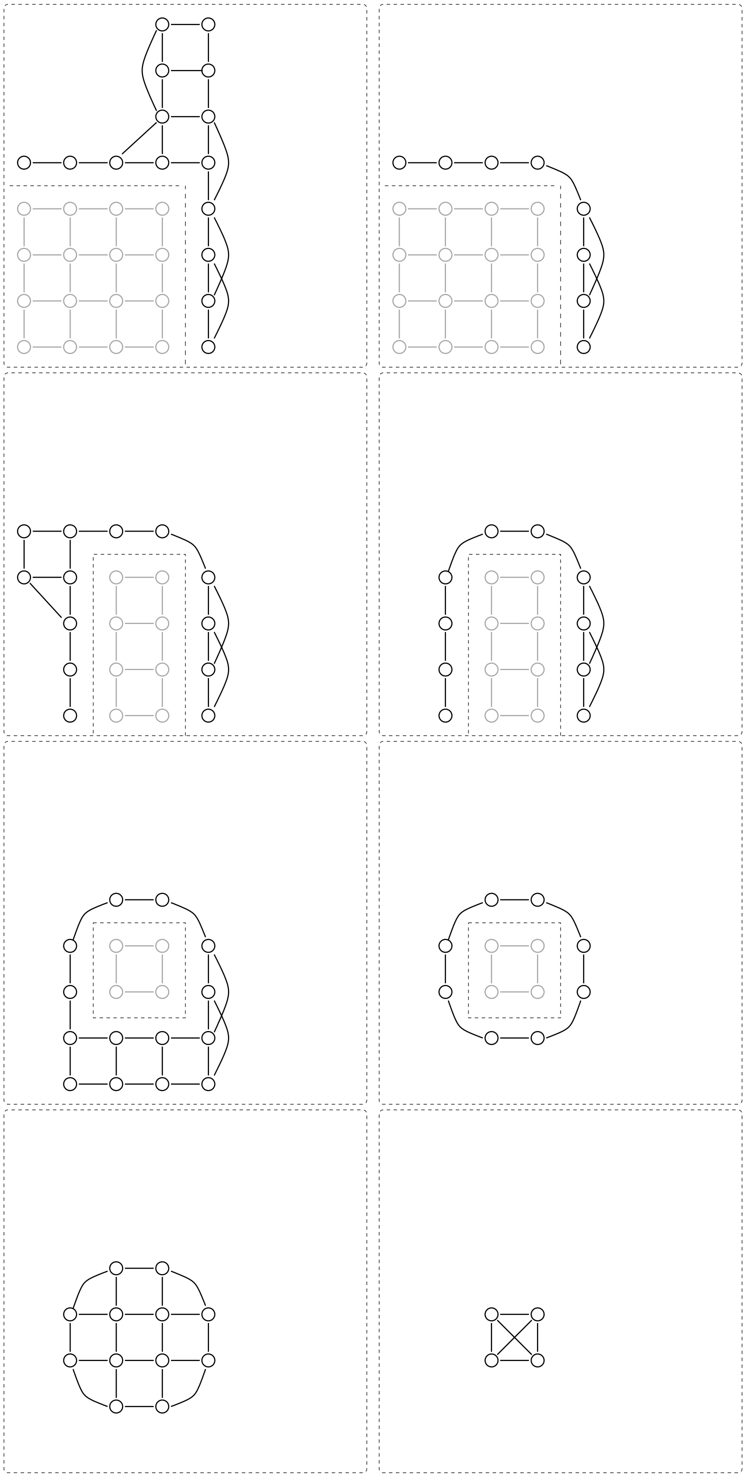


Figure 4-7: Illustration of RCM downward pass (tracing just one path down the dissection tree). This is a recursive procedure defined on the dissection tree (Figure 4-1) for building blanket models at each node of the dissection tree. Blanket models are built form adjacenct cavity models and enclosing blanket models. At the leaves of the dissection tree, marginal models are built by joining subfields with blanket models and eliminating sites in the blanket.

4.2.1 Renormalization

In this subsection we describe an iterative version of the RCM procedure which we call *iterative renormalization*. We give an implementation of this procedure for GMRFs which employs the two-pass implementation described previously as a subroutine. This Gaussian implementation is also identified as performing a Richardson iteration (Young [138], Kelley [81], Sudderth [125]) with RCM playing the role of a preconditioner.

Fundamentally, the accuracy of two-pass RCM hinges on the accuracy of the cavity and blanket models we select. In order to be able to select these models, by tractable m-projections, we have thus far found it necessary to condition on zero state along the boundary of the subfield being approximated. This then allows the m-projection to be carried out entirely in terms of local computations involving just our thinned graphical models. The fundamental idea of renormalization is to allow ourselves instead to select some *estimate* of the state along the boundary while carrying out the model thinning operation. In the context of Gibbs random fields, specified by canonical potentials relative to the so-called ground state x^* of the potential specification (Section 2.1.3), we interpret this as *renormalization* as we now explain.

As discussed in Section 4.1, the model thinning m-projections performed in two-pass RCM implicitly assume ground-state boundary conditions (set to zero in the case of the information representation of GMRFs). This occurs as we only consider partial models, conditional subfield models specified on induced subgraphs of the graphical model, in isolation of the rest of the graphical model. That is, the "conditioning" is really implicit in our choice of potential specification. Hence, in the cavity-modeling phase of the procedure, we are really conditioning on zero boundary conditions (outside of the dissection cell) while selecting our thinned cavity model. Likewise, in the blanket-modeling phase, we are conditioning on zero boundary conditions (inside the dissection cell) while selecting thinned approximations for the blanket. Hence, the approximations introduced by RCM depend upon the choice of ground state (previously always set to zero) implicit in our potential specification. This then begs the question if we might obtain more accurate cavity and blanket models, thereby improving the accuracy of inferred marginal distributions, by appropriate choice of the ground state. Recalling the normalization property of the canonical potential (relative to a given ground state), we call this change of ground state *renormalization*.

We first give a general outline for iterative renormalization (in MRFs) and then specify an implementation of this approach appropriate for GMRFs. In the general formulation, we indicate that either the means or the modes of the (approximate) marginal distributions provide a basis for renormalization (in GMRFs these are identical).[5]

[5] Note that, in finite-state MRFs, the mode renormalization method may be preferable since the means are typically not elements of the state space.

Iterative Renormalization:

- *Input.* Graphical model μ based on canonical potential specifications relative to an arbitrary ground state x^* (for instance, the zero ground-state $x^* = 0$ where applicable).

- *Loop.* Initialize $k = 0$, $\mu^{(0)} = \mu$ and $\hat{x}^{(0)} = x^*$. Do the following until termination is indicated:

  - *RCM Inference.* Run two-pass RCM for graphical model $\mu^{(k)}$ giving approximate marginal distributions $(\mu^{(k)}(x_\gamma), \forall \gamma \in \Gamma)$. Note that the model thinning m-projections in RCM assume boundary conditions specified by $\hat{x}^{(k)}$.
  - *State Estimation.* Generate state estimate $\hat{x}^{(k+1)}$ either by

    $$\hat{x}_\gamma^{(k+1)} = E_{\mu^{(k)}}\{\mathrm{x}_\gamma\}$$

    for *mean-renormalization*, or by

    $$\hat{x}_\gamma^{(k+1)} = \arg\max_{x_\gamma \in \mathcal{X}_\gamma} \mu^{(k)}(x_\gamma)$$

    for *mode-renormalization.* If $\hat{x}^{(k+1)} = \hat{x}^{(k)}$ (to within some tolerance), then terminate loop.
  - *Renormalization.* Construct canonical potential specification $\phi^{(k+1)}$ relative to ground state $x^* = \hat{x}^{(k+1)}$ giving a "renormalized" graphical model $\mu^{(k+1)}$.
  - *Iterate.* Set $k \leftarrow k + 1$ and repeat.

- *Output.* State estimate $\hat{x}^{(k)}$, renormalized model $\mu^{(k)}$ (with correspondingly refined approximation of cavity, blanket and marginal models).

We illustrate this idea in the context of GMRFs where it has an especially simple form. Relative to an arbitrary choice of ground state x^*, we find that the canonical potentials for a GMRF $\mathrm{x} \sim \mathcal{N}^{-1}(h, J)$ with interaction graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ are given by the singleton potential functions

$$\phi_\gamma^*(x_\gamma) = -\frac{1}{2}(x_\gamma - x_\gamma^*)' J_\gamma (x_\gamma - x_\gamma^*) + (h - Jx^*)_\gamma' (x_\gamma - x_\gamma^*) \tag{4.13}$$

for all $\gamma \in \Gamma$, and the pairwise interaction potentials

$$\phi_{\gamma,\lambda}^*(x_\gamma, x_\lambda) = -(x_\gamma - x_\gamma^*)' J_{\gamma,\lambda} (x_\lambda - x_\lambda^*) \tag{4.14}$$

for all $\{\gamma, \lambda\} \in \mathcal{E}_\Gamma$. It is apparent that these interaction potentials satisfy the normalization property such that $\phi_\gamma^*(x_\gamma^*) = 0$, $\phi_{\gamma,\lambda}^*(x_\gamma^*, x_\lambda) = 0$ and $\phi_{\gamma,\lambda}^*(x_\gamma, x_\lambda^*) = 0$.

Setting

$$h^* = h - Jx^* \tag{4.15}$$

gives an exponential family with linear and quadratic statistics in $x - x^*$ and exponential parameters (h^*, J).

$$p(x) \propto \exp\{-\frac{1}{2}(x - x^*)' J (x - x^*) + h^* \cdot (x - x^*)\} \tag{4.16}$$

Note that this reduces to the usual (h, J) information parameterization when $x^* = 0$. On the other hand, setting $x^* = E\{\mathrm{x}\}$ gives $h^* = 0$ (indicating that the linear moment constraints are inactive) corresponding to the mean-centered covariance selection model (parameterized by just the non-zero entries of the inverse covariance J). Hence, for GMRFs, iterative renormalization may be viewed as an attempt to iteratively transition from the (h, J) to the $(0, J)$ representation by adjusting the ground state x^* (implicit in our choice of potential specification) from $x^* = 0$ to $x^* = E\{\mathrm{x}\}$.

Collecting all interaction potentials associated with subfield Λ (and summing these) gives the partial potential function:

$$\phi^*(x_\Lambda) = -\frac{1}{2}(x_\Lambda - x^*_\Lambda)' J_\Lambda (x_\Lambda - x^*_\Lambda) + h^*_\Lambda \cdot (x_\Lambda - x^*_\Lambda) \tag{4.17}$$

This (renormalized) partial potential now specifies the conditional distribution

$$p(x_\Lambda | x^*_{\partial\Lambda}) \propto \exp \phi^*(x_\Lambda)$$

of subfield x_Λ assuming ground-state boundary conditions $\mathrm{x}_{\partial\Lambda} = x^*_{\partial\Lambda}$. Hence, performing RCM with respect to this "renormalized" information parameterization implies that the model-thinning m-projections now assume non-zero boundary conditions as specified in x^*.

Rather then specifying a new version of Gaussian RCM for this new (h^*, J) parameterization relative to x^*, we may instead "shift" the inference procedure as follows. Let $\mathrm{y} = \mathrm{x} - x^*$ such that $\mathrm{y} \sim \mathcal{N}^{-1}(h^*, J)$. Then run the usual RCM inference procedure with information model (h^*, J) producing the estimate $\hat{y}^{(RCM)}$ (the RCM approximation for the expectation $\hat{y} = E\{\mathrm{y}\}$). Then calculate the estimate $\hat{x}^{(RCM)} = \hat{y}^{(RCM)} + x^*$ (the "renormalized" RCM approximation for the expectation $\hat{x} = E\{\mathrm{x}\}$). Then reset the ground state $x^* = \hat{x}^{(RCM)}$ to seed the next iteration. This gives the following iterative procedure:

Gaussian Iterative Renormalization:

- *Input.* Graphical model $\mu = (h, J)$, tolerance τ.
- *Loop.* Initialize $k = 0$, $h^{(0)} = h$ and $\hat{x}^{(0)} = 0$. Do the following until termination is indicated:
  - *RCM Inference.* $\Delta\hat{x}^{(k)} = RCM(h^{(k)}, J) \approx J^{-1}h^{(k)}$.
  - *Update State Estimate.* $\hat{x}^{(k+1)} = \hat{x}^{(k)} + \Delta\hat{x}^{(k)}$.
  - *Renormalization.* Set $h^{(k+1)} = h^{(k)} - J\Delta\hat{x}^{(k)}$ (equivalently, set $h^{(k+1)} = h^{(0)} - J\hat{x}^{(k+1)}$). Set $k \leftarrow k + 1$.
  - *Stopping Condition.* If $||\Delta\hat{x}^{(k)}|| < \tau$, terminate loop. (alternatively, if $||h^{(k)}||/||h^{(0)}|| < \tau$, then terminate). Otherwise, repeat loop.
- *Output.* State estimate $\hat{x}^{(k)} \approx E_\mu\{\mathrm{x}\}$ (renormalized model $(h^{(k)}, J)$ with $h^{(k)} \approx 0$).

This may be viewed as performing a Richardson iteration for solving the sparse linear system $J\hat{x} = h$ for $\hat{x}$ given (h, J) where the RCM procedure takes the place of multiplication by an inverse preconditioner matrix M^{-1} (a tractable linear operator approximating multiplication by J^{-1}). This iteration may be written as

$$\begin{aligned} \hat{x}^{(k+1)} &= (I - M^{-1}J)\hat{x}^{(k)} + M^{-1}h \\ &= \hat{x}^{(k)} + M^{-1}(h - J\hat{x}^{(k)}) \\ &= \hat{x}^{(k)} + M^{-1}h^{(k)} \end{aligned} \tag{4.18}$$

which is equivalent to renormalization if we identify RCM as the preconditioner, defining M such that $M^{-1}h^{(k)} \equiv RCM(h^{(k)}, J)$.

At first glance, it may not be apparent that the estimate $\hat{x}^{(RCM)}$, obtained by running two-pass RCM with inputs (h, J), is actually linear in h. In fact, it is only linear insofar as our moment matching subroutine (employed during each of the model thinning m-projections) is precise.[6] Then, moment matching may be seen as computing $h_\Lambda^{(\text{thin})} = J_\Lambda^{(\text{thin})}\hat{x}_\Lambda$ where $\hat{x}_\Lambda = J_\Lambda^{-1}h_\Lambda$ are the inferred means, of some cavity/blanket model (h_Λ, J_Λ) being thinned, and $J_\Lambda^{(\text{thin})}$ is the thinned information matrix (the computation of which is independent of h_Λ). Hence, in each m-projection, $h_\Lambda^{(\text{thin})} \approx J_\Lambda^{(\text{thin})}J_\Lambda^{-1}h_\Lambda$ which is (approximately) linear in h_Λ (at least upto our moment matching tolerance). Consequently, the global computation of $\hat{x}^{(RCM)} = RCM(h, J)$ is likewise linear in h. Note also, in adjusting the input h we do not change the inverse covariance structure of any of our cavity/blanket/marginal models. Hence, Gaussian iterative renormalization only refines the RCM estimates of the mean parameters.

[6] We could force this by setting our moment-matching tolerance to zero, which would force the iterative moment matching subroutine to continue until all moments are matched to machine precision.

The RCM estimates of marginal covariances remain fixed throughout the iteration.[7]

It is known that the this iteration will converge to the mean if and only the spectral radius[8] of $(I - M^{-1}J)$ is less than one. Moreover, the spectral radius characterizes the asymptotic rate of convergence. In view of our use of m-projections to give a hierarchical, tractable, near-optimal model of the inverse covariance structure of the GMRF, we expect RCM to give a very efficient preconditioner in this regard (some examples in Section 4.3 seem to support this intuition). We also remark the possibility of using Gaussian RCM as a preconditioner for solving arbitrary sparse positive-definite linear systems (not necessarily conceived of as GMRFs), perhaps by accelerated Krylov subspace methods such as the conjugate gradient method [63]. This suggests a much broader range of potential applications for RCM not limited to inference in GMRFs [104, 48, 81, 16]. Also, our renormalization view of this iteration (in GMRFs) suggests a more general approach applicable for other (non-Gaussian) families of MRFs where approximate inference methods, such as two-pass RCM, might be improved upon by iterative adjustment of the ground state implicit in our choice of potential specifications.

We postpone giving examples of this approach to Section 4.3 where we show that this iterative renormalization version of Gaussian RCM has the advantage that the marginal mean parameters $\hat{x}$ (as estimated by RCM) converge to the true means. This iterative approach, however, does not refine the estimated covariance of marginal distributions. This is the related iterative method described next.

4.2.2 Remodeling

This section presents an alternative iterative extension of the RCM approach. The methods discussed thus far provide tractable m-projections by conditioning on some guess for the state along the boundary of a subfield while performing model thinning operations with respect to that subfield. Here, we consider a less heavy-handed type of approximation. Rather then assuming a specific guess for the state of the boundary, we assume a tractable model for the boundary which captures our knowledge of the state while retaining the essential uncertainty of the state. These models are naturally provided by the cavity and blanket models already constructed by our two-pass RCM approach. Here, we show how the model thinning m-projections performed by RCM may be modified so as to exploit these previously constructed cavity and blanket models while selecting refined cavity and blanket models. This results in a

[7] This also indicates that Gaussian RCM could be accelerated by implementing an "on-line" version of the means computation, so as to reuse stored copies of the information matrices of all cavity/blanket models computed earlier in an "off-line" computation based on two-pass RCM ran with zero input, $h = 0$. Then, later iterations may employ the "on-line" computation, replacing each m-projection step $(h_\Lambda, J_\Lambda) \to (h_\Lambda^{(\text{thin})}, J_\Lambda^{(\text{thin})})$ by a simpler calculation: (i) solve $J_\Lambda \hat{x}_\Lambda = h_\Lambda$ for $\hat{x}$ either by recursive inference methods or by employing standard iterative methods (perhaps accelerated using a local preconditioner also computed off-line) and (ii) calculate $h_\Lambda^{(\text{thin})} = J_\Lambda^{(\text{thin})} \hat{x}_\Lambda$.

[8] The *spectral radius* of an $n \times n$ matrix A is defined as the maximum absolute value of the eigenvalues of A. Equivalently, this may be defined as $\rho(A) = \max_{v \in \mathcal{R}^n} \frac{||Av||}{||v||}$ $(v \neq 0)$ where $||v|| = \sqrt{v'v}$ is the Euclidean norm.

modified model thinning step in both the upward and downward passes of our two-pass inference procedure. Applying the modified two-pass procedure repeatedly then gives an iterative method we call *iterative remodeling.*

Cavity Remodeling. Let us first reconsider the upward cavity-modeling pass of RCM. Recall that the intent of each cavity model is to provide approximation for the surface of a subfield for the sake of inference *outside* that subfield. That is, the cavity model gives a simplified model for the interactions within a subfield which is adopted while inferring other parts of the field. Hence, we should take into account the interaction of the subfield with the rest of the field while selecting our approximation for that subfield. A tractable approach is to make use of any existing blanket model for the subfield, together with the known interactions between the surface and the blanket, while selecting our refined cavity model approximation. This idea is illustrated in Figure 4-8. We refer to this figure in the following discussion.

We now describe how the available blanket model can be exploited while thinning the cavity model. Consider the situation shown in Figure 4-8(a). As before, we have some initial cavity model $\hat{\mu}^s_\alpha$ for the surface Λ^s_α of dissection cell Λ_α at node α of the dissection tree. This model was produced by variable elimination, either in the cavity initialization step (Figure 4-3(b)) or the region merging step (Figure 4-4(c)), and has some fill edges we would like to prune. In the remodeling approach, we first join this cavity model $\hat{\mu}^s_\alpha$ with a previously constructed (thin) blanket model $\tilde{\mu}^b_\alpha$ reinstating interactions ψ between these. This gives a joined model

$$\mu^{s,b}_\alpha(x_{\Lambda^s_\alpha}, x_{\Lambda^b_\alpha}) \propto \hat{\mu}^s_\alpha(x_{\Lambda^s_\alpha})\tilde{\mu}^b_\alpha(x_{\Lambda^b_\alpha})\psi(x_{\Lambda^s_\alpha}, x_{\Lambda^b_\alpha}) \tag{4.19}$$

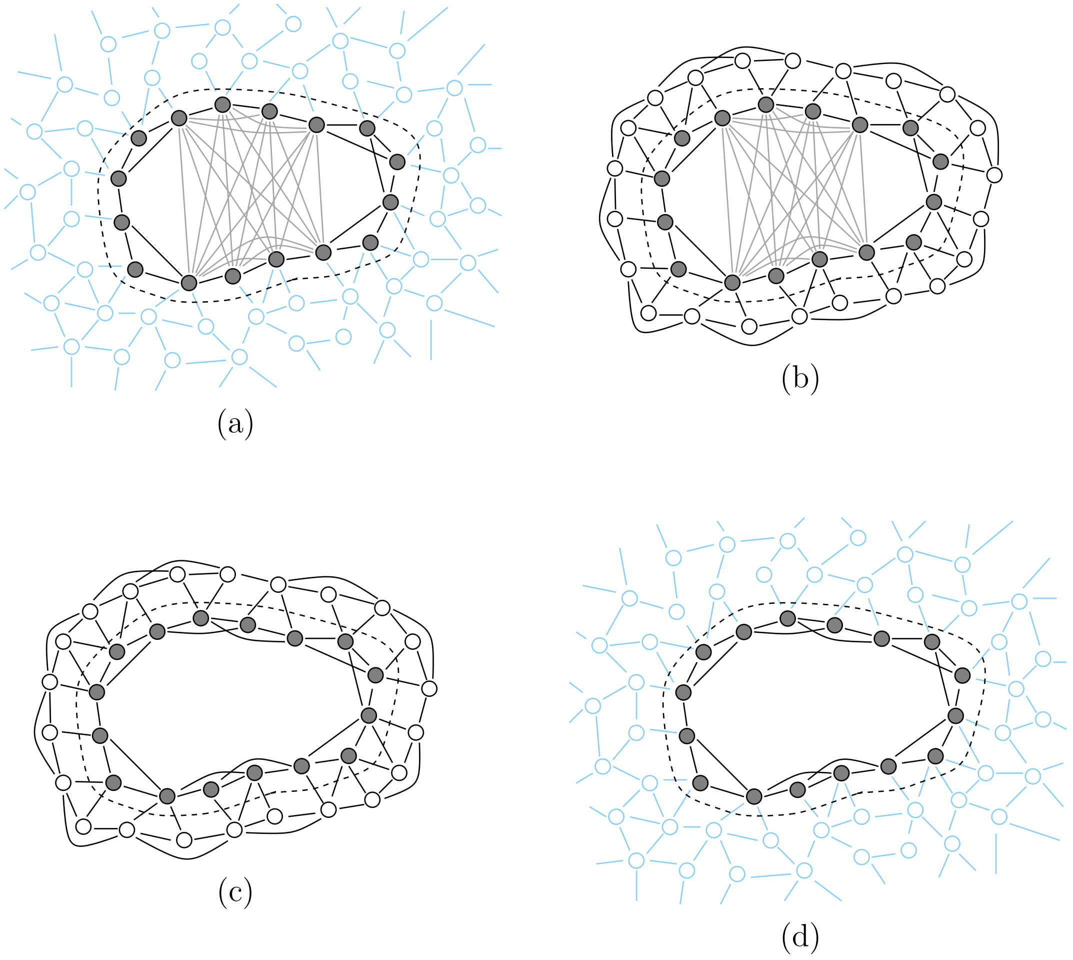
for both the surface and the blanket of the subfield such as shown in Figure 4-8(b). For instance, in GMRFs, joining $\hat{\mu}^s_\alpha = (h^{\text{elim}}_{\Lambda^s_\alpha}, J^{\text{elim}}_{\Lambda^s_\alpha})$ and $\tilde{\mu}^b_\alpha = (h^{\text{thin}}_{\Lambda^b_\alpha}, J^{\text{thin}}_{\Lambda^b_\alpha})$ gives $\mu^{s,b}_\alpha = (h^{\text{join}}, J^{\text{join}})$ with information parameters

$$h^{\text{join}} = \begin{pmatrix} h^{\text{elim}}_{\Lambda^s_\alpha} \\ h^{\text{thin}}_{\Lambda^b_\alpha} \end{pmatrix} \tag{4.20}$$

$$J^{\text{join}} = \begin{pmatrix} J^{\text{elim}}_{\Lambda^s_\alpha} & K \\ K' & J^{\text{thin}}_{\Lambda^b_\alpha} \end{pmatrix} \tag{4.21}$$

with reinstated interactions $K \equiv J_{\Lambda^s_\alpha, \Lambda^b_\alpha}$.

Now we thin this joined model to remove weak interactions from the cavity model. However, since the blanket model is only acting as a “stand-in” for the rest of the (intractable) graphical model, we still only perform moment matching within the cavity. This means that only the parameters of the cavity model are adjusted, but these are adjusted so as to preserve moments computed under the joined model. Hence, the only influence the blanket model has is in the calculation of moments inside the cavity. The model thinning step insures that these moments are the same before and after pruning edges from the model. This gives our thinned cavity model $\tilde{\mu}^s_\alpha$ as illustrated in Figure 4-8(c).



(a)

(b)

(c)

(d)

Figure 4-8: Illustration of model thinning in upward cavity-remodeling pass. A previously constructed blanket model is joined with our cavity model while thinning edges in the cavity model. That is, all moment calculations are performed with the blanket model attached but only the parameters of the cavity model are adjusted during moment matching. Once thinning is complete, the cavity model is extracted and may then be used to infer other parts of the field. Essentially, the intent of this approach is to get from (a) to (d) while incurring minimal KL-divergence. The blanket model functions as a tractable substitute for the exterior field while selecting a favorable cavity model.

In GMRFs, this means that we solve for the thinned joint model $(h^{\text{thin}}, J^{\text{thin}})$ which satisfies the following conditions:

1. The blanket and surface-to-blanket interactions are held fixed.
$$\begin{aligned} h^{\text{thin}}_{\Lambda^b} &= h^{\text{join}}_{\Lambda^b} && (4.22)\\ J^{\text{thin}}_{\Lambda^b} &= J^{\text{join}}_{\Lambda^b} && (4.23)\\ J^{\text{thin}}_{\Lambda^s,\Lambda^b} &= J^{\text{join}}_{\Lambda^s,\Lambda^b} && (4.24)\end{aligned}$$

2. The cavity model is thinned so as to respect a thinned interaction graph $\mathbf{G}^{\text{thin}}_{\Lambda^s} = (\Lambda^s, \mathcal{E}^{\text{thin}}_{\Lambda^s})$. That is, we require
$$J^{\text{thin}}_{\gamma,\lambda} = 0 \qquad (4.25)$$
for all $\gamma, \lambda \in \Lambda^s$ where $\{\gamma,\lambda\}$ *is not* an edge of $\mathbf{G}^{\text{thin}}_{\Lambda^s}$, $\{\gamma,\lambda\} \notin \mathcal{E}^{\text{thin}}_{\Lambda^s}$.

3. The moments of the thinned cavity model are held fixed. That is we match the mean parameters
$$((J^{\text{thin}})^{-1}h^{\text{thin}})_{\Lambda^s} = ((J^{\text{join}})^{-1}h^{\text{join}})_{\Lambda^s} \qquad (4.26)$$
and match the covariance parameters
$$(J^{\text{thin}})^{-1}_{\gamma,\lambda} = (J^{\text{join}})^{-1}_{\gamma,\lambda} \qquad (4.27)$$
for all $\gamma, \lambda \in \Lambda^s$ where either $\gamma = \lambda$ or $\{\gamma,\lambda\}$ *is* an edge of $\mathbf{G}^{\text{thin}}_{\Lambda^s}$, $\{\gamma,\lambda\} \in \mathcal{E}^{\text{thin}}_{\Lambda^s}$.[9]

Together, these conditions uniquely determine the thinned cavity model $\tilde{\mu}^s_\alpha$ specified by the information parameters $(h^{\text{thin}}_{\Lambda^s}, J^{\text{thin}}_{\Lambda^s})$. This corresponds to performing m-projection to the e-flat submanifold specified by conditions (1) and (2). We solve for this m-projection by performing our LIS moment-matching technique within this exponential family until the moment constraints (3) are met.[10] This also gives the maximum-entropy distribution subject to just (1) and (3).

Note that, once the model thinning step is completed, the blanket model is "deleted", only the thinned cavity model for the surface of the cell is retained. This, in turn, is used to infer other parts of the field such as suggested in Figure 4-8(d). Essentially, in going from (a) to (d), we only employ the blanket model as a temporary tractable substitute for the rest of the field in order to guide our selection of a thinned cavity model.

Blanket Remodeling. Likewise, in the downward blanket-modeling pass of RCM, we design each blanket model to yield a simplified model of the interactions *outside*

[9] Note that we do not condition on any state of the blanket in the remodeling approach. The moments preserved during model thinning actually correspond (approximately) to the actual moments calculated under the full graphical model.

[10] This actually is a very minor modification of the moment matching procedure used in the initial cavity modeling procedure. The only additional complexity is in the moments calculation which requires inference of the joined model of both the surface and the blanket.

of a given subfield for the sake of performing tractable inference *inside* that subfield. Hence, in selecting our thinned approximation for the blanket model, we should take into account the interaction with that subfield. A tractable approach is to make use of any existing cavity model for the subfield, together with the known interactions between the surface and blanket of the subfield, while selecting our refined blanket model approximation. This idea is illustrated in Figure 4-9.

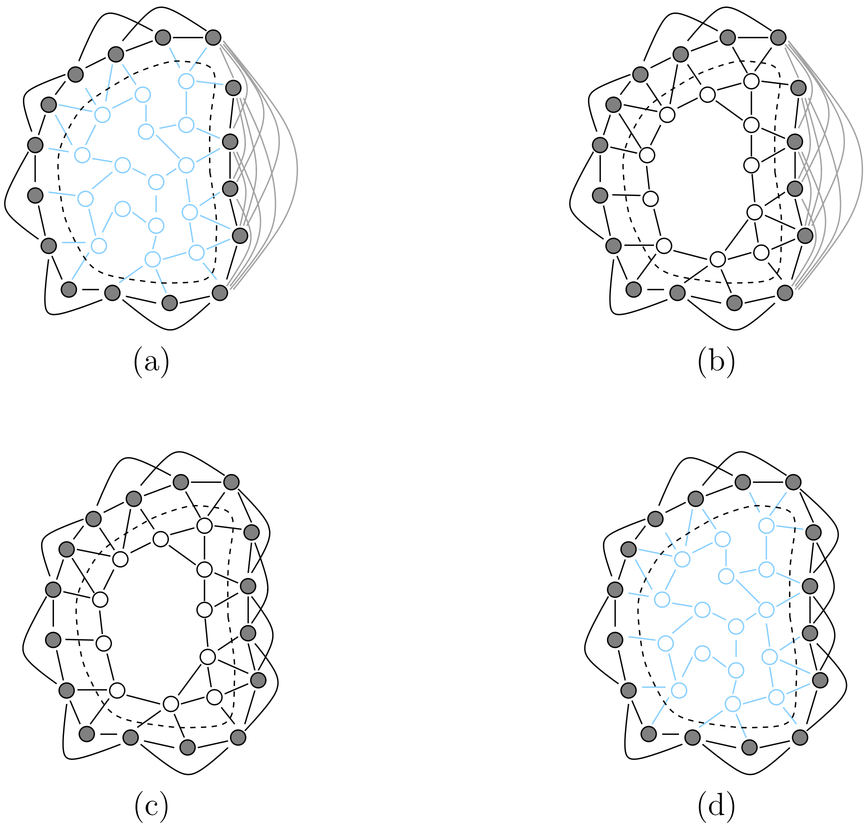
The model thinning operation in the downward blanket-remodeling pass is modified in a similar manner as in the upward cavity-remodeling pass but where the roles of the cavity and blanket models are reversed. In Figure 4-9(a), we have an initial blanket model we would like to thin. While selecting a thinned model of the blanket, we substitute a previously constructed cavity model for the enclosed subfield as shown in Figure 4-9(b). This joined cavity-blanket model is then thinned while holding the parameters of the cavity model fixed, only adjusting the parameters of the blanket so as to hold fixed the corresponding moment characteristics. This yields a thinned blanket-cavity model such as shown in Figure 4-9(c). Finally, the original subfield may be substituted in place of the cavity model so that our thinned blanket model supports inference of the enclosed subfield such as indicated in Figure 4-9(d).

This completes specification of our remodeling version of two-pass RCM. Iterating this two-pass remodeling approach then gives an iterative version of RCM where we might hope that the method converges to a stable set of cavity/blanket models giving improved approximations of the marginal models provided by RCM. We examine this further in some simulated examples given at the end of the next section.

4.3 Simulations

In this section we demonstrate the RCM approach to inference in GMRFs for some simulated image processing examples. This both clarifies and motivates the methods developed in this thesis and also shows that RCM can provide a tractable yet near-optimal inference approach. We first specify a prior image model and pose three image restoration problems based on noisy, partial observations of random images distributed according to this prior model. We then show the results of applying the basic two-pass version of RCM for three such simulated examples. Some diagrams are given to show the thinned structure of the cavity and blanket models developed during this procedure. The quality of the estimated pixel values (conditional means) and uncertainties (square-roots of conditional variances) are compared to the exact values computed by recursive inference methods.[11] Next, we examine the performance of the iterative renormalization method and show that this iteratively refines the quality of the mean estimates to machine precision. The estimated uncertainties, however, are not refined by this procedure. Finally, we look at the performance of the iterative remodeling approach which has the advantage that *both* the mean estimates and the

[11] Here, for the sake of verification, we look at an image processing example which is sufficiently small to allow exact *recursive* inference. Exact calculations are performed by running RCM without any model thinning ($\delta = 0$) which is much slower than in the thinned version of the algorithm but still much faster than the "brute-force" calculation $(\hat{x}, P) = (J^{-1}h, J^{-1})$.


(a)

(b)

(c)

(d)

Figure 4-9: Illustration of model thinning in downward blanket-remodeling pass. A previously constructed cavity model is joined with our blanket model for the purpose of thinning edges in the blanket model. That is, all moment calculations are performed with the cavity model attached but only parameters of the blanket model are adjusted during moment matching. Once thinning is complete, the blanket model is extracted and may then be used to infer the enclosed subfield. Essentially, the intent of this approach is to get from (a) to (d) while incurring minimal KL-divergence. The cavity model functions as a tractable substitute for the enclosed subfield while selecting a favorable blanket model.

uncertainty estimates are refined.

Prior Image Model. First, we specify our prior image model. In these examples, we consider a 64×64 image with real-valued scalar pixels. We model our image as being a sample of a GMRF x_Γ with sites $\Gamma = \{(i,j) | i = 1, \ldots, 64,\ j = 1, \ldots, 64\}$ arranged on a 2-D square grid. We assume nearest-neighbor interactions such that x_Γ is Markov with respect to the graph $\mathbf{G}_\Gamma = (\Gamma, \mathcal{E}_\Gamma)$ with edge set $\mathcal{E}_\Gamma$ comprised of vertical edges $\{(i,j),(i,j+1)\}$ and horizontal edges $\{(i,j),(i+1,j)\}$ linking adjacent vertices in the grid. Our prior image model has zero-mean and sparse information matrix (inverse covariance) $J = P^{-1}$ designed to respect the conditional independencies imposed by $\mathbf{G}_\Gamma$. We construct J as a sum of pairwise interactions

$$J^E = \begin{pmatrix} 1.0 & -0.99 \\ -0.99 & 1.0 \end{pmatrix} \tag{4.28}$$

for each edge $E = \{\gamma, \lambda\} \in \mathcal{E}_\Gamma$. That is,

$$J_{\gamma,\lambda} = \sum_{E \in \mathcal{E}_\Gamma} (J^E)_{\gamma,\lambda} \tag{4.29}$$

where $(J^E)_{\gamma,\gamma} = (J^E)_{\lambda,\lambda} = 1.0$, $(J^E)_{\gamma,\lambda} = (J^E)_{\gamma,\lambda} = -0.99$, and is otherwise zero. Hence, J has diagonal values of 4.0 (at sites corresponding to interior nodes of the grid), 3.0 (along the edges of the grid), and 2.0 (at the corners of the grid); and has off-diagonal values of either -0.99 (at locations corresponding to edges) or zero (between non-adjacent sites). In this GMRF, the partial correlation coefficients between interior sites are $\rho = 0.2475.$[12]. This gives a symmetric, positive-definite matrix with reciprocal condition number ≈ 0.004 (this is the ratio between the smallest and largest eigenvalues of J such that small values indicate a nearly singular matrix). We design our model in this way in order to yield a GMRF with positive, nearly-uniform interactions where the strength of interactions are made about as strong as possible while still giving a non-singular information matrix.

All examples are based on a randomly generated sample of our prior image model. We simulate a sample of this image model by the following method due to Rue [119]. First, an ordering is adopted for the the sites Γ so as to give a low bandwidth representation of the information matrix.[13] Next, the Cholesky factorization of J is computed. This gives a sparse, low bandwidth, upper triangular matrix R such that $J = R'R$. Then, we simulate a vector of 64 independent, identically distributed standard Gaussian deviates $w = (w_k, k = 1, \ldots, 64^2)$ such that $\mathrm{w} \sim \mathcal{N}(0, I)$. Finally, we solve $Rx = w$ employing standard iterative methods.[14] This gives a zero-mean Gaussian random vector with covariance $E\{\mathrm{xx}'\} = R^{-1}E\{\mathrm{ww}'\}(R')^{-1} = R^{-1}(R')^{-1} =$

[12] The partial correlation coefficients are somewhat larger at the edges and corners of the grid (i.e., $\rho = 0.29, 0.33, 0.40$).

[13] We used the symmetric reverse Cuthill-McKee permutation, computed by the Matlab subroutine symrcm.

[14] We used the generalized minimum-residual method available as a Matlab subroutine gmres.

$(R'R)^{-1} = J^{-1}$ as desired. A sample image generated in this fashion is shown at top-left in Figure 4-10. We use this same sample image in all three of the following image restoration examples.

Three Image Restoration Examples. We now pose three image-restoration problems based on the preceding prior image model. In each of our three examples, we generate a noisy measurement of the state of some site γ according to the measurement model,

$$\mathrm{y}_\gamma = \mathrm{x}_\gamma + \mathrm{v}_\gamma \tag{4.30}$$

where the measurement noise is generated according to the zero-mean Gaussian distribution $v_\gamma \sim \mathcal{N}(0, \sigma^2)$ with standard deviation σ. We then consider the following image restoration problems:

1. *All Pixels Observed.* First, we consider the case where we have noisy observations of every pixel in the image. In this case, we set $\sigma = 1.0$ such that there is a significant level of noise relative to the variance in the prior image model. A set of measurements simulated in this way is shown at top-right in Figure 4-10.

2. *Sparse Observations.* Second, we consider the case where we have a high rate of "data drop-out" such that only a fraction of the pixels in the image are observed. We model the occurrence of observations as being independent at each pixel and occurring with probability 0.05 such that about five percent of the pixels are observed. In this case, we set $\sigma = 0.1$ so that these sparse observations are more accurate than in the preceding fully-observed case. A set of measurements simulated in this way is shown at bottom-left in Figure 4-10.

3. *Boundary Observations.* Lastly, we consider the case where we have noisy measurements only along the boundary of the image. That is, we have observations at just those site along the edges of the grid. In this case we consider near-perfect observations with $\sigma = 0.01$. A set of measurements simulated in this way are shown at bottom-right in Figure 4-10.

In all three examples we simulate noisy observations of the same underlying image (the previously simulated sample of our prior image model) but the measurement noise is independently simulated in each case. We then wish to calculate, in each example, the conditional marginal distribution $p(x_\gamma|y)$ of the state at each pixel x_γ conditioned on all available observations y. This is a Gaussian distribution $\mathrm{x}_\gamma|y \sim \mathcal{N}(\hat{x}_\gamma(y), \hat{\sigma}^2_\gamma)$ specified by the conditional mean $\hat{x}_\gamma(y) = E\{\mathrm{x}_\gamma|y\}$ and the conditional variance $\hat{\sigma}^2_\gamma = E\{(\mathrm{x}_\gamma - \hat{x}_\gamma(y))^2\}$. This may also be posed as marginalization of the conditional distribution $p(x_\Gamma|y) \propto \exp\{-\frac{1}{2}x'_\Gamma \hat{J} x_\Gamma - \hat{h}' x_\Gamma\}$ where $\hat{h}$ is specified by the measurements

$$\hat{h}_\gamma = \begin{cases} y_\gamma/\sigma^2, & \gamma \,\text{observed.} \\ 0, & \gamma \,\text{not observed.} \end{cases} \tag{4.31}$$

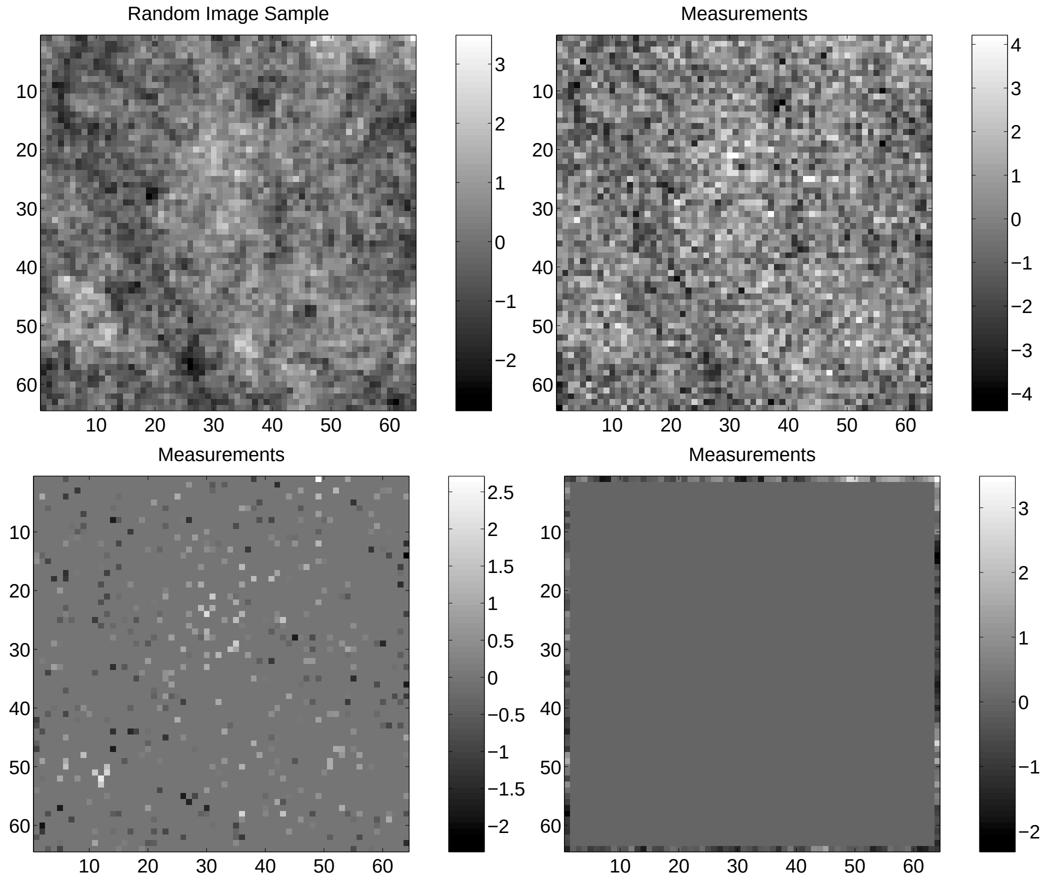


Figure 4-10: Three simulated image restoration problems. First, we simulate a sample of our prior image model (top left). Then, we simulate three noisy and/or partially observed versions of this image either having: noisy measurements ($\sigma = 1.0$) of each pixel (top right); sparse measurements ($\sigma = 0.1$) for a randomly selected subset of pixels (bottom left), or just measurements ($\sigma = 0.01$) of those "boundary" pixels going around the perimeter of the image (bottom right). The measurement noise is independent in each of these three simulations. Given any one of the observation images we should like to estimate (restore) the underlying sample image and also estimate the pixel-by-pixel uncertainty in our restoration.

and where $\hat{J}$ has diagonal elements given by

$$\hat{J}_{\gamma,\gamma} = \begin{cases} J_{\gamma,\gamma} + 1/\sigma^2 & \gamma \, \text{observed.} \\ J_{\gamma,\gamma}, & \gamma \, \text{not observed.} \end{cases} \tag{4.32}$$

and off-diagonal elements given by $\hat{J}_{\gamma,\lambda} = J_{\gamma,\lambda}$ for all $\gamma \neq \lambda$. We may also use RCM in place of an exact marginalization procedure.

Two-Pass "Block" RCM. We now discuss how we use the RCM inference procedure to perform image restoration (e.g., estimation of the underlying image from the available observations) in these examples. We have actually found it somewhat beneficial to perform a "block" version of RCM for GMRFs with scalar-valued states. This means that we first convert the original 64 × 64 GMRF (having scalar-valued states at each site) into an equivalent 16 × 16 GMRF (having 16-D vector-valued states corresponding to 4×4 subfields of the original GMRF). We then run RCM for this latter GMRF having 16-D states at each site which produces a marginal model for each 4×4 patch of the image (corresponding to a site in the "blocked" GMRF). It is then straight-forward to calculate the single-pixel distributions within each of the 4×4 marginal models yielding the desired estimates of the means $\hat{x}_\gamma$ and associated uncertainties $\hat{\sigma}_\gamma$ for every pixel $\gamma \in \Gamma$ of the original 64 × 64 image.

Cavity and Blanket Models. We have executed this blocked version of RCM for each of our three examples with precision parameter $\delta = 10^{-4}$ controlling the complexity of the cavity and blanket models developed during the procedure and with moment matching tolerance $\epsilon = 10^{-12}$ controlling the precision of our iterative moment matching subroutines. We show selected "snapshots" of this procedure for the fully-observed example in Figure 4-11 (the other examples are similar). In all of these "blocked" examples, the choice of precision parameter $\delta = 10^{-4}$ leads to the selection of mostly singly-connected chains and loops going around the boundary of each subregion (but sometimes adds an extra edge at the "corners" of these subregions).

However, in this blocked version of RCM, each node actually represents a 4×4 patch of the image. Hence, the cavity and blanket models shown are actually rather more complex than they appear. Each node represents a fully-parameterized 4×4 Gaussian subfield (represented by a 16-D influence vector and 16×16 symmetric information matrix) and each edge actually represents 16^2 interactions between the coupled 4×4 subfields. Hence, in the moment-matching subroutines, iterative scaling adjusts all of these parameters to preserve a corresponding set of moments including cross-covariances between every pair of pixels contained in either the same 4×4 block or in adjacent 4×4 blocks. Note also that, in this blocked approach, RCM essentially treats the Markov blanket of a given subfield as the "fat" boundary of that subfield of width 4 (i.e., all pixels within 4 steps of that subfield in the original 64×64 model). Essentially, our cavity and blanket models are actually augmented by three extra layers of latent variables. Hence, the cavity and blanket models shown, while having rather simple graphical structure, actually represent very rich, precisely refined models for these "fat" Markov blankets.

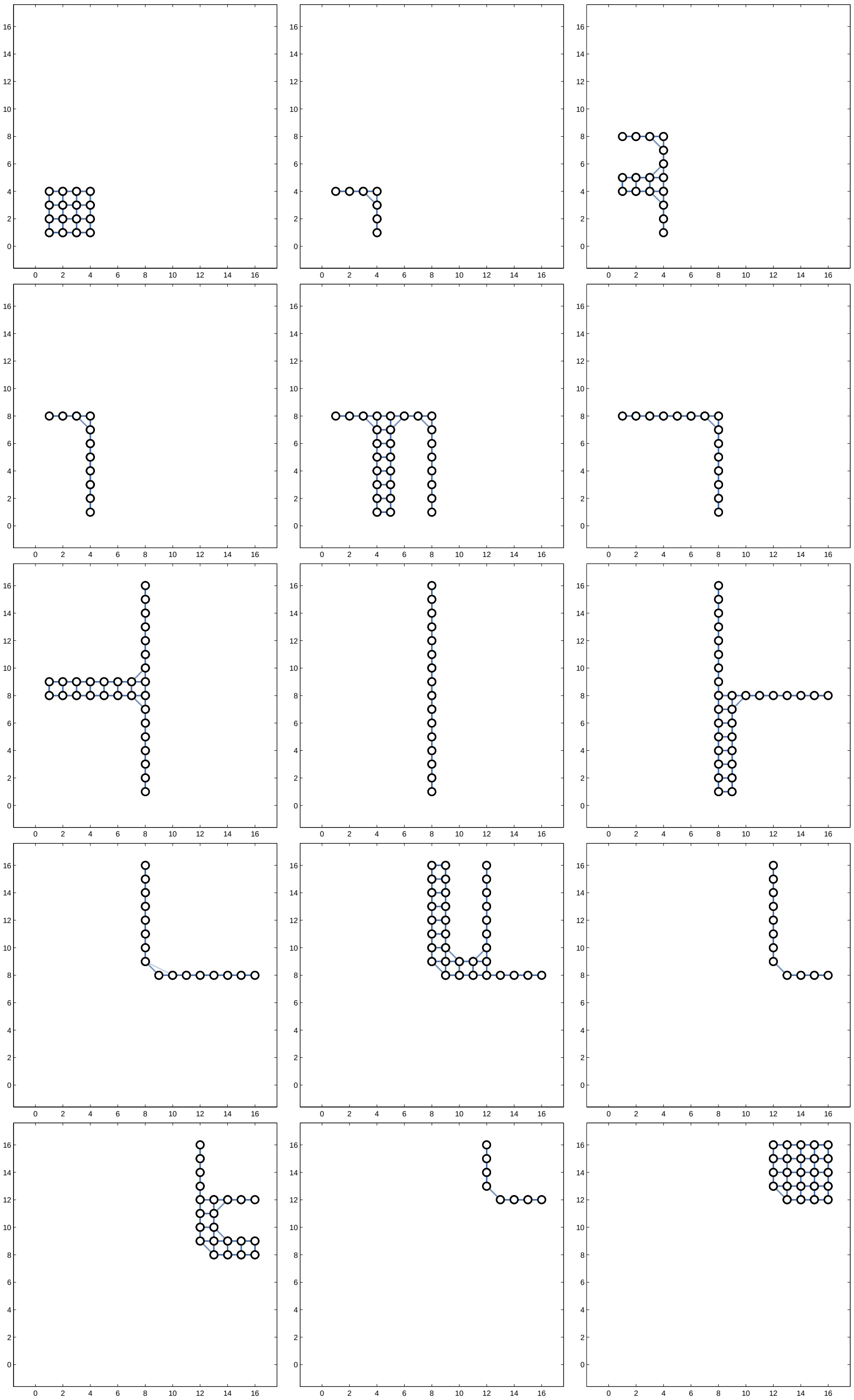


Figure 4-11: Selected snapshots of Two-Pass RCM in 16 × 16 blocked GMRF (each node represents a 4 × 4 patch of the image). We display 8 frames from the upward pass followed by 7 frames from the downward pass. This shows how RCM propagates information from the lower-left corner to the upper-right corner.

Choice of Block Size. Of course, selection and inference of such rich families of cavity/blanket comes at a heavy computational expense which limits how large a block size we can use. While the actual number of "block" computations[15] is reduced (roughly by a factor of $4^2 = 16$), each such operation now requires roughly $4^3 = 64$ floating point calculations. Hence, one might expect 4×4 blocking to increase run-time by about a factor of 4. Paradoxically, we actually find that a *moderate* amount of blocking actually tends to *reduce* the execution time of RCM. Apparently, this is due to the vector pipelining effect which allows a "pipelined" sequence of floating-point calculations, performed sequentially without interruption on data stored in contiguous memory, to execute in an accelerated manner.[16] Hence, the cost of any "extra" floating point calculations arising due to blocking is not nearly as onerous as one might initially expect. We should also remark that this effect is probably exaggerated due to our Matlab implementation of RCM.[17] Hence, perhaps smaller block sizes would be preferable in a more efficient compiled version of RCM but we would still expect some blocking to prove beneficial. In any case, once we have made the block size large enough to saturate these effects, further increase of block size is not warranted and will only slow down our RCM approach. For instance, in these examples, the 4×4 block size apparently saturates these effects. Indeed, switching to 8×8 blocks increases the run-time of RCM by about a factor $4^3 = 64$.

Computational Complexity. Based on the observation that, in these image processing examples, RCM constructs very thin graphical models for cavity and blanket models, coming close to either a Markov chain (for those dissection cells along the boundary of the image) or a "long loop" (for those dissection cells embedded in the interior of the image), let us give a "back of the envelope" estimate of the computational complexity of RCM as a function of the dimension $N = W \times W$ of the image (N is the number of pixels in a $W \times W$ image).[18] For simplicity, let us consider images where $W = 2^k$ (for $k = 1, 2, \ldots$) and estimate the computational complexity of

[15] For instance, elimination of a 4×4 block in the thinning and inference subroutines or the computation and accrual of an IPF update for a 4×4 block (or pair of adjacent blocks) in the moment-matching subroutine.

[16] Modern digital processors are adept at optimizing such pipelined calculations by breaking down the floating point operation (flop) into a sequence of $N_{\text{flop}} \approx 16$ suboperations each implemented by a separate unit on the microchip. For a single flop, these suboperations must be performed sequentially (requiring N_{flop} cycles per flop). Yet, these suboperations may be performed *in parallel* for a pipelined sequence of flops (requiring an average of $1/N_{\text{flop}}$ cycles per pipelined flop). Essentially, this means that N_{flop}^2 pipelined flops require about the same amount number of cycles as N_{flop} non-pipelined flops. This is the motivation for the use of block representations of sparse matrices and the associated "block" versions of matrix computation algorithms often employed in iterative methods (see Golub and Van Loan [63]).

[17] Essentially, blocking allows efficient compiled subroutines to do most of the work placing less burden on higher-level control mechanisms of RCM executed by the Matlab interpreter. As the Matlab interpreter is notoriously slower than compiled C code, this displacement of computation towards lower-level compiled subroutines can also significantly reduce execution times.

[18] The effect of vector-pipelining, however, is neglected in this analysis. Essentially, an appropriate amount of blocking can at best reduce execution time by a factor of $1/N_{\text{flop}}$ where N_{flop} indicates the degree to which floating point computations can be done in parallel.

RCM for such images assuming such thin chain-like and loop-like cavity and blanket models.

The brunt of the computation occurs in the iterative moment matching subroutine used to calculate our m-projections (and associated moment calculations). Each iteration of our LIS technique calls an inference subroutine for the cavity or blanket model being calculated, computes a set of local IPF updates, and fuses these IPF updates to yield the LIS update for the parameters of the graphical model. Assuming maximal cliques in all of our cavity and blanket models stay below some specified size, each of these steps is linear in the size of the cavity or blanket model (the total number of nodes in the cavity or blanket model).[19] So let us assume that each iteration of moment matching requires $\mathcal{O}(\alpha(\delta)s)$ computations where s is the size of the cavity or blanket model and $\alpha(\delta)$ indicates the dependence of the scaling factor on our choice of δ parameter. For GMRFs, we estimate $\alpha(\delta) \sim \mathcal{O}(m^3(\delta))$ where $m(\delta)$ is the state dimension of maximal cliques in our cavity and blanket models. This maximal clique size will grow as δ becomes small. Very roughly, we estimate that $m(\delta) \sim \mathcal{O}(\log \delta^{-1})$ as δ approaches zero.[20]

In the case of singly-connected Markov chains, our LIS method actually is exact after just one iteration. In the case of loopy graphs, several iterations are usually required to obtain a given tolerance. Hence, let us simply suppose that a constant number $c(\epsilon)$ of iterations always suffices to match moments to tolerance ϵ. Since, in loopy graphs, LIS seems to gain significant digits linearly in the number of iterations (see experiments in Chapter 3), the author roughly estimates that $c(\epsilon) \sim \mathcal{O}(\log \epsilon^{-1})$ as ϵ approaches zero. Hence, we estimate that $\mathcal{O}(c(\epsilon)m^3(\delta)s) \approx \mathcal{O}(\log \epsilon^{-1} \log^3 \delta^{-1} s)$ computations are required to calculate a cavity or blanket model with s nodes.

Under these assumptions, we conclude that the total computation of RCM performed at each level of dissection (for fixed δ, ϵ) is linear in the number of "surface" nodes produced by dissection. Roughly speaking, the total number of surface nodes doubles at each level of dissection, with $\mathcal{O}(W)$ surface nodes at the first level of dissection and $\mathcal{O}(N)$ surfaces nodes at the last level of dissection. Hence, we estimate that the total computation is proportional to $N + N/2 + N/4 + \cdots + N/W < N(1 + 1/2 + 1/4 + \cdots) = 2N$. That is, most of the computation is performed at or near the finest level of dissection and the total computation is bounded above by an $\mathcal{O}(N)$ function.

Hence, we arrive at our estimate $\mathcal{O}(c(\epsilon)m^3(\delta)N) \approx \mathcal{O}(\log \epsilon^{-1} \log^3 \delta^{-1} N)$ of the computational complexity of RCM for a square image with N pixels. This indicates a scalable approach to inference which is linear in the number of pixels N in the image. This also shows that doubling the moment matching precision (replacing $\epsilon \ll 1$ by ϵ^2) will approximately double run-times. However, we expect that likewise doubling the precision of our model selection criterion (replacing $\delta \ll 1$ by δ^2) will

[19] This requires that we implement the inference subroutine in an appropriate recursive manner to take advantage of the thin structure of these cavity and blanket models.

[20] This estimate is based on the observation that doubling the precision of RCM (replacing δ by δ^2) appears to typically have the effect of doubling the "range" of interactions retained in our model thinning procedures. In the context of these 2-D image processing examples, this has the effect of doubling the size of maximal clique in our essentially 1-D boundary models.

increase run-times approximately by a factor of eight.[21] Hence, this does limit how small we can make δ (thereby limiting the accuracy we can achieve with RCM) while still giving a tractable inference procedure (with a small amount of computation per node). Nevertheless, we shall see in our examples that we can achieve some very accurate results while still keeping the computation per node at a reasonable level.

Accuracy. The results of performing this Block-RCM procedure for each of our three examples are shown in Figures 4-12, 4-13 and 4-14 respectively. In each case, we show the actual image (a random sample of our prior image model), the available noisy/partial observations, the estimated conditional means $\hat{x}_\gamma^{(RCM)}(y)$, and the estimated conditional uncertainties $\hat{\sigma}_\gamma^{(RCM)}$. We have also computed the exact conditional means $\hat{x}_\gamma(y)$ and uncertainties $\hat{\sigma}_\gamma$ for the sake of comparison. The deviation of the RCM estimates from these optimal values are displayed as relative errors (normalized by the actual uncertainty).

$$\begin{aligned}
\text{re1}_\gamma &= \frac{\hat{x}_\gamma^{(RCM)}(y) - \hat{x}_\gamma(y)}{\hat{\sigma}_\gamma} && (4.33)\\
\text{re2}_\gamma &= \frac{\hat{\sigma}_\gamma^{(RCM)} - \hat{\sigma}_\gamma}{\hat{\sigma}_\gamma} && (4.34)\\
&= \left(\frac{\hat{\sigma}_\gamma^{(RCM)}}{\hat{\sigma}_\gamma}\right) - 1 && (4.35)
\end{aligned}$$

These are also displayed as images for each of our three examples.

In all three examples our RCM state estimates agree with the optimal state estimates to at least three or four digits. The uncertainty estimates appear to be somewhat more precise agreeing with the exact values to at least four or five digits. It is also interesting to note that RCM consistently underestimates uncertainties. This is not too surprising since all inferences in RCM are based on approximate cavity and blanket models which are selected by *conditional* m-projections (conditioned on zero boundary conditions). Since conditioning decreases uncertainty, we might expect any inferences made with these cavity and blanket models might tend to make overconfident predictions leading to underestimates of the final uncertainty in each state estimate.

Blocky Artifacts. It is apparent that the relative errors in our estimates tend to be largest near the boundaries and corners of our smallest dissection cells. This is to be expected since the inference calculations performed within each dissection cell are otherwise exact except for the use of an approximate blanket model. Typically, errors due to a misleading blanket model are most apparent at the surface of the

[21] Actually, in order for δ to be indicative of accuracy, we presumably should keep $\epsilon < \delta$ (for instance, by setting $\epsilon = \delta \times 10^{-8}$) so that the cost of δ-precision RCM is actually $\mathcal{O}(\log^4 \delta^{-1} N)$ as δ approaches zero. Then, doubling the precision of RCM actually increases run-time by about a factor of 16.

cell. However, we point out that these "blocky artifacts" are nevertheless sufficiently small so as not to be apparent in the restored images and associated images of the uncertainty. In fact, visually comparing the RCM estimates (and uncertainties) to the optimal estimates (and uncertainties), the author could not visually discern the differences between these images. This is in contrast to seemingly comparable methods in the state-reduced multiscale modeling approach (see examples given in Luettgen [91], Irving [72] and Frakt [55]). Also, the author wishes to emphasize that the severity of our blocky artifacts do not seem to be a function of which "cut" we are near. For instance, the errors near the coarsest cut, where the field is first cut into two halves, do not appear to be any more substantial than those at the finest level of dissection. We think this shows the advantage of our thinned Markov-blanket approach over those state-reduced multiscale modeling methods (see discussion in Section 2.3.3).

It should also be remarked that we could make the relative errors gradually vanish by decreasing the δ parameters in these experiments (at the expense of additional computation). For instance, setting $\delta = 10^{-8}$ produces state and uncertainty estimates accurate to around 9 or 10 digits. Yet, decreasing δ in this way causes the cavity and blanket models developed by the method to gradually become more fully connected[22] so that the computational advantage of thinning is gradually lost.[23] As an alternative approach for improving accuracy (without decreasing δ), we consider our two iterative refinement procedures in the remainder of the section.

Iterative Renormalization. In this section we show the result of applying the iterative renormalization technique to improve the state estimates produced by RCM.

Recall that this corresponds to performing a Richardson iteration with RCM playing the role of a preconditioner. This is initialized by running RCM for our graphical model (h, J) and setting $\hat{x}^{(0)} = \hat{x}^{(RCM)}$. This inference is then iterated generating a sequence of improved image estimates $\hat{x}^{(k)}$ for $k = 1, 2, 3, \ldots$ by rerunning RCM with inputs $(h^{(k)}, J)$ where $h^{(k)} = h - J\hat{x}^{(k-1)}$ is the residual error image of the previous iterate. This produces a correction term $\Delta\hat{x}^{(k)}$ which is added to our image estimate $\hat{x}^{(k)} = \hat{x}^{(k-1)} + \Delta\hat{x}^{(k)}$ seeding the next iteration.

Here, we again perform RCM with parameters $(\delta = 10^{-4}, \epsilon = 10^{-12})$ as in the previous experiments. In each example, we have performed 12 iterations of iterative renormalization. In Figure 4-15, we show the convergence in all three examples by plotting the relative residual error,

$$e^{(k)} = \frac{\|h - J\hat{x}^{(k)}\|}{\|h\|}, \tag{4.36}$$

which indicates how close we have come to solving $J\hat{x} = h$. Note that all three

[22] For instance, in our 4×4 blocked examples, setting $\delta = 10^{-8}$ tends to produce cavity and blanket model where each block is coupled to the two nearest blocks in the boundary on either side of that block producing richer models which require more computation in the inference and moment matching subroutines.

[23] Nevertheless, in much larger fields, we could still allow a substantial amount of fill while keeping computation well below that of the exact version of the algorithm.

examples converge to a relative residual error of about 10^{-16}, indicating convergence to the machine's relative floating point accuracy.[24] Moreover, after the 2nd iteration, the convergence seems to be linear in our logarithmic plots indicating exponential decay of the residual error until machine precision is reached. Each iteration reduces the relative residual error by about three orders of magnitude (a factor of 1000) giving very rapid convergence.

In Figures 4-16, 4-17 and 4-18, we also show images of the relative state estimation errors and of the residual error images $h^{(k)}$ of selected iterates (in these figures, the first pair of images is the result of two-pass RCM shown for comparison). Note that both error images are reduced during the iteration approaching machine precision.

Iterative Remodeling. In this final set of experiments, we consider the iterative remodeling procedure for improving both the state estimates and the uncertainty estimates produced by RCM.

Recall that, in this approach, we modify the RCM procedure to take advantage of the blanket and cavity models calculated on previous iterations of RCM while calculating new cavity and blanket models. We illustrate this procedure for the upward and downward passes of the remodeling procedure in Figures 4-19 and 4-20. The upward pass proceeds as before except that the cavity model thinning step is modified to take advantage of the blanket model from the preceding downward pass. This means that, while thinning, we join the cavity model with the blanket model so that that the moments preserved during m-projection are those computed assuming the blanket model (rather than assuming zero boundary conditions as in two-pass RCM or estimating the state of the boundary as in renormalization). Note, however, that we still only match moments within the cavity. This may be interpreted as approximating a global m-projection where the calculated moments now approximate the true moments (such as would be given by global inference of the entire GMRF were this feasible) but we still only refine parameters within the cavity model to match a corresponding set of locally-defined moments. Once moment matching is complete, the blanket model is deleted yielding our new cavity model and the upward pass continues. Likewise, in the downward remodeling pass, we exploit the previously constructed cavity model (from the preceding upward pass) while thinning the new blanket model. Alternating these upward and downward remodeling passes gives an iterative method. We should like to see if this iteration converges to a stable set of cavity and blanket models and if this yields improved estimates of the pixel-level marginal distributions.

We have executed this iterative remodeling procedure, performing 6 iterations with parameters fixed at $(\delta = 10^{-4}, \epsilon = 10^{-12})$, for all three of our image restoration examples. The result of these iterations are shown in Figures 4-21, 4-22, 4-23 and 4-24. We show the errors in the estimated means and uncertainties relative to the correct means and uncertainties (computed previously by exact recursive inference). In all three examples, the method appears to reach a stable point after 2-4 iterations.

[24] About 2.2×10^{-16}. This is the distance from 1.0 to the next largest number which can be represented exactly by the machines finite-precision floating point number system.

Both the estimated means and uncertainties are improved as shown by the decrease in relative errors with most of the improvement occurring on the first iteration. The improvement in the estimated means is much more substantial than the improvement in the estimated uncertainties. The estimated means are improved by about 5-7 orders of magnitude. However, unlike the iterative renormalization method, the estimated means do not converge to machine precision. The improvement in the estimated uncertainties is moderate in the fully and sparsely observed examples (the largest errors are reduced by about 15%) but more substantial in the example with just boundary observations (the largest error are reduced by about an order of magnitude).

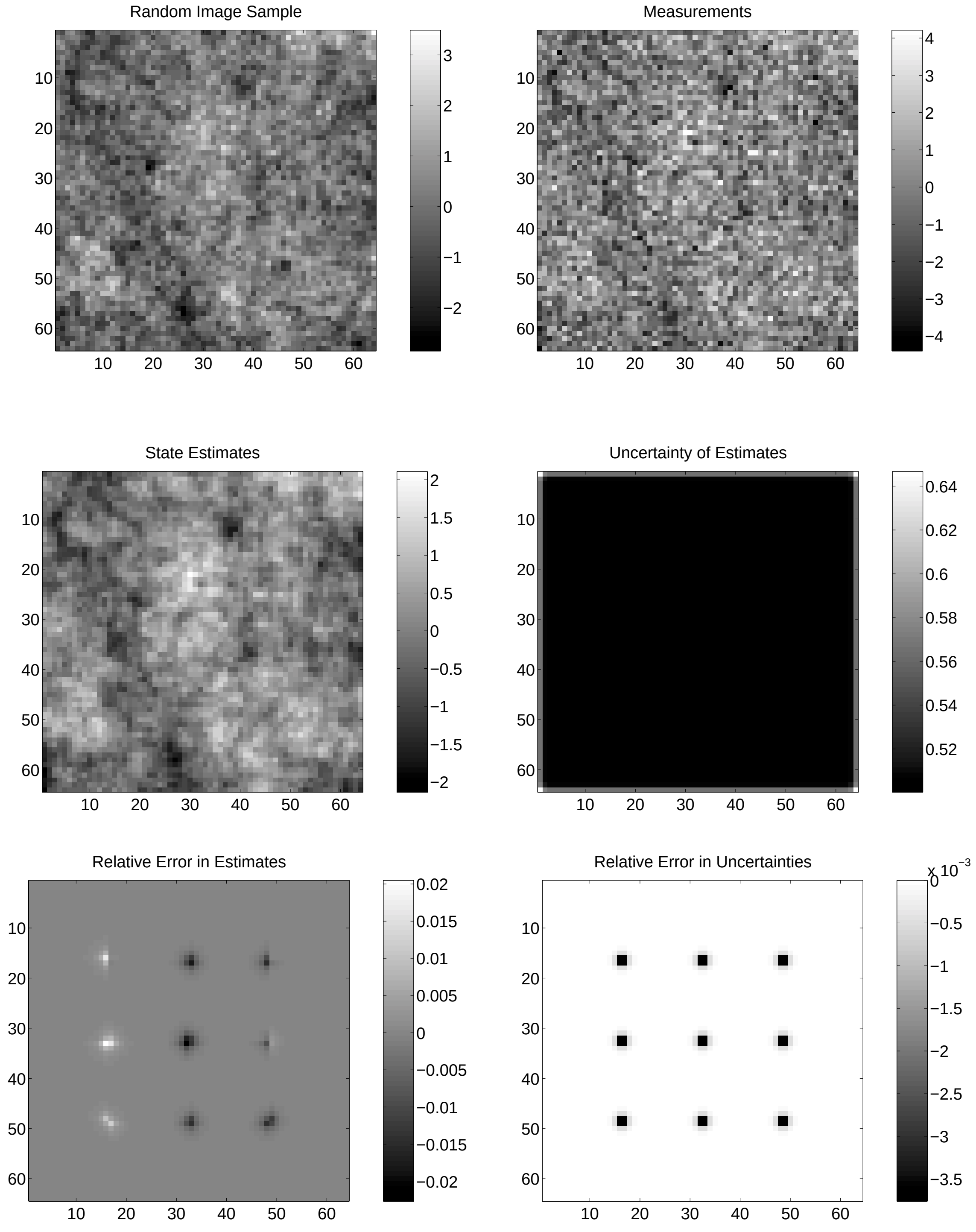


Figure 4-12: Fully-observed image restoration example. A sample of our GMRF image model is shown (top left). based on this sample, we simulated a set of noisy measurements of each pixel in the sample image (top right). Given these observations, we then estimate the image using our RCM inference technique. This yields the restored image (middle left) and associated pixel-by-pixel uncertainties (middle right). Finally, at bottom, we compare these RCM estimates of the pixel means (bottom left) and uncertainties (bottom right) to the optimal values.

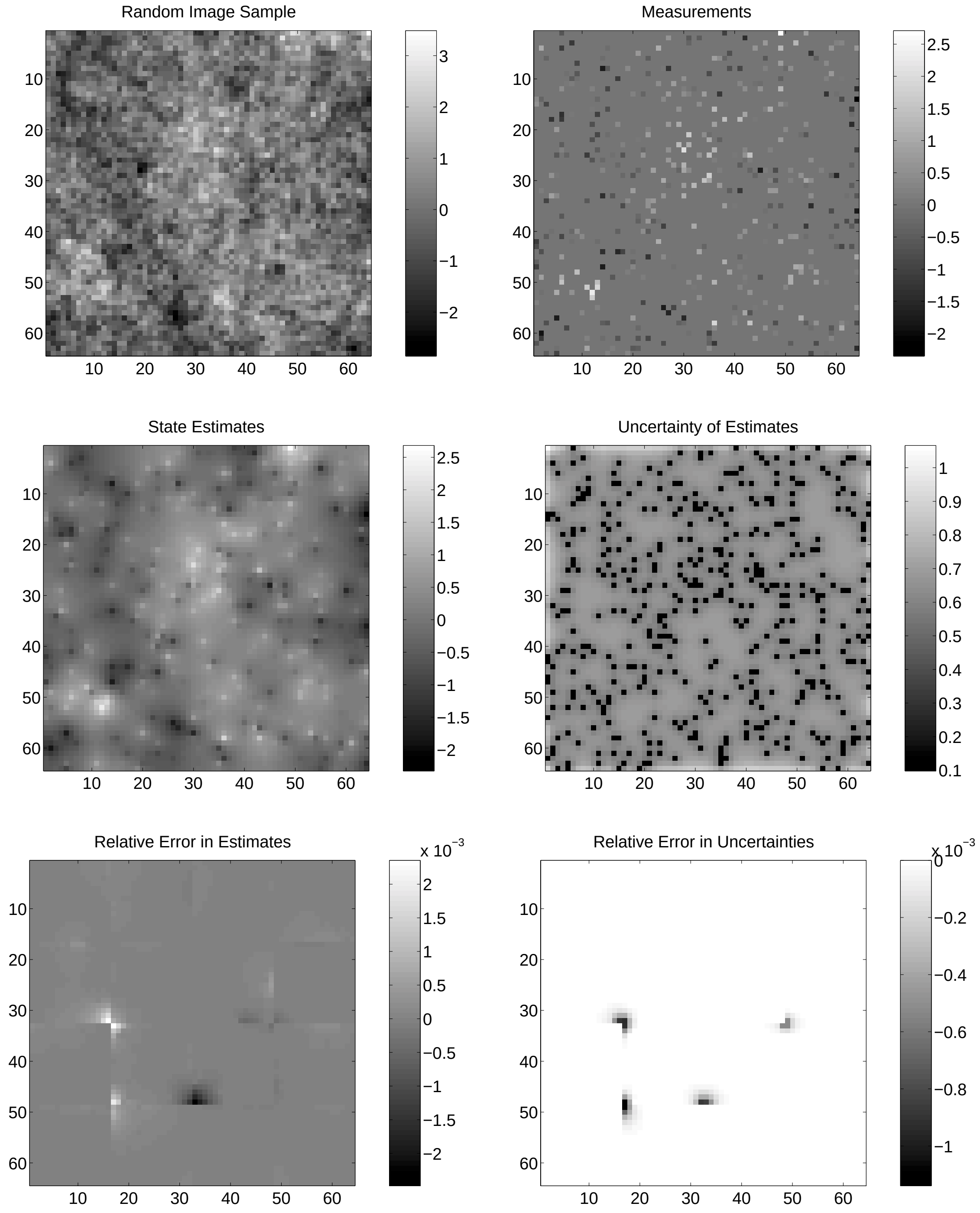


Figure 4-13: Sparsely-observed image restoration example. A sample of our GMRF image model is shown (top left). based on this sample, we simulated a set of noisy measurements of a randomly selected subset of pixels (top right). Given these observations, we then estimate the original image using our RCM inference technique. This yields the restored image (middle left) and associated pixel-by-pixel uncertainties (middle right). Finally, at bottom, we compare these RCM estimates of the pixel means (bottom left) and uncertainties (bottom right) to the optimal values.

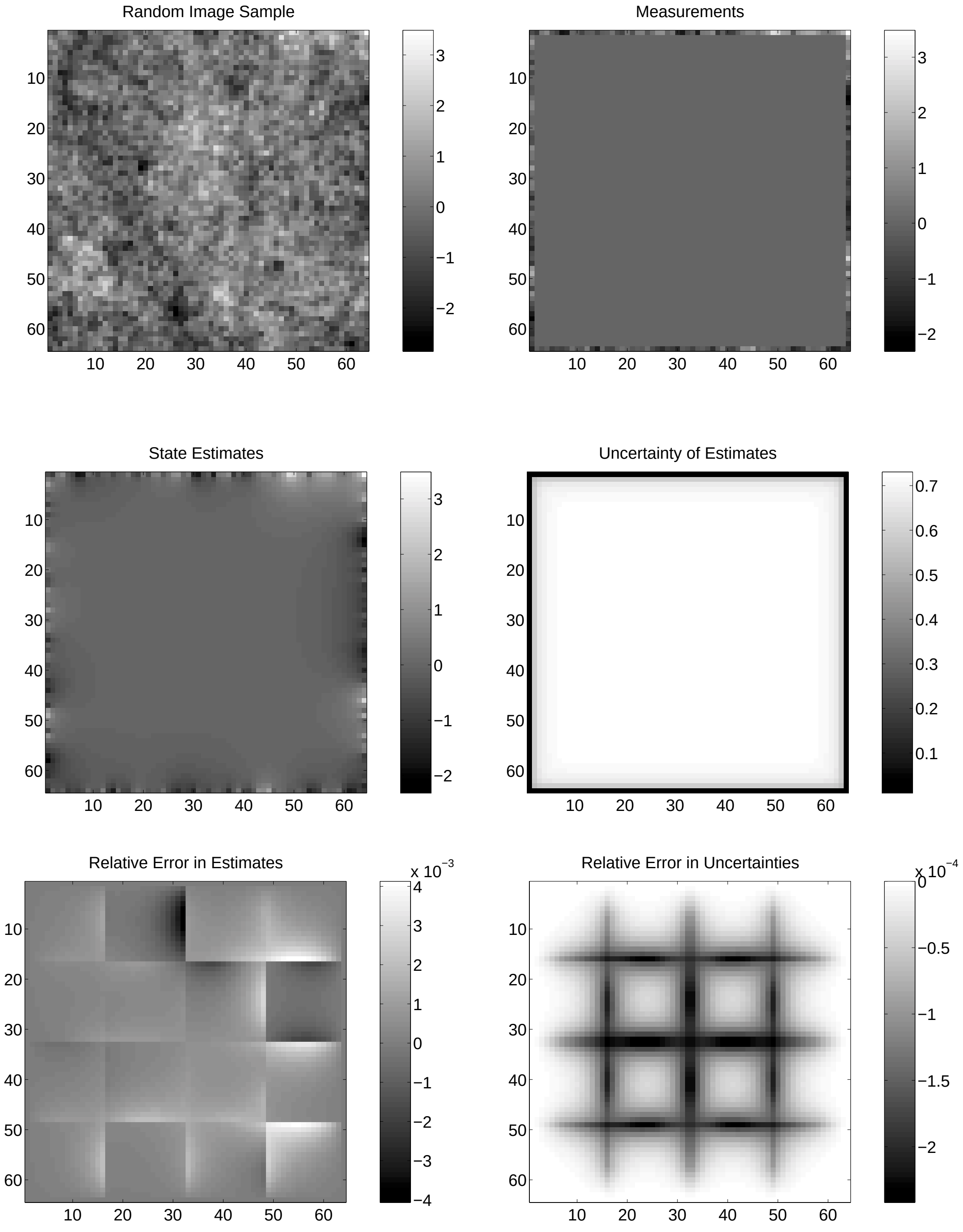


Figure 4-14: Boundary-observed image restoration example. A sample of our GMRF image model is shown (top left). Based on this sample, we simulated a set of noisy measurements of just those pixels going around the perimeter of the image (top right). Given these observations, we then estimate the original image using our RCM inference technique. This yields the restored image (middle left) and associated pixel-by-pixel uncertainties (middle right). Finally, at bottom, we compare these RCM estimates of the pixel means (bottom left) and uncertainties (bottom right) to the optimal values.

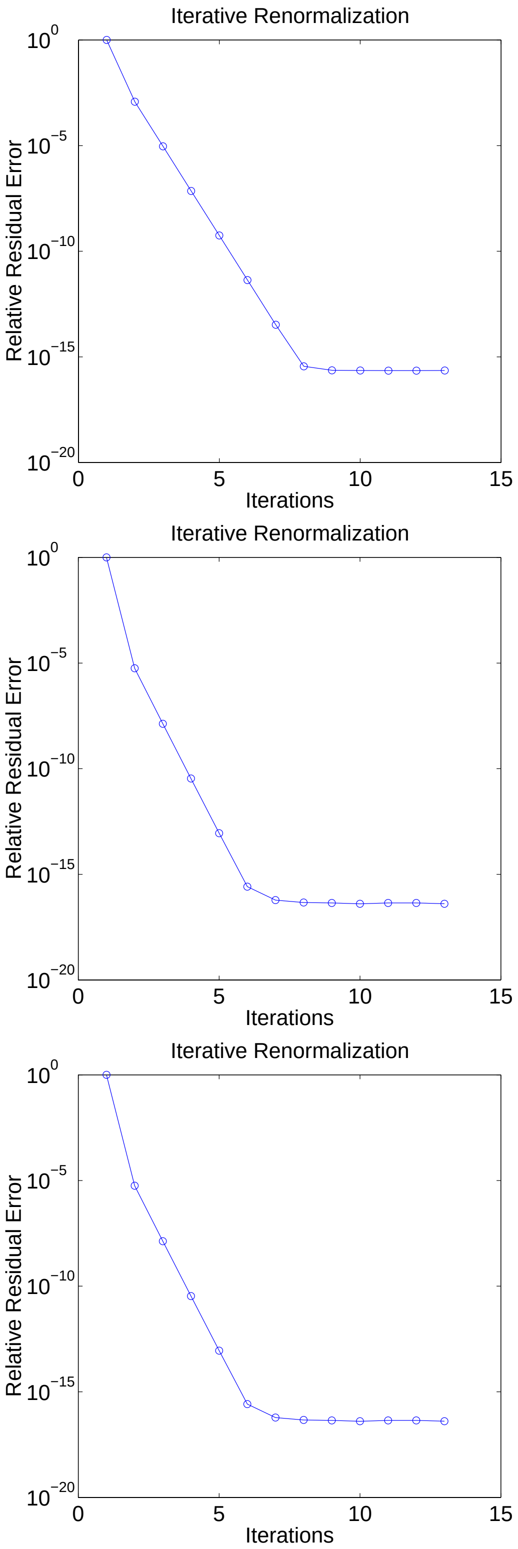


Figure 4-15: Convergence of iterative renormalization in the relative residual error for each of our three image restoration examples: fully-observed (top), sparsely-observed (middle) and boundary-observed (bottom).

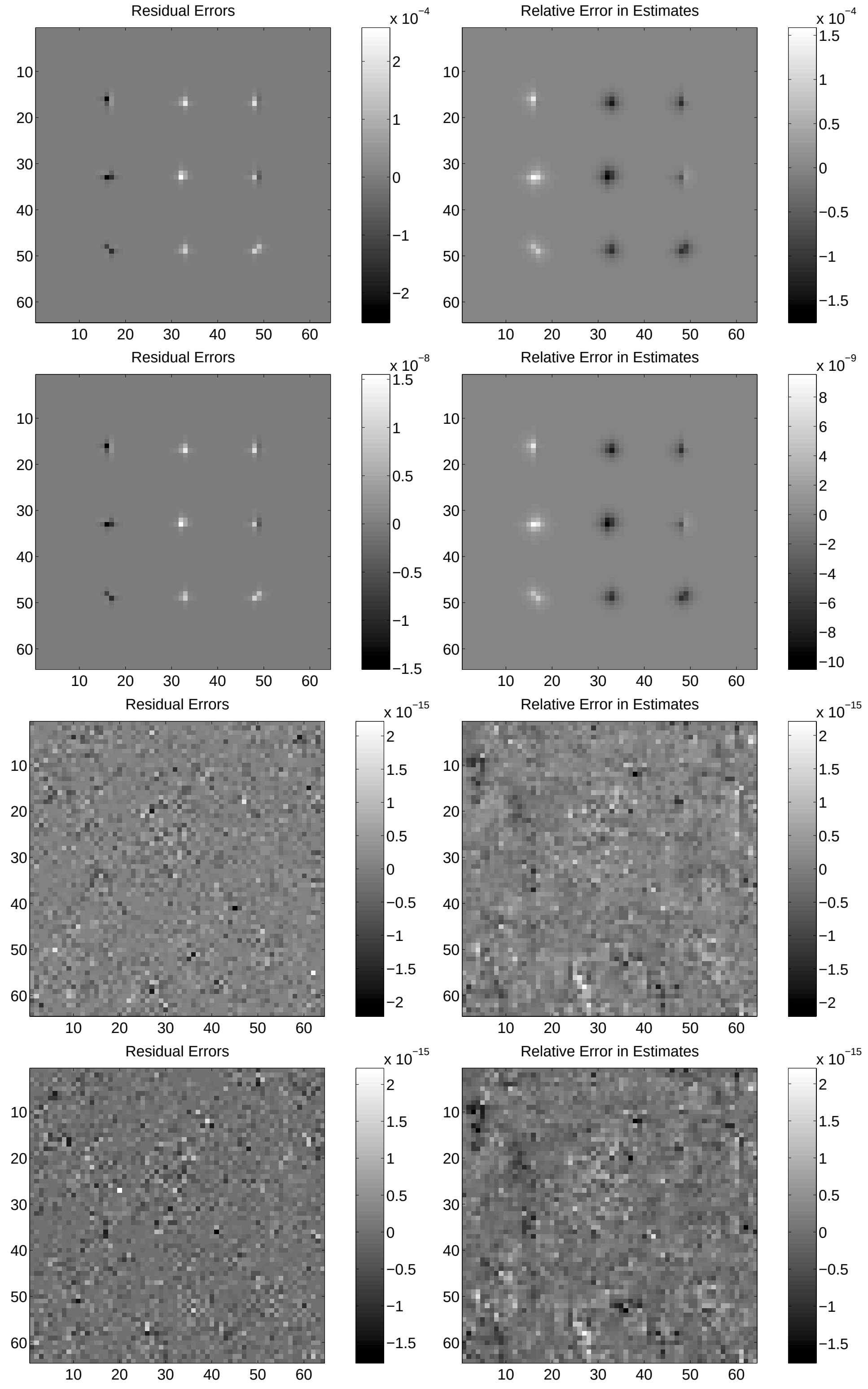


Figure 4-16: Iterative renormalization in fully-observed image restoration example. Errors arising in iterative renormalization approach with noisy measurements at every pixel. The residual errors (left column) and relative state estimation errors (right column) after 2 iterations (top row), 4 iterations (2nd row), 8 iterations (3rd row), and 12 iterations (bottom row).

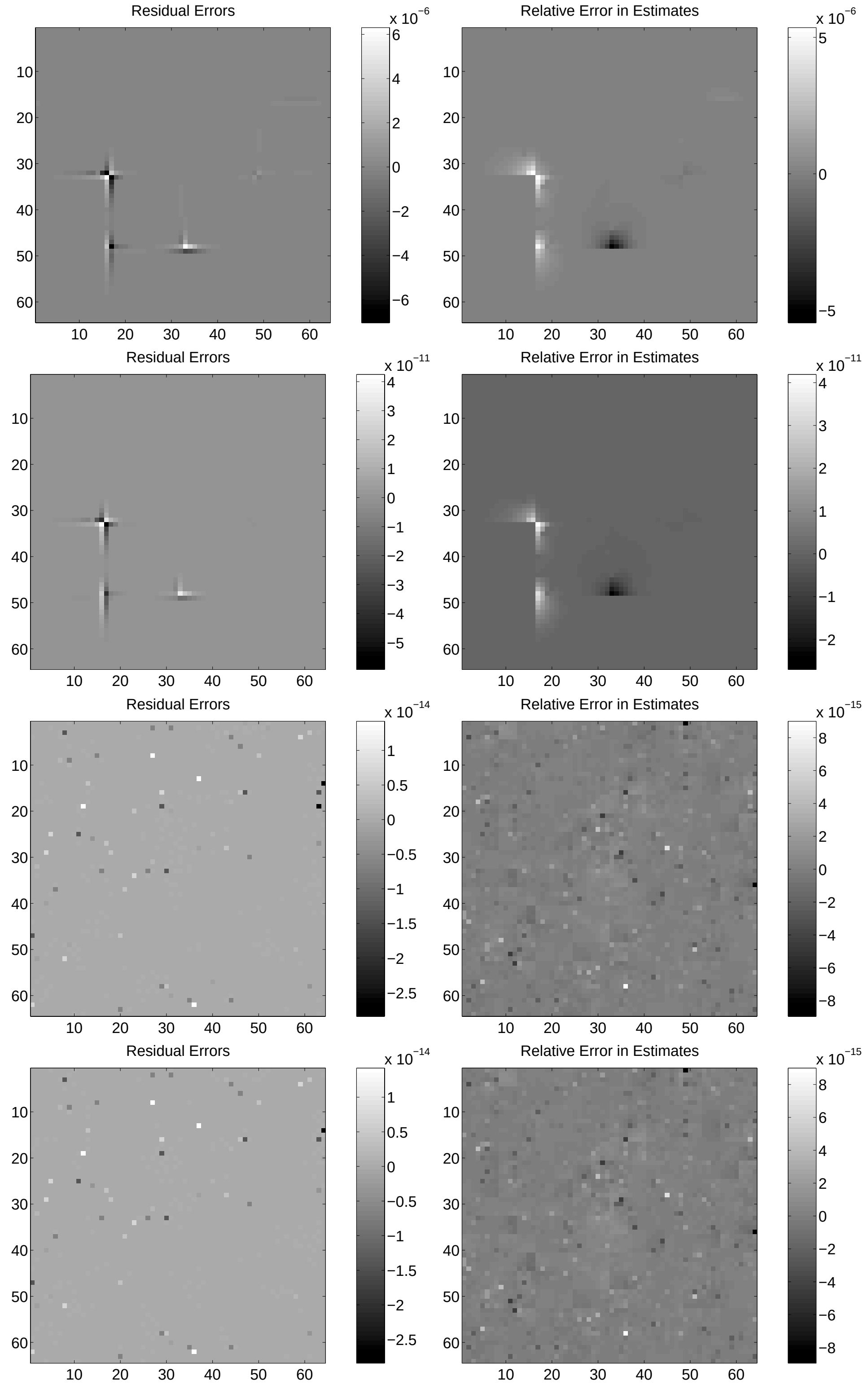
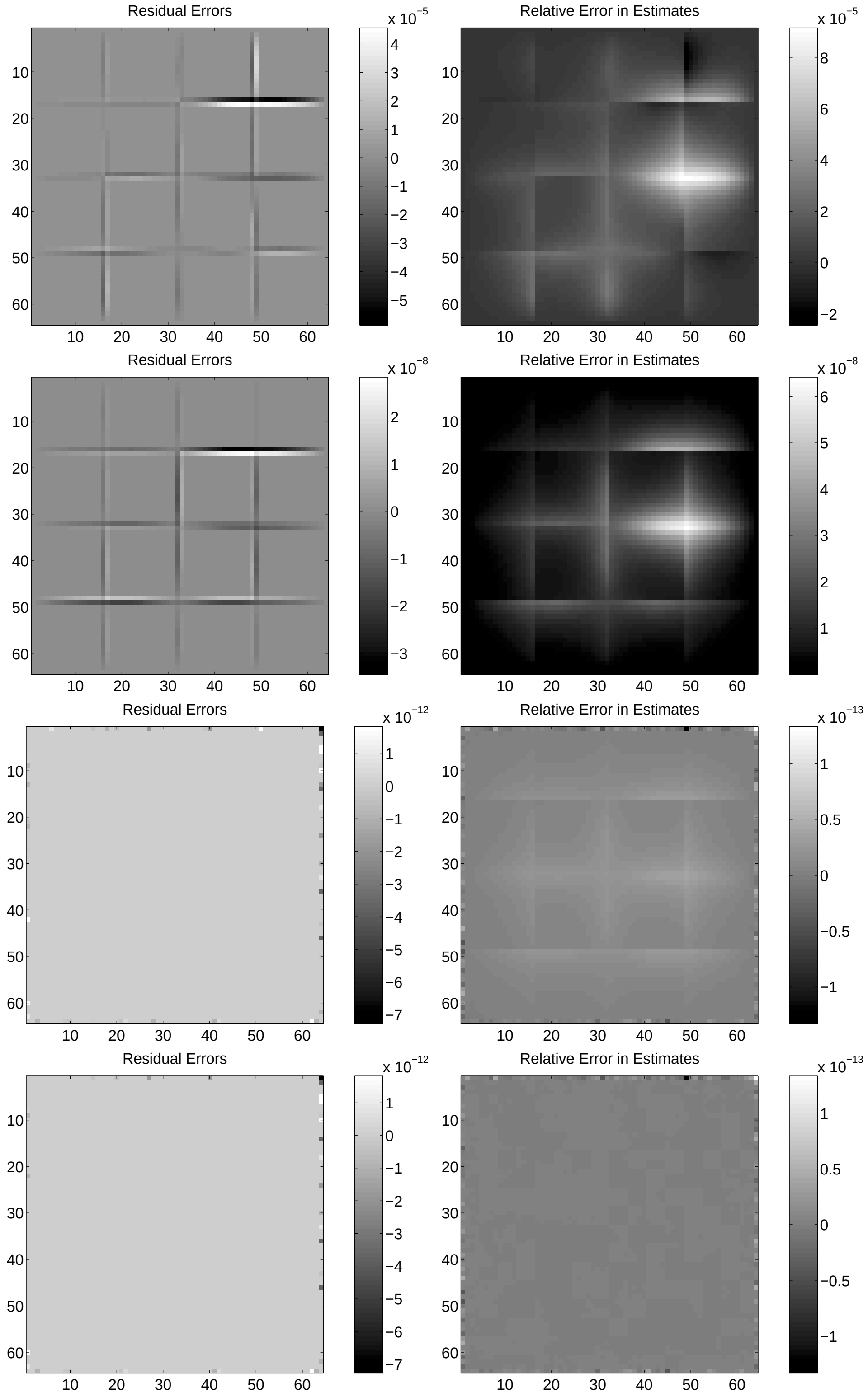


Figure 4-17: Iterative renormalization in sparsely-observed image restoration example. The residual errors (left column) and relative state estimation errors (right column) after 2 iterations (top row), 4 iterations (2nd row), 8 iterations (3rd row), and 12 iterations (bottom row).



Residual Errors

Relative Error in Estimates

Figure 4-18: Iterative renormalization in boundary-observed image restoration example. The residual errors (left column) and relative state estimation errors (right column) after 2 iterations (top row), 4 iterations (2nd row), 8 iterations (3rd row), and 12 iterations (bottom row).

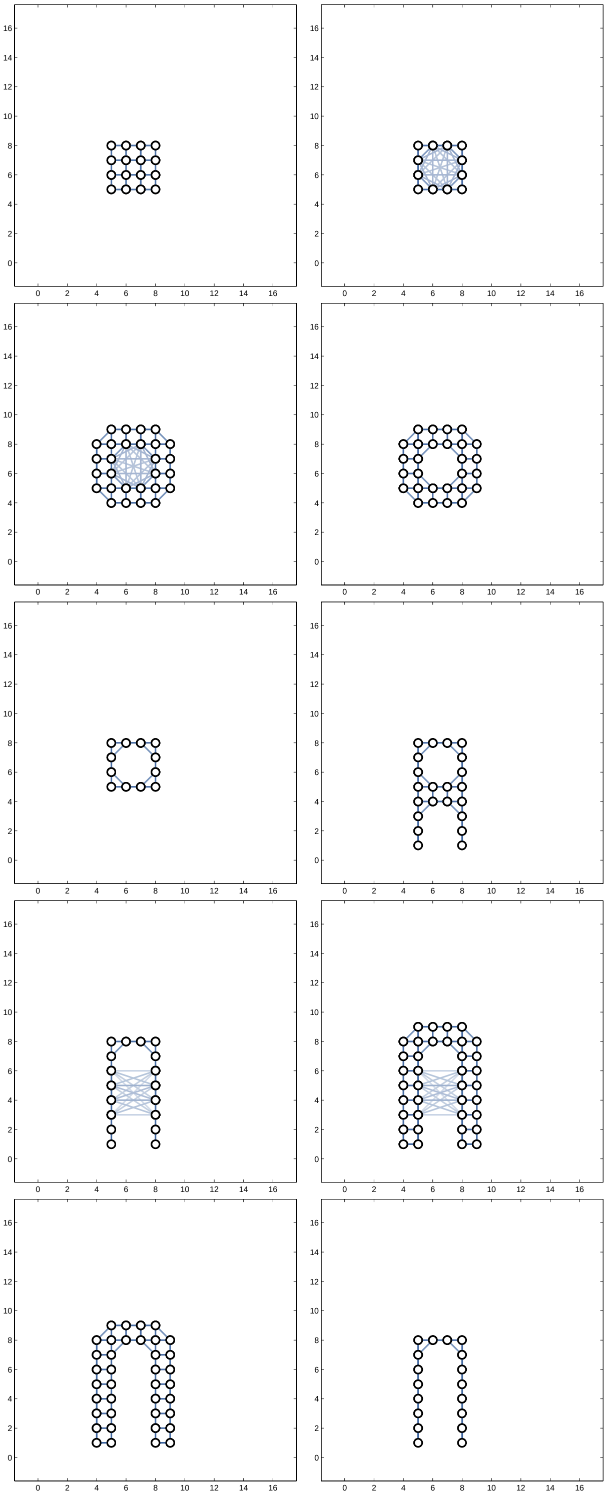


Figure 4-19: Selected snapshots of upward pass of iterative remodeling in 16 × 16 blocked GMRF (each node corresponds to 4×4 patch of the image). This shows the construction of two cavity models.

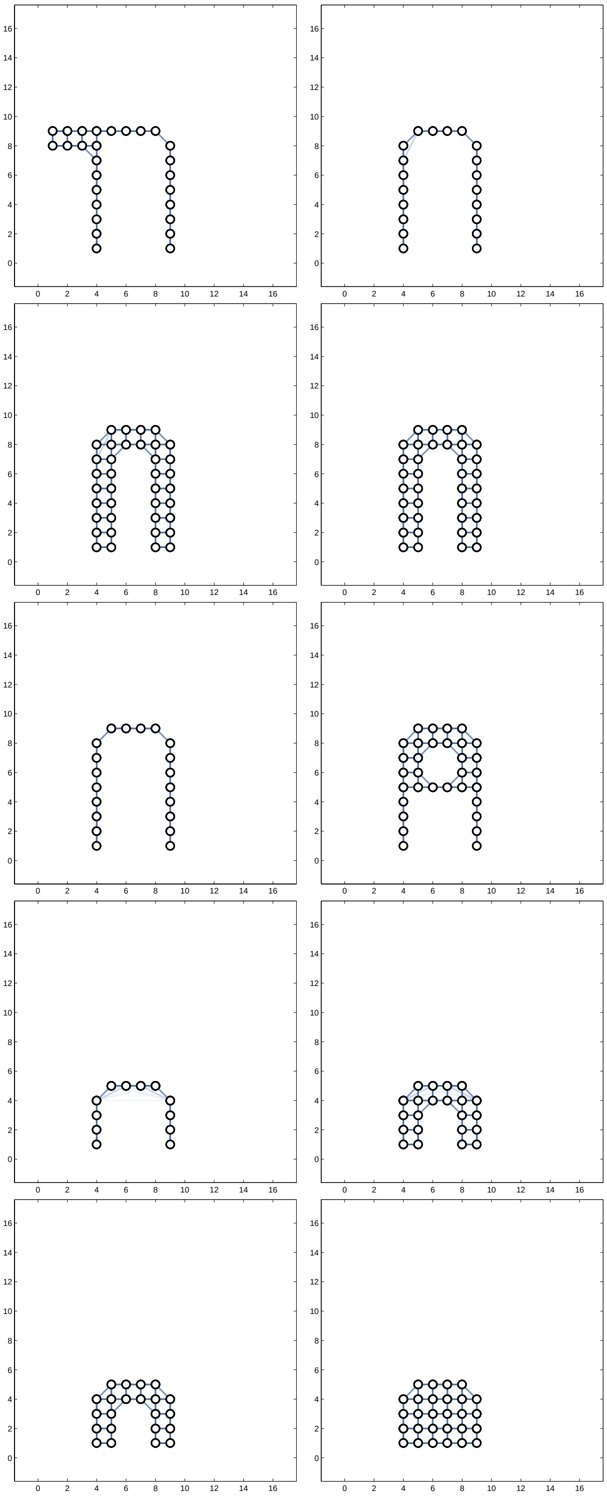


Figure 4-20: Selected snapshots of downward pass of iterative remodeling in 16 × 16 blocked GMRF (each node corresponds to 4 × 4 patch of the image). This shows construction of two blanket models and inference of a subfield from enclosing blanket model.

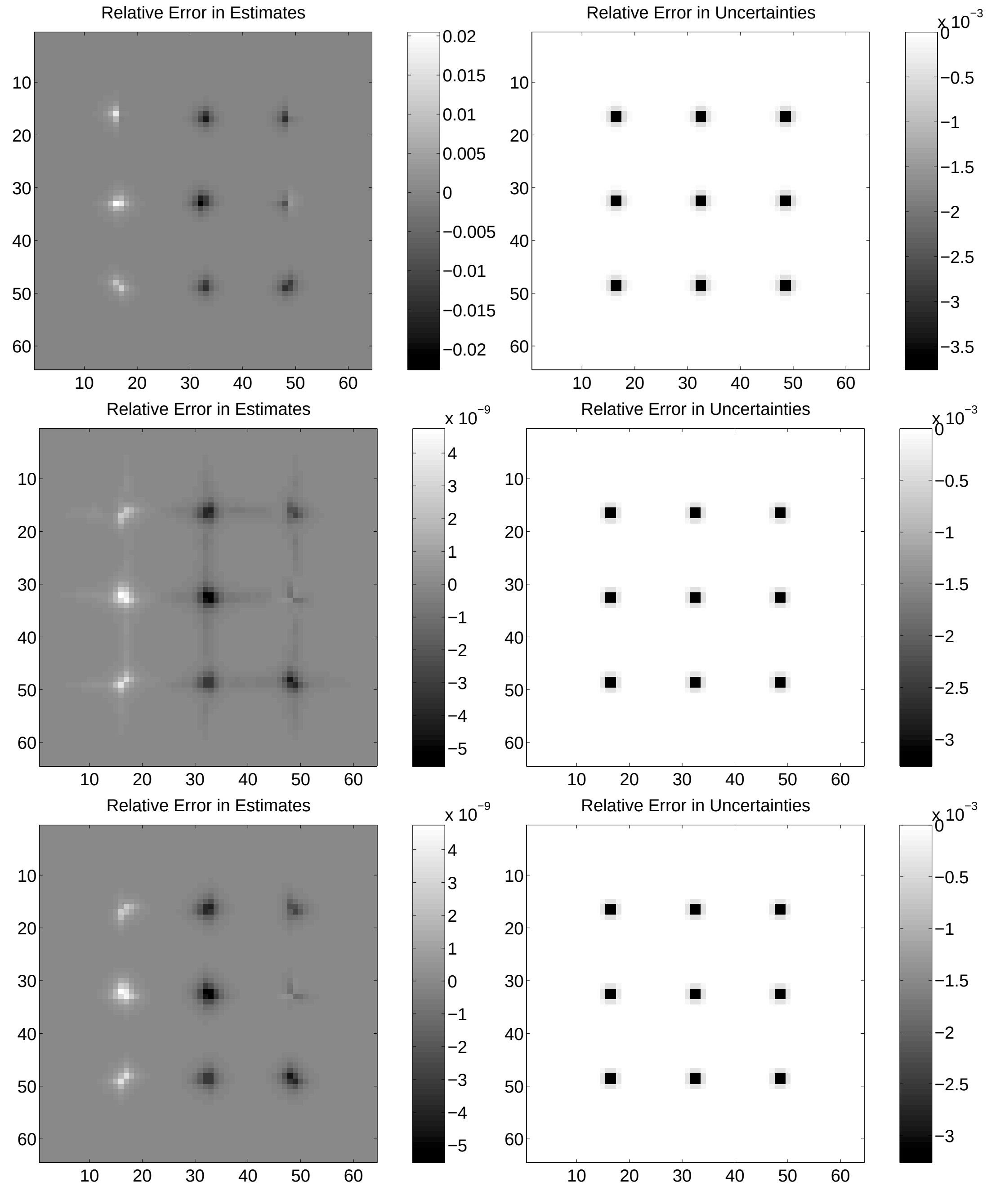


Figure 4-21: Iterative remodeling in the fully-observed image restoration example. Images of the relative mean-state estimation errors (left column) and relative uncertainty estimation errors (right column) arising in our iterative remodeling approach.

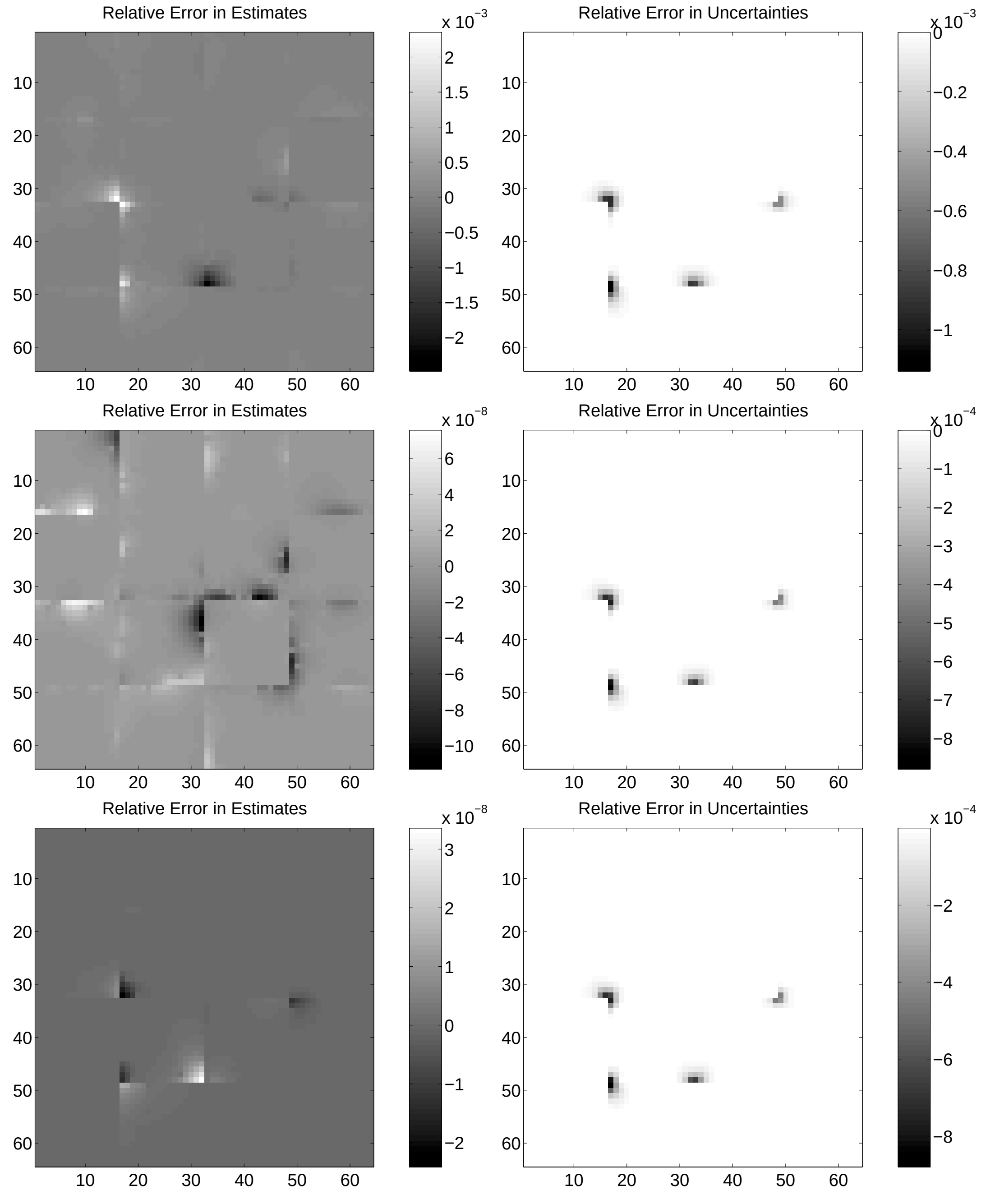


Figure 4-22: Iterative remodeling in sparsely-observed image restoration example. Images of the relative mean-state estimation errors (left column) and relative uncertainty estimation errors (right column) arising in our iterative remodeling approach.

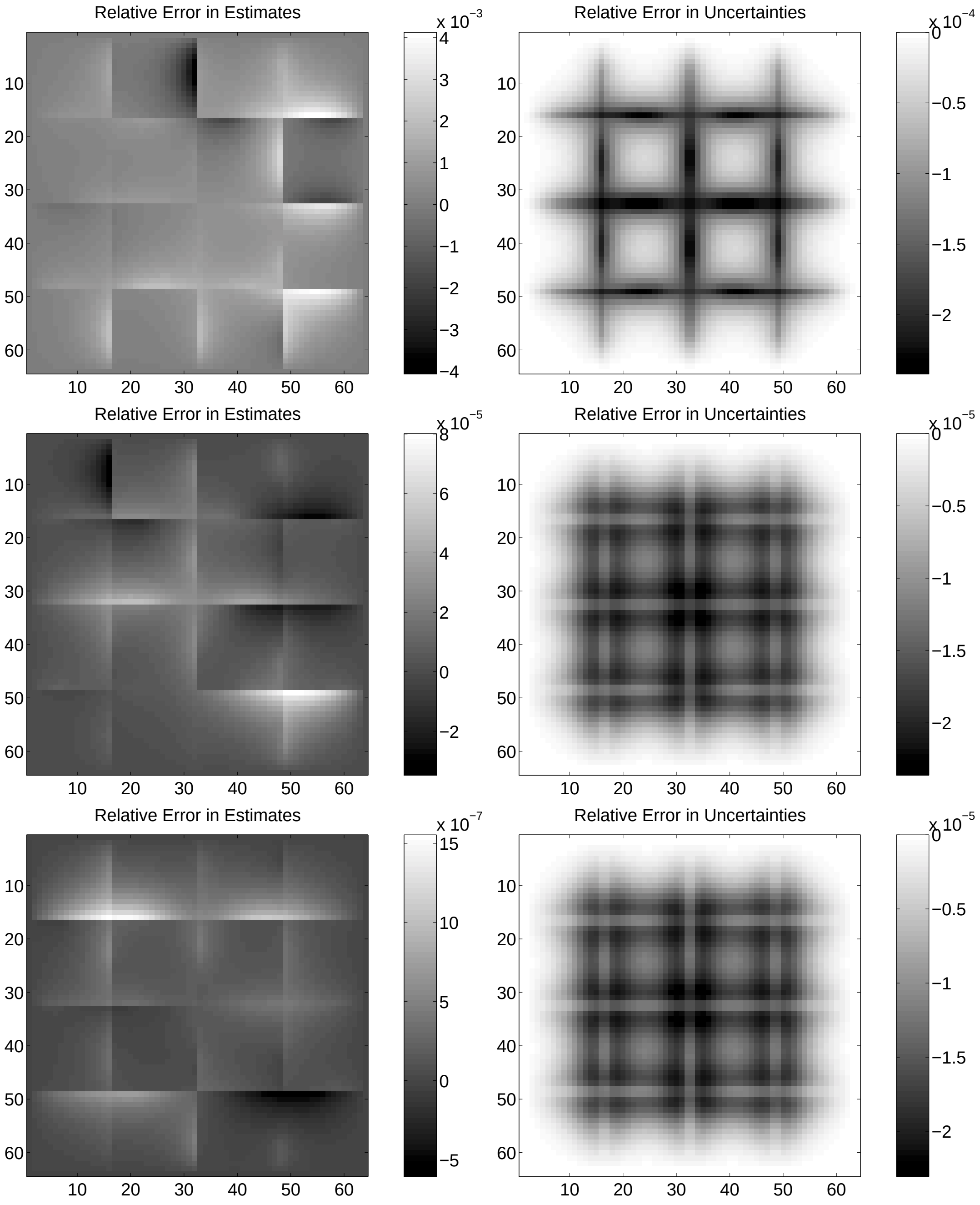


Figure 4-23: Iterative remodeling in the boundary-observed image restoration example (continued in Figure 4-24). Images of the relative mean-state estimation errors (left column) and relative uncertainty estimation errors (right column) arising in our iterative remodeling approach after 1 and 2 iterations (middle and bottom rows).

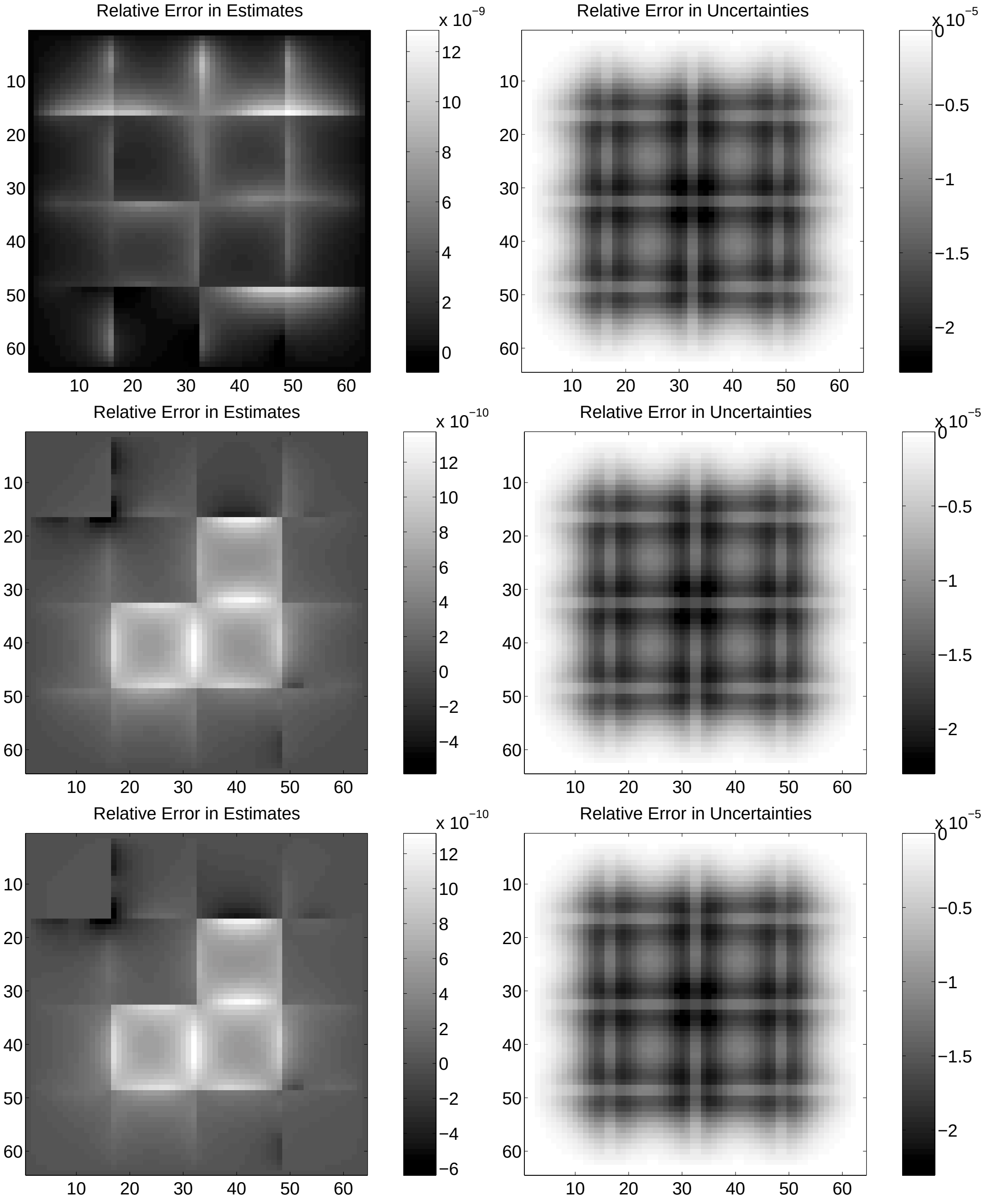


Figure 4-24: Iterative remodeling in the boundary-observed image restoration example (continued from Figure 4-23). Images of the relative mean-state estimation errors (left column) and relative uncertainty estimation errors (right column) arising in our iterative remodeling approach after 3, 4 and 5 iterations (top, middle and bottom rows).

Chapter 5

Conclusion

This thesis has presented a general framework, recursive cavity modeling, for tractable approximate computation of the marginal distributions of Markov random fields and has detailed implementation of this method for Gauss-Markov random fields in particular.

In the next section we give a summary of our approach and identify the original contributions of the thesis. In Section 5.2, we outline some promising directions for further research and development.

5.1 Contributions

First, the main ideas underlying the approach are summarized:

- *Nested dissection* gives a tree-structured decomposition of the graphical model. This entails recursively partitioning the sites of the field into smaller and smaller subfields called *dissection cells*. The nested structure of this dissection procedure is recorded as a *dissection tree* where the root of the tree corresponds to the entire field and the leaves correspond to the smallest (final) dissection cells.

- *Recursive inference* methods are adopted to provide a two-pass tree-structured (non-loopy) message-passing inference procedure structured according to the dissection tree:

  - An *upward pass* recursively builds *cavity models* for each dissection cell, a compact yet faithful graphical model for the surface of each cell sufficient (or nearly so) for inference outside that cell. Each cavity model is constructed from cavity models of subcells.

  - A *downward pass* recursively builds complementary *blanket models* for each dissection cell, a compact graphical model for the blanket of each cell sufficient (or nearly so) for inference inside that cell. Each blanket model is constructed from an adjacent cavity model and an enclosing blanket model.

Blanket models for the smallest dissection cells (at the leaves of the tree) then support approximation of the *marginal models* for each subfield.

- *Model thinning* of cavity and blanket models is introduced to provide tractable yet near-optimal inference. Some key features of our model thinning technique are emphasized:

  - *M-Projection.* We thin cavity and blanket models by m-projection to lower-order exponential families (minimizing KL-divergence) thereby pruning selected weak interactions from the model. This also gives the *maximum-entropy model* subject to a correspondingly reduced set of moment constraints so that RCM may be viewed as a "forgetful" inference approach where only a subset of the moment characteristics are preserved through the computation.
  - *Moment Matching.* The necessary m-projections are evaluated by *moment matching* within the lower-order exponential family using our *loopy iterative scaling* (LIS) moment-matching technique. This is loosely based on iterative scaling (e-projection) techniques such as iterative proportional fitting (IPF). Our approach, however, is more aggressive in that we fuse IPF updates for all maximal cliques (in the thinned cavity/blanket model), in a manner which attempts to correct for "overcounted" intersections, so as to obtain more rapid convergence to the desired moments.
  - *Model Selection.* An AIC-like information criterion, balancing model compactness against model fidelity, is adopted to guide our selection of thinned cavity and blanket models in a principled manner. We use an inductive approach to model selection which allows the effect of pruning weaker interactions to be determined before deciding what other interactions might also be pruned. Model thinning continues until we can no longer identify m-projections to lower-order families while keeping the information loss per removed parameter less than the precision parameter δ of our information criterion.

  Hence, our inference approach is really as much about modeling as inference and relies heavily on ideas borrowed from information geometric modeling techniques for the selection of compact yet faithful cavity and blanket models.

- *Iterative extensions* of this approach allow refinement of the approximations introduced by RCM giving improved approximation of marginal distributions. Having shown that the model-thinning m-projections performed in RCM are performed relative to certain "ground-state" boundary conditions (implicit in our choice of potential specifications), we have proposed two iterative extensions of RCM performing modified m-projections so as to refine these approximations.

  - *Renormalization.* In the first approach, we use two-pass RCM to generate a state estimate and then "renormalize" the potential specification, resetting the ground state to this estimate. Then, subsequent RCM approximations are conditioned on the estimate of the state on the boundary of each subfield being approximated. In GMRFs, this reduces to a Richardson iteration with RCM playing the role of a preconditioner. The means of our approximate marginal models then converge to the true means.

  - *Remodeling.* In this approach, rather than conditioning on some state estimate while thinning, we instead incorporate a previously constructed thinned model of the boundary (one of our cavity or blanket models). Then, the m-projections performed by RCM correspond (at least approximately) to global (unconditional) m-projections where we thin a cavity model by matching moments computed under the joint cavity-blanket model. We find that, in GMRFs, this can improve the covariance estimates of marginal distributions. The means are also improved but, while the improvement is substantial, these do not converge to machine precision as in the renormalization approach.

The main innovative aspect of this approach is the manner in which RCM marries recursive inference with adaptive model thinning techniques. While some related methods have been developed along these lines, RCM offers a much more general and ambitious approach in this regard. We point out the following distinctive features of RCM distinguishing our approach from these other methods:

- The model thinning technique used in RCM offers an alternative perspective for controlling the computational complexity of recursive inference in comparison to those multiscale modeling techniques based on (approximate) state-reduced Markov trees (Luettgen [91], Irving [72], Frakt [55]). We find, in the simulations of Chapter 4, that our model thinning technique enables a scalable approach for inference in 2-D MRFs with negligible "blocky artifacts" such as have plagued many multiscale modeling methods.

- In contrast to other linear domain-decomposition methods (in GMRFs), such as the methods of Taylor [126] and Daniel [36], RCM selects *optimal approximations* of the thinned boundary models arising in our approach, employing the machinery of information geometry to minimize the KL-divergence introduced by our approximations. That is, those parameters not pruned by our method are optimized in an effort to minimize any ill-effects of thinning. Also, we adaptively select which parameters to prune so as to keep this KL-divergence small.

- In a sense, RCM generalizes various projection filtering approaches for Markov chains (Kullhavý [86], Brigo et al [27], Heskes and Zoeter [70]) for more general (loopy) Markov random fields. The main idea RCM introduces in this regard, is that of performing model thinning by tractable m-projections based on a collection of local interaction potentials. By considering canonical potential specifications, we are able to interpret this as a conditional m-projection relative to "ground-state" boundary conditions for the subfield being thinned. Thus, these local m-projections are well-posed and have a well-understood interpretation.

- In comparison to some related Markov-blanket filtering methods for MRFs (Boyen and Koller [22], Murphy and Weiss [99], Minka [95]), RCM is more general in several regards:

   1. RCM considers representation of the MRF as a *Markov tree* rather than as a *Markov chain*. Where other approaches propagate a "frontier model" back and forth across the MRF, our hierarchical approach instead recursively builds cavity and blanket models. Hence, RCM introduces a new "region merging" step into the Markov-blanket filtering approach (borrowing this idea from multiscale modeling methods) where thinned models of subfields are combined in a tractable manner to obtain a thinned model for larger subfields. This replaces a "frontier propagation" step in the related Markov-chain approach.
   2. We do not impose a priori structure constraints on our cavity and blanket models. For instance, these other approaches assume either: completely disconnected frontier models (Murphy and Weiss), partially disconnected frontier models consisting of a set of smaller, fully-connected components (Boyen and Koller), or singly-connected frontier models (Minka). None of these constraints are placed on our analogous cavity and blanket models. Rather, we employ *adaptive* model-thinning techniques which tend to produce low tree-width models (which we may convert to Markov tree representations while still keeping state dimensions small) thereby supporting tractable computation.

- In contrast to Minka's general expectation propagation (EP) framework [96, 95], which also generalizes the projection filtering approach, RCM offers a more methodical, structured approach in two regards:

   1. All message passing in RCM parallels an exact (non-loopy) recursive inference procedure. This is in contrast to Minka's general EP framework which allows arbitrary "message passing protocols" including "loopy" propagation schemes. Our approach closely follows an exact inference procedure so as to assure that, by making the precision δ of our inference sufficiently small, we can improve the accuracy of our method so as to approach that of exact inference.
   2. Minka also suggests "structured" versions of EP, where additional interactions are included in the model (in addition to those present in the original MRF) so as to allow his method to come closer to exact inference. Yet, as Minka points out, a systematic approach for selecting these additional interactions and subsequent message-passing protocol is lacking. In this regard, RCM may be seen as a more structured form of EP. Our hybrid variable-elimination/model-thinning approach adaptively determines which interactions are added to the graphical model in the course

of the inference so as to (hopefully) remain near the corresponding exact inference.

- Finally, we remark that RCM may be understood as essentially a thinned junction-tree approach. In comparison to the junction-tree thinning procedure of Kjærulff [82, 83], RCM offers a more tractable approach in that we never consider the fully-triangulated representation in the first place. Rather, we infer what additional interactions should be added to the model by analysis in our thinned models of subfields. On the other hand, our method of thinning cavity and blanket models is rather similar to Kjærullf's global method for thinning a junction tree. In RCM, however, we do not restrict ourselves to triangulated Markov-blanket models and hence may prune any weak interaction of the model regardless of whether or not this corresponds to a "removable" edge in Kjærullf's approach.

Also, our approach to model thinning is itself innovative in several regards. The work-horse of our model thinning approach is our LIS moment-matching subroutine used to calculate all m-projections required in RCM. In singly-connected Markov chains or trees, this actually reduces to an exact (non-iterative) m-projection technique suggested by Minka for m-projection to Markov trees [96]. But in loopy-graphs, LIS gives an iterative procedure which appears to consistently outperform the iterative proportional fitting technique and to converge linearly so that KL-divergences from the desired marginal distributions vanish exponentially quickly. Moreover, this technique seems especially appropriate for matching moments in "long loops" such as often occur in RCM[1]. In addition to this moment matching technique, our inductive approach to model selection, based on an AIC-like information criterion with precision parameter δ, seems a very natural approach to model thinning which is motivated from the perspective of information geometry (Amari [4]). While other edge-pruning approaches to model thinning have been developed (Brand [23], Smith and Whittaker [123]), apparently none follow this information geometric perspective. However, there are some related approaches for *building* graphical model by adding features to a graphical model be a sequence of e-projections (Della Pietra et al [106], Bach and Jordan [6]). Essentially, our (m-projection) model thinning method is dual to these (e-projection) model building methods.

In short, the author believes that the RCM approach – by unifying, formalizing and building upon ideas drawn from these various earlier efforts – represents a significant step towards the elusive goal of obtaining reliable, scalable inference for far more general families of MRFs than have previously been considered tractable. Embedding adaptive modeling algorithms within a recursive inference framework appears to give a powerful and flexible approach to tractable yet near-optimal inference. Also, the author hopes that the "cavity modeling" picture developed here, together with the emphasis on information geometry, provides a helpful, intuitive way of thinking about recursive inference and approximation techniques more generally. Yet, much work

[1] See comments on Test Case 1 and 2 in Section 3.4

remains to realize the full potential of this approach and to understand and more precisely characterize the reliability of the method.

5.2 Recommendations

In this section, recommendations for further research and development are given. The aim of these recommendations is to analyze, refine, and extend the RCM framework introduced in this thesis.

Model Thinning. As is probably apparent at this point, the ideas underlying the RCM approach to inference are really as much about modeling as inference. The tractability of this inference approach relies heavily on modeling ideas for the selection of compact (and hence tractable) cavity and blanket models to provide a tractable basis for inference of very large MRFs. In Chapter 3, we outlined one possible approach for implementing the model selection subroutine of the RCM inference procedure. Yet, there are many possible variations of our approach which may be adopted without substantially modifying the basic structure and interpretation of RCM. We briefly consider some promising alternatives.

- *Moment-Relaxation Approach to M-projection.* Currently, we solve for the m-projection to an exponential family by moment matching within that family. This requires that we generate an initial guess for the m-projection to seed the moment-matching subroutine. These model-thinning m-projections also have the interpretation of releasing some moment constraints (corresponding to pruned interactions) and maximizing entropy. This suggests that perhaps other m-projection techniques might be recommended where we instead gradually relax moment constraints (maximizing entropy) so as to trace the $\mathcal{I}$-orthogonal m-geodesic connecting the given model to the desired m-projection. This presumably would take the form of a double-loop algorithm where the outer loop gradually relaxes moments (maximizing entropy) while the inner-loop maintains active moment constraints (so as to stay near the m-geodesic). This may prove to be a more economical approach, requiring less overall computation, since we avoid large steps away from the m-geodesic which might substantially perturb those moments we are trying to hold fixed during thinning.

- *Feature-Induction Approach to Model Selection.* Alternatively, it may also be possible to abandon the m-projection approach all together, and instead use the dual e-projection approach for building cavity/blanket models such as in some related modeling techniques (Della Pietra et al [106], Bach and Jordan [6]). The advantage of this approach would be to avoid the variable elimination steps in RCM and also to consider only very thin cavity/blanket models since we "build up" to the final thinned model rather then "thin down" from a more complex model. The challenge here, of course, is how to select what interactions to add to the cavity/blanket model.

- *Latent-Variable Markov-Blanket Models.* Finally, we note that the philosophy of RCM described thus far only considers cavity/blanket models based on just those sites of the graphical model which are in the surface of the subfield being approximated. We suspect that the accuracy of the cavity/blanket models might be substantially improved (while staying within low-order, thin families of graphical models) by also allowing some latent variables to be included inside of each cavity model (or outside of each blanket model). In the context of RCM, these latent variables might play a useful role for (approximately) capturing in an aggregate manner much of the statistical influence of the eliminated subfield which would otherwise be lost in the model thinning step. The challenge here is how to select and refine such latent-variable models.

Generalization of RCM. We now give recommendation for further research and development focusing on extending the basic RCM framework. While we have only detailed implementation of RCM for GMRFs, the main ideas should apply for much more general families of MRFs. We briefly indicate some especially interesting possibilities.

- *Nonlinear Interactions.* Our Gaussian implementation of RCM could perhaps be extended to treat more general MRFs, also having continuous-valued states, but where the interaction potentials are specified by nonlinear functions (rather than just the linear and quadratic interactions present in GMRFs). We could perform approximate inference in such MRFs by adaptively "linearizing" the MRF, fitting linear and quadratic statistics to the actual nonlinear statistics in the vicinity of some state estimate. Essentially, this would correspond to generalization of the extended Kalman filter and related smoothing methods (developed for Markov chains) to more general (loopy, undirected, non-Gaussian) MRFs.

- *Finite-State MRFs.* RCM could also be readily developed for finite-state MRFs, where each site of the random field has a (small) finite number of states. These may also be described as exponential families but with a discrete state-space and (in general) require higher-order interactions (involving more than two sites of the field). Inference in finite-state MRFs is more challenging than in GMRFs because variable elimination introduces higher-order interactions (even if the original model has only pairwise interactions). The model-thinning techniques employed by RCM seem well-suited for controlling the computational complexity of inference for these models as well. For instance, we could m-project intractable cavity/blanket models to families of more tractable models represented by sparse, lower-order interactions (letting our information criterion decide which interactions to keep).

- *Compound Gaussian MRFs.* More generally still, we could consider hybrid-state MRFs, having sites with both discrete and continuous state components. These may also be described as exponential families. If the interaction potentials between continuous states are restricted to linear and quadratic statistics, such models are then *conditionally Gaussian* so that, conditioning on any discrete-state configuration of the system, the continuous-states are Gaussian distributed (e.g. Lauritzen [88]). In addition to possibly having higher-order interactions, inference in such models is further complicated by the fact that variable elimination produces an exponential number of Gaussian modes in the blanket of an eliminated subfield (one mode for each discrete joint-state configuration of the combined blanket and subfield). Here again, RCM's model thinning approach might play a useful role for thinning such models so as to control the computational complexity of inference while attempting to remain nearly optimal.

Stability and Reliability of RCM. The author has attempted to motivate the RCM approach to inference from the perspective of information geometry by emphasizing the *local optimality* (m-projection interpretation) of each of our modeling thinning steps and also by indicating the *maximum-entropy interpretation* of these model-thinning steps (essentially, RCM as a forgetful approach to inference regularized by the maximum-entropy principle). However, we would like to make more concrete claims as to the reliability of this apparently greedy inference procedure. Essentially, this corresponds to analyzing the stability of the projection filtering approach in the context of Markov trees. That is, we would like to show that keeping the incurred KL-divergence small in each of our approximation steps insures that the *cumulative* KL-divergence in each of our cavity/blanket models remains small relative to the corresponding exact models (constructed without any thinning). The work of Boyen and Koller [22] provides some promising results indicating stability of projection filtering in causal Markov chain models. But there seems to be much room for further analysis in the context of noncausal, tree-structured inference procedures. Ultimately, we would like to be able to provide estimates of and/or upper-bounds on the KL-divergence in each of the marginal models produced by RCM.

RCM for Model Identification. The utility of RCM for practical applications might be greatly enhanced by the development of corresponding model identification techniques. For instance, in image processing applications we would like to be able to fit 2-D MRF models to data characteristic of the image we wish to model. In the case of exponential family graphical models, maximum-likelihood parameter estimation from a collection of independent, fully-observed samples of the process reduces to moment matching. That is, we may employ iterative scaling techniques, such as IPF or our LIS approach, so as to iteratively adjust the parameters of the family until the characteristic moments of the model match the corresponding empirical moments of the data. Unfortunately, these moment matching techniques require an inference subroutine for computation of the moments of the model being adjusted. For many MRFs, inference is intractable so that optimal maximum-likelihood model identification is likewise intractable. This suggests that we instead consider suboptimal model identification where we use RCM in place of an exact moment calculation. By combining RCM with iterative scaling techniques (for instance our LIS variant), we should then be able to estimate the maximum-likelihood model in moment coordinates at a level accuracy dictated by the precision of the RCM moments calculation. Many more sophisticated modeling techniques, such as the expectation-maximization (EM) algorithm for maximum-likelihood parameter estimation in partially-observed MRFs (Dempster et al [42]), also require inference as a subroutine. Hence, the range of applicability of these methods could likewise be extended with the aid of RCM. This very simple idea indicates the enabling role RCM might play in extending the range of applications to which model-based methods might be applied. For instance, we would like to consider even more complex models appropriate for more general 2-D and 3-D random fields (not limited to nearest-neighbor MRFs) by introducing higher-order interactions and/or latent variables to model more complex, global interactions of the field. Potentially, this could lead to generalization of the multiscale modeling technique, previously restricted to quad-tree models, to also incorporate multigrid/multipole methods (Briggs [26], Rokhlin [116], Greengard and Rokhlin [65], Fieguth [51]).

As these ideas show, the realm of potential applications for the general RCM framework is much broader then has been explored by this thesis. The basic idea of developing a hierarchical representation of the probability distribution of a MRF, where interactions between subfields are approximated by thinned graphical models for the surfaces of subfields, appears to supply a powerful and flexible framework for near-optimal computation in many MRFs which previously would have been dismissed as intractable.

Bibliography

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Parzen et al. [102].

[2] H. Akaike. A new look at the statistical model identification. In Parzen et al. [102].

[3] S. Amari. Differential geometry of curved exponential families – curvature and information loss. *Annals of Statistics*, 10(2):357–385, June 1982.

[4] S. Amari. Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory*, 47(5):1701–1711, July 2001.

[5] S. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *AMS Translations of Mathematical Monographs.* Oxford University Press, 1993.

[6] F.R. Bach and M.I. Jordan. Thin junction trees. In *Advances of Neural and Information Processing Systems*, 2001.

[7] O. Barndorff-Nielsen. *Information and Exponential Families.* Wiley series in probability and mathematical statistics. John Wiley, 1978.

[8] M. Basseville, A. Benveniste, K. Chou, S. Golden, R. Nikoukhah, and A. Willsky. Modeling and estimation of multiresolution stochastic processes. *IEEE Transactions of Information Theory*, 38:766–784, 1992.

[9] M. Basseville, A. Benveniste, and A.S. Willsky. Multiscale autoregressive processes, parts 1 and 2. *IEEE Transactions on Signal Processing*, 40:1915–1954, 1992.

[10] H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*, 3(4):615–647, 2001.

[11] H.H. Bauschke and J.M. Bowein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4, 1997.

[12] M. Bazant. Lattice inversion problems with applications in solid state physics. MIT Mathematics Lecture Series, January 1999.

[13] M. Bazant. Möbius, Chebychev and Césaro on series inversion. January 1999.

[14] A. Benveniste, R. Nikoukhah, and A.S. Willsky. Multiscale system theory. In *IEEE Conference on Decision and Control*, Honolulu, Hawaii, December 1990.

[15] C. Berge. *Graphs and hypergraphs.* North-Holland Publishing Company, Amsterdam, 1976.

[16] J.G. Berryman. Analysis of approximate inverses in tomography ii: Iterative inverses. *Optimization and Engineering*, 1:437–473, 2000.

[17] D.P. Bertsekas, Angelic Nedic, and Asuman E. Ozdaglar. Convex analysis and optimization, 2003.

[18] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, 36:192–236, 1974.

[19] J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48(3):259–302, 1986.

[20] L. Boltzmann. *Wiener Ber.*, 76, 1877.

[21] R. Bowley and M. Sánchez. *Introductory Statistical Mechanics.* Oxford University Press, 1996.

[22] X. Boyen and D. Koller. Tractable inference for complex stochastic processes. In *Proc. of the Conf. on Uncertainty in AI*, volume 14, pages 33–42, 1998.

[23] M.E. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, pages 1155–1182, July 1999.

[24] L.M. Bregman. The relaxation method of finding the common point of convex sets and its applications to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, 1967.

[25] P. Brémaud. *Markov chains, Gibbs fields, Monte Carlo simulation, and queues.* Springer-Verlag, 1999.

[26] W.L. Briggs. A multigrid tutorial. *SIAM*, 1987.

[27] D. Brigo, B. Hanzon, and Francois LeGland. A differential geometric approach to nonlinear filtering: The projection filter. *IEEE Transactions on Automatic Control*, 43(2), February 1998.

[28] N.N. Chentsov (Čencov). Statistical decision rules and optimal inference. In *Translations of Mathematical Monographs*, volume 53. Americal Mathematical Society, 1982. (Originally published in Russian, Nauka, Moscow, 1972).

[29] N.N. Chentsov. A systematic theory of exponential families of probability distributions. *Theory of Probability and its Applications*, 11:425–425, 1966.

[30] K.C. Chou. *A Stochastic Modeling Approach to Multiscale Signal Processing.* PhD thesis, Laboratory for Information and Decision Systems, MIT, May 1991. LIDS-TH-2036.

[31] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[32] R. Cowell. Advanced inference in Bayesian networks. In Jordan [76], pages 27–50.

[33] R. Cowell. Introduction to inference for Bayesian networks. In Jordan [76], pages 9–26.

[34] I. Csiszár. *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, February 1975.

[35] I. Csiszár. A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling. *Annals of Statistics*, 17:1409–1413, 1989.

[36] M.M. Daniel. Parallel algorithms for 2-D boundary value systems. Master's thesis, Laboratory for Information and Decision Systems, MIT, February 1993. LIDS-TH-2164.

[37] J.N. Darroch, S.L. Lauritzen, and T.P. Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8(3):522–539, May 1980.

[38] J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43:1470–1480, 1972.

[39] A.P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, 41:1–31, 1979.

[40] A.P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2:25–36, 1992.

[41] A. Dembo and O. Zeitouni. *Lage Deviations Techniques and Applications, 2nd ed.* Springer, New York, 1998.

[42] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[43] A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, March 1972.

[44] R.L. Dobrushin. The description of a random field by means of conditional probabilities and condition on its regularity. *Theory of Probability and its Applications*, 13:197–224, 1968.

[45] R.L. Dobrushin. Prescribing a system of random variables by conditional distribution. *Theory of Probability and its Applications*, 15:458–486, 1970.

[46] B. Efron. Defining the curvature of a statistical problem. *The Annals of Statistics*, 3:1189–1242, 1975.

[47] B. Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

[48] R. Barrett et al. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, Philadelphia, 1994.

[49] W. Fenchel. On conjugate convex functions. *Canadian Journal of Mathematics*, 1:73–77, 1949.

[50] W. Fenchel. Convex cones, sets, and functions, 1951. unpublished notes.

[51] P.W. Fieguth. Multipole-motivated reduced-state estimation. In *International Conference on Image Processing (ICIP '98)*, volume 1, pages 4–7, 1998.

[52] P.W. Fieguth, W.C. Karl, A.S. Willsky, and C. Wunsch. Multiresolution optimal interpolation and statistical analysis of topex/poseidon satellite altimetry. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):280–292, March 1995.

[53] R.A. Fisher. Theory of statistical estimation. *Proc. Cambridge Philos. Trans.*, 122:700–725, 1925.

[54] R.A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, 1956.

[55] A.B. Frakt. *Internal Multiscale Autoregressive Processes, Stochastic Realization, and Covariance Extension*. PhD thesis, Laboratory for Information and Decision Systems, MIT, August 1999. LIDS-TH-2456.

[56] A.B. Frakt, H. Lev-Ari, and A.S. Willsky. Graph-theoretic results in covariance extension with applications in maximum entropy multiresolution modeling. *in review*, 2002.

[57] A.B. Frakt and A.S. Willsky. A scale-recursive method for constructing multiscale stochastic models. *Multidimensional Signal Processing*, 12:109–142, 2001.

[58] D.C. Fraser. A new technique for the optimal smoothing of data. Technical report, Massachusetts Institute of Technology, August 1967.

[59] B. Frey, editor. *Graphical models for machine learning and digital communication*. MIT Press, 1998.

[60] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

[61] A. George. Nested dissection of a regular finite element mesh. *SIAM Journal of Numerical Analysis*, 10(2):345–363, 1973.

[62] W. Gibbs. *Elementary Principles of statistical mechanics*. Yale University Press, New Haven, 1902.

[63] G.H. Golub and C.F. Van Loan. *Matrix Computations, 3rd ed.* Johns Hopkins University Press, 1996.

[64] I.J. Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Annals of Mathematical Statistics*, 34(3):911–934, September 1963.

[65] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. *Journal of Computational Physics*, 73:325–348, 1987.

[66] G.R. Grimmett. A theorem about random fields. *Bulletin of the London Mathematical Society*, 5:81–84, 1973.

[67] X. Guyon. *Random Fields on a Network: modeling, statistics, and applications.* Springer-Verlag, 1995.

[68] J.M. Hammersley and P.E. Clifford. *Markov fields on finite graphs and lattices.* 1971. unpublished manuscript.

[69] T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Advances of Neural and Information Processing Systems*, 2002.

[70] T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proc. of the Conf. on Uncertainty in AI*, 2002.

[71] C.T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 55:179–188, 1968.

[72] W.W. Irving. *Multiresolution Stochastic Realization and Model Identification with Applications to Large-Scale Estimation Problems.* PhD thesis, Laboratory for Information and Decision Systems, MIT, September 1995. LIDS-TH-2310.

[73] W.W. Irving and A.S. Willsky. Multiscale stochastic realization using canonical correlations. *IEEE Transactions on Automatic Control*, September 2001.

[74] T.S. Jaakkola. Machine learning seminar. Lecture notes, Massachusetts Institute of Technology, February 16 – March 17 1999.

[75] E.T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, 1968.

[76] M.I. Jordan, editor. *Learning in Graphical Models.* Adaptive Computation and Machine Learning Series. MIT Press, 1999.

[77] M.I. Jordan. *Introduction to Graphical Models.* MIT Press, in preparation.

[78] R. Kalman. A new approach to linear filtering and prediction problems. *The American Society of Mechanical Engineers: Basic Engineering*, 82:35–45, March 1960.

[79] R. Kalman and R. Bucy. New results in linear filtering and prediction theory. *The American Society of Mechanical Engineers: Basic Engineering*, 83:95–108, March 1961.

[80] H. Kappen and W. Wiegerinck. A novel iteration scheme for the cluster variation method. In *Advances in Neural Information Processing Systems*, volume 14, 2002.

[81] C.T. Kelley. *Iterative Methods for Linear and Nonlinear Equations.* SIAM, Philadelphia, 1995.

[82] U. Kjærulff. Approximation of Bayesian networks through edge removals. Research Report IR-93-2007, Department of Mathematics and Computer Science, Aalborg University, Denmark, August 1993.

[83] U. Kjærulff. Reduction of computational complexity in Bayesian networks through removal of weak dependences. In *Proc. of the Conf. on Uncertainty in AI*, volume 10, pages 374–384, 1994.

[84] S. Kullback. *Information Theory and Statistics.* John Wiley, 1959. Dover reprint, 1997.

[85] S. Kullback and R.A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, March 1951.

[86] R. Kullhavý. *Recursive Nonlinear Estimation: A Geometric Approach.* Springer Verlag, 1996.

[87] S. Lakshamanan and H. Derin. Simultaneous parameter estimation and segmentation of Gibbs random fields using simulated annealing. *IEEE Transactions on PAMI*, 11(8), August 1989.

[88] S.L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford University Press, 1996.

[89] S.L. Lauritzen and D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50:157–224, 1988.

[90] R. Leahy, T. Herbert, and R. Lee. Application of Markov random fields in medical imaging. In *Information Processing in Medical Imaging*, pages 1–14, 1989.

[91] M.R. Luettgen. *Image Processing with Multiscale Stochastic Models*. PhD thesis, Laboratory for Information and Decision Systems, MIT, May 1993. LIDS-TH-2178.

[92] M.R. Luettgen, W.C. Karl, A.S. Willsky, and R.R. Tenney. Multiscale representations of Markov random fields. *IEEE Transactions on Signal Processing*, 41(12):3377–3395, 1993.

[93] M.R. Luettgen and A.S. Willsky. Likelihood calculation for a class of multiscale stochastic models, with application to texture discrimination. *IEEE Transactions on Image Processing*, 4(2):194–207, 1995.

[94] A.A. Markov. Extensions of the law of large numbers to dependent events (in Russian). *Bull. Soc. Phys. Math. Kazan*, 2(15):155–156, 1906.

[95] T.P. Minka. Expectation propagation for approximate Bayesian inference. In *Proc. of the Conf. on Uncertainty in AI*, volume 17, 2001.

[96] T.P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, January 2001.

[97] S.K. Mitter. Modelling and estimation for random fields. Technical Report LIDS-P-2167, Laboratory for Information and Decision Systems, MIT, November 1992.

[98] A.F. Möbius. Uber eine besondere art von umkehrung der reihen. *Journal fur die Reine und Angewandte Mathematik*, 9, 1832.

[99] K. Murphy and Y. Weiss. The factored frontier algorithm for approximate inference in DBNs. Technical report, Computer Science Department, UC Berkeley, 2000.

[100] M. Opper and D. Saad, editors. *Advanced Mean Field Methods: Theory and Practice*. Neural Information Processing Series. MIT Press, 2001.

[101] G. Parisi. *Statistical field theory*. Addison-Wesley, 1988.

[102] E. Parzen, K. Tanabe, and G. Kitagawa, editors. *Selected Papers of Hirotuga Akaike*. Springer Series in Statistics. Springer-Verlag, 1998.

[103] Y. Pawitan. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press, 2001.

[104] D.W. Peaceman and H.H. Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the SIAM*, 3(1):28–41, March 1955.

[105] J. Pearl. *Probabilistic inference in intelligent systems*. Morgan Kaufmann, 1988.

[106] S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):28–41, March 1997.

[107] C. Preston. *Random Fields*. Springer-Verlag, 1974.

[108] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[109] C.R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Culcutta Mathematical Society*, 37:81–91, 1945.

[110] H. Rauch, F. Tung, and C. Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, 3(8):1445–1450, August 1965.

[111] J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11(2):416–431, June 1983.

[112] J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14(3):1080–1100, September 1986.

[113] J. Rissanen. Stochastic complexity. *Journal of the Royal Statistical Society (Series B)*, 49(3):223–239, 1987.

[114] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[115] R.T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, 1974.

[116] V. Rokhlin. Rapid solution of integral equations of classical potential theory. *Journal of Computational Physics*, 60:187–207, 1983.

[117] S. Roweis and Z. Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345, 1999.

[118] Y.A. Rozanov. Gaussian random fields with given conditional distributions. *Thoery of Probability and its Applications*, 12:381–391, 1967.

[119] H. Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338, 2001.

[120] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[121] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[122] P.P. Shenoy and G.R. Shafer. Axioms for probability and belief-function propagation. In *Uncertainty in artificial intelligence IV*, pages 169–198. North-Holland, Amsterdam, 1990.

[123] P. Smith and J. Whittaker. Edge exclusion tests for graphical models. In Jordan [76], pages 555–574.

[124] T.P. Speed and H.T. Kiiveri. Gaussian Markov distributions over finite graphs. *The Annals of Statistics*, 14(1):138–150, March 1986.

[125] E. Sudderth. Multiscale modeling and estimation using graphs with loops. Master's thesis, Laboratory for Information and Decision Systems, MIT, February 2002.

[126] D. Taylor. *Parallel Estimation of One and Two Dimensional Systems*. PhD thesis, Laboratory for Information and Decision Systems, MIT, February 1992. LIDS-TH-2092.

[127] D. Taylor and A.S. Willsky. Parallel smoothing algorithms for causal and acausal systems. Technical Report LIDS-P-2027, Laboratory for Information and Decision Systems, MIT, March 1991.

[128] D.S. Tucker. *Multiresolution Modeling from Data and Partial Specifications*. PhD thesis, Laboratory for Information and Decision Systems, MIT, in preparation 2003.

[129] M.J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, Dept. of Electrical Engineering and Computer Science, MIT, January 2002.

[130] R.B. Washburn, W.W. Irving, J.K. Johnson, D.S. Avtgis, J.W. Wissinger, R.R. Tenney, and A.S. Willsky. Multiresolution image compression and image fusion algorithms. Technical report, Alphatech, Inc., February 1996.

[131] Y. Weiss and W.T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural Computation*, 13:2173–2200, 2001.

[132] A.S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, August 2002.

[133] J.W. Woods. Two-dimensional discrete Markov random fields. *IEEE Transactions on Information Theory*, 18(2):232–240, March 1972.

[134] J.W. Woods. Markov image modeling. *IEEE Transactions on Automatic Control*, 23(5):846–850, October 1978.

[135] J.S. Yedidia. An idiosyncratic journey beyond mean field thoery. In Opper and Saad [100], pages 21–36.

[136] J.S. Yedidia, W.T. Freeman, and Y. Weiss. Generalized belief propagation. In *Neural Information Processing Systems 13*, pages 689–695. MIT Press, 2001.

[137] R.W. Yeung, T.T. Lee, and Z. Ye. Information-theoretic characterizations of conditional mutual independence and Markov random fields. *IEEE Transactions on Information Theory*, 48(7):1996–2011, July 2002.

[138] D.M. Young. *Iterative Solution of Large Linear Systems*. Academic Press, New York, 1971.

[139] A. Yuille. A double-loop algorithm to minimize the Bethe and Kikuchi free energies. *Neural Computation*, 2001.