Necessary and sufficient conditions for high-dimensional salient feature subset recovery

Citation: Tan, Vincent Y. F., Matthew Johnson, and Alan S. Willsky. “Necessary and Sufficient Conditions for High-dimensional Salient Feature Subset Recovery.” IEEE International Symposium on Information Theory Proceedings (ISIT), 2010. 1388–1392. ©2010 IEEE

As Published: http://dx.doi.org/10.1109/ISIT.2010.5513598

Publisher: Institute of Electrical and Electronics Engineers (IEEE)

Persistent URL: http://hdl.handle.net/1721.1/73588

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Necessary and Sufficient Conditions for High-Dimensional Salient Feature Subset Recovery

Vincent Y. F. Tan, Matthew Johnson, and Alan S. Willsky

Stochastic Systems Group, LIDS, MIT, Cambridge, MA 02139, Email: {vtan,mattjj,willsky}@mit.edu

***Abstract*—We consider recovering the salient feature subset for distinguishing between two probability models from i.i.d. samples. Identifying the salient set improves discrimination performance and reduces complexity. The focus in this work is on the high-dimensional regime where the number of variables d, the number of salient variables k and the number of samples n all grow. The definition of saliency is motivated by error exponents in a binary hypothesis test and is stated in terms of relative entropies. It is shown that if n grows faster than $\max\{ck\log((d-k)/k), \exp(c'k)\}$ for constants c, c', then the error probability in selecting the salient set can be made arbitrarily small. Thus, n can be much smaller than d. The exponential rate of decay and converse theorems are also provided. An efficient and consistent algorithm is proposed when the distributions are graphical models which are Markov on trees.**

***Index Terms*—Salient feature subset, High-dimensional, Error exponents, Binary hypothesis testing, Tree distributions.**

I. Introduction

Consider the following scenario: There are 1000 children participating in a longitudinal study in childhood asthma of which 500 of them are asthmatic and the other 500 are not. 10^6 measurements of possibly relevant features (*e.g.* genetic, environmental, physiological) are taken from each child but only a very small subset of these (say 30) is useful in predicting whether the child has asthma. This example is, in fact, modeled after a real-life large-scale experiment — the Manchester Asthma and Allergy Study (http://www.maas.org.uk/). The correct identification and subsequent interpretation of this *salient* subset is important to clinicians for assessing the susceptibility of other children to asthma. We expect that by focusing only on the 30 salient features, we can improve discrimination and reduce the computational cost in coming up with a decision rule. Indeed, when the salient set is small compared to the overall dimension (10^6), we also expect to be able to estimate the salient set with a small number of samples.

This general problem is also known as *feature subset selection* [1]. In this paper, we derive, from an information-theoretic perspective, necessary and sufficient conditions so that the salient set can be recovered with arbitrarily low error probability in the high-dimensional regime, *i.e.*, when the *number of samples* n, the *number of variables* d and the *number of salient variables* k grow. Intuitively, we expect that if k and d do not grow too quickly with n, then consistent recovery is possible. However, in this paper, we focus on the interesting case where $d \gg n, k$, which is most relevant to problems such as the asthma example above.

Motivated by the Chernoff-Stein lemma [2, Ch. 11] for a binary hypothesis test under the Neyman-Pearson framework, we define the notion of *saliency* for distinguishing between two probability distributions. We show that this definition of saliency can also be motivated by the same hypothesis testing problem under the Bayesian framework, in which the overall error probability is minimized. For the asthma example, intuitively, a feature is salient if it is useful in predicting whether a child has asthma and we also expect the number of salient features to be very small. Also, conditioned on the salient features, the non-salient ones should not contribute to the distinguishability of the classes. Our mathematical model and definition of saliency in terms of the KL-divergence (or Chernoff information) captures this intuition.

There are three main contributions in this work. Firstly, we provide *sufficient* conditions on the scaling of the model parameters (n, d, k) so the salient set is recoverable asymptotically. Secondly, by modeling the salient set as a uniform random variable (over all sets of size k), we derive a *necessary* condition that *any* decoder must satisfy in order to recover the salient set. Thirdly, in light of the fact that the exhaustive search decoder is computationally infeasible, we examine the case in which the underlying distributions are Markov on trees and derive efficient tree-based combinatorial optimization algorithms to search for the salient set.

The literature on feature subset selection (or variable extraction) is vast. See [1] (and references therein) for a thorough review of the field. The traditional methods include the so-called *wrapper* (assessing different subsets for their usefulness in predicting the class) and *filter* (ranking) methods. Our definition of saliency is related to the minimum-redundancy, maximum-relevancy model in [3] and the notion of Markov blankets in [4] but is expressed using information-theoretic quantities motivated by hypothesis testing. The algorithm suggested in [5] shows that the generalization error remains small even in the presence of a large number of irrelevant features, but this paper focuses on exact recovery of the salient set given scaling laws on (n, d, k). This work is also related to [6] on sparsity pattern recovery (or compressed sensing) but does not assume the linear observation model. Rather, samples are drawn from two arbitrary discrete multivariate probability distributions.

This work is supported by a AFOSR funded through Grant FA9559-08-1-1080, a MURI funded through ARO Grant W911NF-06-1-0076 and a MURI funded through AFOSR Grant FA9550-06-1-0324. V. Tan and M. Johnson are supported by A*STAR, Singapore and the NSF fellowship respectively.

978-1-4244-7892-7/10/$26.00 ©2010 IEEE

II. Notation, System Model and Definitions

Let $\mathcal{X}$ be a finite set and let $\mathcal{P}(\mathcal{X}^d)$ denote the set of discrete distributions supported on $\mathcal{X}^d$. Let $\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}$ be two sequences of distributions where $P^{(d)}, Q^{(d)} \in \mathcal{P}(\mathcal{X}^d)$, are the distributions of d-dimensional random vectors $\mathbf{x}, \mathbf{y}$ respectively. For a vector $\mathbf{x} \in \mathcal{X}^d$, $\mathbf{x}_A$ is the length-$|A|$ subvector that consists of the elements in A. Let $A^c := V_d \setminus A$. In addition, let $V_d := \{1, \ldots, d\}$ be the *index set* and for a subset $A \subset V_d$, let $P_A^{(d)}$ be the marginal of the subset of random variables in A, i.e., the random vector $\mathbf{x}_A$. Each index $i \in V_d$, associated to marginals $(P_i^{(d)}, Q_i^{(d)})$, will be generically called a *feature*.

We assume that for each pair $(P^{(d)}, Q^{(d)})$, there exists a set of n i.i.d. samples $(\mathbf{x}^n, \mathbf{y}^n) := (\{\mathbf{x}^{(l)}\}_{l=1}^n, \{\mathbf{y}^{(l)}\}_{l=1}^n)$ drawn from $P^{(d)} \times Q^{(d)}$. Each sample $\mathbf{x}^{(l)}$ (and also $\mathbf{y}^{(l)}$) belongs to $\mathcal{X}^d$. Our goal is to distinguish between $P^{(d)}$ and $Q^{(d)}$ using the samples. Note that for each d, this setup is analogous to binary classification where one does not have access to the underlying distributions but only samples from the distribution. We suppress the dependence of $(\mathbf{x}^n, \mathbf{y}^n)$ on the dimensionality d when the lengths of the vectors are clear from the context.

A. Definition of The Salient Set of Features

We now motivate the notion of saliency (and the salient set) by considering the following binary hypothesis testing problem. There are n i.i.d. d-dimensional samples $\mathbf{z}^n := \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)}\}$ drawn from either $P^{(d)}$ or $Q^{(d)}$, *i.e.*,

$$H_0 : \mathbf{z}^n \sim P^{(d)}, \qquad H_1 : \mathbf{z}^n \sim Q^{(d)}. \tag{1}$$

The Chernoff-Stein lemma [2, Theorem 11.8.3] says that the error exponent for (1) under the Neyman-Pearson formulation is

$$D(P^{(d)} \,||\, Q^{(d)}) := \sum_{\mathbf{z}} P^{(d)}(\mathbf{z}) \log \frac{P^{(d)}(\mathbf{z})}{Q^{(d)}(\mathbf{z})}. \tag{2}$$

More precisely, if the probability of false alarm $P_{\mathrm{FA}} = \Pr(\hat{H}_1|H_0)$ is kept below α, then the probability of mis-detection $P_{\mathrm{M}} = \Pr(\hat{H}_0|H_1)$ tends to zero exponentially fast as $n \to \infty$ with exponent given by $D(P^{(d)} \,||\, Q^{(d)})$ in (2).

In the Bayesian formulation, we seek to minimize the overall probability of error $\Pr(\mathrm{err}) = \Pr(H_0)P_{\mathrm{FA}} + \Pr(H_1)P_{\mathrm{M}}$, where $\Pr(H_0)$ and $\Pr(H_1)$ are the prior probabilities of hypotheses H_0 and H_1 respectively. It is known [2, Theorem 11.9.1] that in this case, the error exponent governing the rate of decay of $\Pr(\mathrm{err})$ with the sample size n is the *Chernoff information* between $P^{(d)}$ and $Q^{(d)}$:

$$D^*(P^{(d)}, Q^{(d)}) := -\min_{t\in[0,1]} \log \sum_{\mathbf{z}} (P^{(d)}(\mathbf{z}))^t (Q^{(d)}(\mathbf{z}))^{1-t}. \tag{3}$$

Similar to the KL-divergence, $D^*(P^{(d)}, Q^{(d)})$ is a measure of the separability of the distributions. It is a symmetric quantity in the distributions but is still not a metric. Given the form of the error exponents in (2) and (3), we would like to identify a size-k subset of features $S_d \subset V_d$ that "maximally distinguishes" between $P^{(d)}$ and $Q^{(d)}$. This motivates the following definitions:

Definition 1 (KL-divergence Salient Set): A subset $S_d \subset V_d$ of size k is *KL-divergence salient* (or simply salient) if

$$D(P^{(d)} \,||\, Q^{(d)}) = D(P_{S_d}^{(d)} \,||\, Q_{S_d}^{(d)}), \tag{4}$$

Thus, conditioned on the variables in the salient set S_d (with $|S_d| = k$ for some $1 \leq k \leq d$), the variables in the complement S_d^c do not contribute to the distinguishability (in terms of the KL-divergence) of $P^{(d)}$ and $Q^{(d)}$.

Definition 2 (Chernoff information Salient Set): A subset $S_d \subset V_d$ of size k is *Chernoff information salient* if

$$D^*(P^{(d)}, Q^{(d)}) = D^*(P_{S_d}^{(d)}, Q_{S_d}^{(d)}), \tag{5}$$

Thus, given the variables in S_d, the remaining variables in S_d^c do not contribute to the Chernoff information defined in (3). A natural question to ask is whether the two definitions above are equivalent. We claim the following lemma.

Lemma 1 (Equivalence of Saliency Definitions): For a subset $S_d \subset V_d$ of size k, the following are equivalent:

S1: S_d is KL-divergence salient.
S2: S_d is Chernoff information salient.
S3: $P^{(d)}$ and $Q^{(d)}$ admit the following decompositions:

$$P^{(d)} = P_{S_d}^{(d)} \cdot W_{S_d^c|S_d}, \quad Q^{(d)} = Q_{S_d}^{(d)} \cdot W_{S_d^c|S_d}. \tag{6}$$

Lemma 1 is proved using Hölder's inequality and Jensen's inequality.[1]

Observe from (6) that the conditionals $W_{S_d^c|S_d}$ of both models are identical. Consequently, the *likelihood ratio test* (LRT) [7, Sec. 3.4] between $P^{(d)}$ and $Q^{(d)}$ depends solely on the marginals of the salient set S_d, *i.e.*,

$$\frac{1}{n}\sum_{l=1}^{n} \log \frac{P^{(d)}(\mathbf{z}^{(l)})}{Q^{(d)}(\mathbf{z}^{(l)})} = \frac{1}{n}\sum_{l=1}^{n} \log \frac{P_{S_d}^{(d)}(\mathbf{z}_{S_d}^{(l)})}{Q_{S_d}^{(d)}(\mathbf{z}_{S_d}^{(l)})} \underset{\hat{H}=H_1}{\overset{\hat{H}=H_0}{\gtrless}} \gamma_n, \tag{7}$$

is the most powerful test [2, Ch. 11] of fixed size α for threshold γ_n.[2] Also, the inclusion of *any* non-salient subset of features $B \subset S_d^c$ keeps the likelihood ratio in (7) exactly the same, *i.e.*, $P_{S_d}^{(d)}/Q_{S_d}^{(d)} = P_{S_d\cup B}^{(d)}/Q_{S_d\cup B}^{(d)}$. Moreover, correctly identifying S_d from the set of samples $(\mathbf{x}^n, \mathbf{y}^n)$ results in a reduction in the number of relevant features, which is advantageous for the design of parsimonious and efficient binary classifiers.

Because of this equivalence of definitions of saliency (in terms of the Chernoff-Stein exponent in (2) and the Chernoff information in (3)), if we have successfully identified the salient set in (4), we have also found the subset that maximizes the error exponent associated to the overall probability of error $\Pr(\mathrm{err})$. In our results, we find that the characterization of saliency in terms of (4) is more convenient than its equivalent characterization in (5). Finally, we emphasize that the number of variables and the number of salient variables $k = |S_d|$ can grow as functions of n, *i.e.*, $d = d(n), k = k(n)$. In the

[1] Due to space constraints, the proofs of the results are not included here but can be found at http://web.mit.edu/vtan/www/isit10.

[2] We have implicitly assumed that the distributions $P^{(d)}, Q^{(d)}$ are nowhere zero and consequently the conditional $W_{S_d^c|S_d}$ is also nowhere zero.

sequel, we provide necessary and sufficient conditions for the asymptotic recovery of S_d as the model parameters scale, *i.e.*, when (n, d, k) all grow.

B. Definition of Achievability

Let $\mathfrak{S}_{k,d} := \{A : A \subset V_d, |A| = k\}$ be the set of cardinality-k subsets in V_d. A *decoder* is a set-valued function $\psi_n : (\mathcal{X}^d)^n \times (\mathcal{X}^d)^n \to \mathfrak{S}_{k,d}$. Note in this paper that the decoder is given k.[3] In the following, we use the notation $\widehat{P}^{(d)}, \widehat{Q}^{(d)}$ to denote the *empirical distributions* (or types) of $\mathbf{x}^n, \mathbf{y}^n$ respectively.

Definition 3 (Exhaustive Search Decoder): The *exhaustive search decoder* (ESD) $\psi_n^* : (\mathcal{X}^d)^n \times (\mathcal{X}^d)^n \to \mathfrak{S}_{k,d}$ is given as

$$\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \;\in\; \underset{S'_d \in \mathfrak{S}_{k,d}}{\operatorname{argmax}} \; D(\widehat{P}^{(d)}_{S'_d} \,||\, \widehat{Q}^{(d)}_{S'_d}). \tag{8}$$

If the argmax in (8) is not unique, output any set $S'_d \in \mathfrak{S}_{k,d}$ that maximizes the objective. The ESD is closely related to the maximum-likelihood (ML) decoder, and can be viewed as an approximation that is tractable for analysis. For a discussion, please see the supplementary material.

We remark that, in practice, the ESD is computationally infeasible for large d and k since it has to compute the *empirical KL-divergence* $D(\widehat{P}^{(d)}_{S'_d} || \widehat{Q}^{(d)}_{S'_d})$ for all sets in $\mathfrak{S}_{k,d}$. In Section IV, we analyze how to reduce the complexity of (8) for tree distributions. Nonetheless, the ESD is *consistent* for fixed d and k. That is, as $n \to \infty$, the probability that a non-salient set is selected by ψ_n^* tends to zero. We provide the exponential rate of decay in Section III-B. Let $\mathbb{P}^n := (P^{(d)} \times Q^{(d)})^n$ denote the n-fold product probability measure of $P^{(d)} \times Q^{(d)}$.

Definition 4 (Achievability): We say that the sequence of model parameters $\{(n, d, k)\}_{n \in \mathbb{N}}$ is *achievable* for the sequence of distributions $\{P^{(d)}, Q^{(d)} \in \mathcal{P}(\mathcal{X}^d)\}_{d \in \mathbb{N}}$ if there exists a decoder ψ_n such that to every $\epsilon > 0$, there exists a $N_\epsilon \in \mathbb{N}$ for which the error probability

$$p_n(\psi_n) := \mathbb{P}^n(\psi_n(\mathbf{x}^n, \mathbf{y}^n) \neq S_d) < \epsilon, \; \forall\, n > N_\epsilon. \tag{9}$$

Thus, if $\{(n, d, k)\}_{n \in \mathbb{N}}$ is achievable, $\lim_n p_n(\psi_n) = 0$.
In the sequel, we provide achievability conditions for the ESD.

III. Conditions for the High-Dimensional Recovery of Salient Subsets

In this section, we state three assumptions on the sequence of distributions $\{P^{(d)}, Q^{(d)}\}_{d \in \mathbb{N}}$ such that under some specified scaling laws, the triple of model parameters (n, d, k) is achievable with the ESD as defined in (9). We provide both *positive* (achievability) and *negative* (converse) sample complexity results under these assumptions. That is, we state when (9) holds and also when the sequence $p_n(\psi_n)$ is uniformly bounded away from zero.

A. Assumptions on the Distributions

In order to state our results, we assume that the sequence of probability distributions $\{P^{(d)}, Q^{(d)}\}_{d \in \mathbb{N}}$ satisfy the following three conditions:

A1: (*Saliency*) For each pair of distributions $P^{(d)}, Q^{(d)}$, there exists a salient set $S_d \subset V_d$ of cardinality k such that (4) (or equivalently (5)) holds.

A2: (*η-Distinguishability*) There exists a constant $\eta > 0$, independent of (n, d, k), such that for all $d \in \mathbb{N}$ and for all non-salient subsets $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$, we have

$$D(P^{(d)}_{S_d} \,||\, Q^{(d)}_{S_d}) - D(P^{(d)}_{S'_d} \,||\, Q^{(d)}_{S'_d}) \geq \eta > 0. \tag{10}$$

A3: (*L-Boundedness of the Likelihood Ratio*) There exists a $L \in (0, \infty)$, independent of (n, d, k), such that for all $d \in \mathbb{N}$, we have $\log[P^{(d)}_{S_d}(\mathbf{x}_{S_d}) / Q^{(d)}_{S_d}(\mathbf{x}_{S_d})] \in [-L, L]$ for all $\mathbf{x}_{S_d} \in \mathcal{X}^k$.

Assumption A1 pertains to the existence of a salient set. Assumption A2 allows us to employ the large deviation principle [7] to quantify error probabilities. This is because all non-salient subsets $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$ are such that their divergences are uniformly smaller than the divergences on S_d, the salient set. Thus, for each d, the associated salient set S_d is *unique* and the error probability of selecting any non-salient set S'_d decays exponentially. A2 together with A3, a regularity condition, allows us to prove that the *exponents* of all the possible error events are *uniformly* bounded away from zero. In the next subsection, we formally define the notion of an error exponent for the recovery of salient subsets.

B. Fixed Number of Variables d and Salient Variables k

In this section, we consider the situation when d and k are constant. This provides key insights for developing achievability results when (n, d, k) scale. Under this scenario, we have a large deviations principle for the error event in (9). We define the *error exponent* for the ESD ψ_n^* as

$$C(P^{(d)}, Q^{(d)}) := -\lim_{n \to \infty} n^{-1} \log \mathbb{P}^n(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) \neq S_d). \tag{11}$$

Let $J_{S'_d|S_d}$ be the *error rate* at which the non-salient set $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$ is selected by the ESD, *i.e.*,

$$J_{S'_d|S_d} := -\lim_{n \to \infty} n^{-1} \log \mathbb{P}^n\left(\psi_n^*(\mathbf{x}^n, \mathbf{y}^n) = S'_d\right). \tag{12}$$

For each $S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}$, also define the set of distributions

$$\begin{aligned} \Gamma_{S'_d|S_d} := \; & \{(P, Q) \in \mathcal{P}(\mathcal{X}^{2|S_d \cup S'_d|}) : \\ & D(P_{S_d} \,||\, Q_{S_d}) = D(P_{S'_d} \,||\, Q_{S'_d})\}. \end{aligned} \tag{13}$$

Proposition 2 (Error Exponent as Minimum Error Rate): Assume that the ESD ψ_n^* is used. If d and k are constant, then the error exponent (11) is given as

$$C(P^{(d)}, Q^{(d)}) = \min_{S'_d \in \mathfrak{S}_{k,d} \setminus \{S_d\}} J_{S'_d|S_d}, \tag{14}$$

where the error rate $J_{S'_d|S_d}$, defined in (12), is

$$J_{S'_d|S_d} = \min_{\nu \in \Gamma_{S'_d|S_d}} D(\nu \,||\, P^{(d)}_{S_d \cup S'_d} \times Q^{(d)}_{S_d \cup S'_d}). \tag{15}$$

Furthermore if A2 holds, $C(P^{(d)}, Q^{(d)}) > 0$ and hence the error probability in (9) decays exponentially fast in n.
This result is proved using Sanov's Theorem and the contraction principle [7, Ch. 4] in large deviations.

[3] We will discuss the recovery of S_d without knowledge of k in a longer version of this paper.

C. An Achievability Result for the High-Dimensional Case

We now consider the high-dimensional scenario when (n, d, k) all scale and we have a sequence of salient set recovery problems indexed by n for the probability models $\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}$. Thus, $d = d(n)$ and $k = k(n)$ and we are searching for how such dependencies must behave (scale) such that we have achievability. This is of interest since this regime (typically $d \gg n, k$) is most applicable to many practical problems and modern datasets such as the motivating example in the introduction. Before stating our main theorem, we define the *greatest lower bound (g.l.b.) of the error exponents* as

$$B := B(\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}) := \inf_{d\in\mathbb{N}} C(P^{(d)}, Q^{(d)}), \tag{16}$$

where $C(P^{(d)}, Q^{(d)})$ is given in (14). Clearly, $B \geq 0$ by the non-negativity of the KL-divergence. In fact, we prove that $B > 0$ under assumptions A1 – A3, *i.e.*, the exponents in (14) are *uniformly* bounded away from zero. For $\epsilon > 0$, define

$$g_1(k, \epsilon) := \exp\left(\frac{2k \log|\mathcal{X}|}{1-\epsilon}\right), \tag{17}$$

$$g_2(d, k) := \frac{k}{B} \log\left(\frac{d-k}{k}\right). \tag{18}$$

Theorem 3 (Main Result: Achievability): Assume that A1 – A3 hold for the sequence of distributions $\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}$. If there exists an $\epsilon > 0$ and an $N \in \mathbb{N}$ such that

$$n > \max\{g_1(k, \epsilon), g_2(d, k)\}, \quad \forall\, n > N, \tag{19}$$

then $p_n(\psi_n^*) = O(\exp(-nc))$, where the exponent

$$c := B - \limsup_{n\to\infty} \frac{k}{n} \log \frac{d-k}{k} > 0. \tag{20}$$

In other words, the sequence $\{(n, d, k)\}_{n\in\mathbb{N}}$ of parameters is achievable if (19) holds. Furthermore, the exhaustive search decoder in (8) achieves the scaling law in (19).

The key elements in proof include applications of large deviations bounds (*e.g.*, Sanov's theorem), asymptotic behavior of binomial coefficients and most crucially demonstrating the positivity of the g.l.b. of the error exponents B defined in (16). We now discuss the ramifications of Theorem 3.

Firstly, $n > g_1(k, \epsilon)$ means that k, the number of salient features, is only allowed to grow logarithmically in n. Secondly, $n > g_2(d, k)$ means that if k is a constant, the number of redundant features $|S_d^c| = d - k$ can grow exponentially with n, and p_n still tends to zero exponentially fast. This means that recovery of S_d is asymptotically possible even if the data dimension is extremely large (compared to n) but the number of salient ones remain a small fraction of the total number d. We state this observation formally as a corollary of Theorem 3.

Corollary 4 (Achievability for constant k): Assume A1 – A3. Let $k = k_0$ be a constant and fix $R < R_1 := B/k_0$. Then if there exists a $N \in \mathbb{N}$ such that $n > (\log d)/R$ for all $n > N$, then the error probability obeys $p_n(\psi_n^*) = O(\exp(-nc'))$, where the exponent is $c' := B - k_0 R$.

This result means that we can recover the salient set even though the number of variables d is much larger (exponential) in the number of samples n as in the asthma example.

D. A Converse Result for the High-Dimensional Case

In this section, we state a converse theorem (and several useful corollaries) for the high-dimensional case. Specifically, we establish a condition on the scaling of (n, d, k) so that the probability of error is uniformly bounded away from zero for *any* decoder. In order to apply standard proof techniques (such as Fano's inequality) for converses that apply to all possible decoders ψ_n, we consider the following slightly modified problem setup where S_d is random and not fixed as was in Theorem 3. More precisely, let $\{\widetilde{P}^{(d)}, \widetilde{Q}^{(d)}\}_{d\in\mathbb{N}}$ be a fixed sequence of distributions, where $\widetilde{P}^{(d)}, \widetilde{Q}^{(d)} \in \mathcal{P}(\mathcal{X}^d)$. We assume that this sequence of distributions satisfies A1 – A3, namely there exists a salient set $\widetilde{S}_d \in \mathfrak{S}_{k,d}$ such that $\widetilde{P}^{(d)}, \widetilde{Q}^{(d)}$ satisfies (4) for all d.

Let Π be a permutation of V_d chosen uniformly at random, *i.e.*, $\Pr(\Pi = \pi) = 1/(d!)$ for any permutation operator $\pi : V_d \to V_d$. Define the sequence of distributions $\{P^{(d)}, Q^{(d)}\}_{d\in\mathbb{N}}$ as

$$\pi \sim \Pi, \qquad P^{(d)} := \widetilde{P}^{(d)}_{\pi}, \qquad Q^{(d)} := \widetilde{Q}^{(d)}_{\pi}. \tag{21}$$

Put simply, we permute the indices in $\widetilde{P}^{(d)}, \widetilde{Q}^{(d)}$ (according to the realization of Π) to get $P^{(d)}, Q^{(d)}$, *i.e.*, $P^{(d)}(x_1 \ldots x_d) := \widetilde{P}^{(d)}(x_{\pi(1)} \ldots x_{\pi(d)})$. Thus, once π has been drawn, the distributions $P^{(d)}$ and $Q^{(d)}$ of the random vectors $\mathbf{x}$ and $\mathbf{y}$ are completely determined. Clearly the salient sets S_d are drawn *uniformly at random (u.a.r.)* from $\mathfrak{S}_{k,d}$ and we have the Markov chain:

$$S_d \xrightarrow{\varphi_n} (\mathbf{x}^n, \mathbf{y}^n) \xrightarrow{\psi_n} \widehat{S}_d, \tag{22}$$

where the length-d random vectors $(\mathbf{x}, \mathbf{y}) \sim P^{(d)} \times Q^{(d)}$ and $\widehat{S}_d$ is any estimate of S_d. Also, φ_n is the *encoder* given by the random draw of π and (21). ψ_n is the decoder defined in Section II-B. We denote the entropy of a random vector $\mathbf{z}$ with pmf P as $H(\mathbf{z}) = \mathcal{H}(P)$ and the conditional entropy of $\mathbf{z}_A$ given $\mathbf{z}_B$ as $H(\mathbf{z}_A|\mathbf{z}_B) = \mathcal{H}(P_{A|B})$.

Theorem 5 (Converse): Assume that the salient sets $\{S_d\}_{d\in\mathbb{N}}$ are drawn u.a.r. and encoded as in (21). If

$$n < \frac{\lambda k \log(\frac{d}{k})}{\mathcal{H}(P^{(d)}) + \mathcal{H}(Q^{(d)})}, \quad \text{for some } \lambda \in (0, 1), \tag{23}$$

then $p_n(\psi_n) \geq 1 - \lambda$ for any decoder ψ_n.

The converse is proven using Fano's inequality [2, Ch. 1]. Note from (23) that if the non-salient set S_d^c consists of uniform random variables independent of those in S_d then $\mathcal{H}(P^{(d)}) = O(d)$ and the bound is never satisfied. However, the converse is interesting and useful if we consider distributions with additional structure on their entropies. In particular, we assume that most of the non-salient variables are redundant (or processed) versions of the salient ones. Again appealing to the asthma example in the introduction, there could be two features in the dataset "body mass index" (in S_d) and "is obese" (in S_d^c). These two features capture the same basic information and are thus redundant, but the former may be more informative to the asthma hypothesis.

Corollary 6 (Converse with Bound on Conditional Entropy): If there exists a $M < \infty$ such that

$$\max\{\mathcal{H}(P^{(d)}_{S_d^c|S_d}), \mathcal{H}(Q^{(d)}_{S_d^c|S_d})\} \le Mk \qquad (24)$$

for all $d \in \mathbb{N}$, and

$$n < \frac{\lambda \log(\frac{d}{k})}{2(M + \log|\mathcal{X}|)}, \quad \text{for some } \lambda \in (0,1), \qquad (25)$$

then $p_n(\psi_n) \ge 1 - \lambda$ for any decoder ψ_n.

Corollary 7 (Converse for constant k): Assume the setup in Corollary 6. Fix $R > R_2 := 2(M + \log|\mathcal{X}|)$. Then if k is a constant and if there exists an $N \in \mathbb{N}$ such that $n < (\log d)/R$ for all $n > N$, then there exist a $\delta > 0$ such that error probability $p_n(\psi_n) \ge \delta$ for all decoders ψ_n.

We previously showed (cf. Corollary 4) that there is a rate of growth R_1 so that achievability holds if $R < R_1$. Corollary 7 says that, under the specified conditions, there is also another rate R_2 so that if $R > R_2$, recovery of S_d is no longer possible.

IV. Specialization to Tree Distributions

As mentioned previously, the ESD in (8) is computationally prohibitive. In this section, we assume Markov structure on the distributions (also called graphical models [8]) and devise an efficient algorithm to reduce the computational complexity of the decoder. To do so, for each d and k, assume the following:

A4: (*Markov tree*) The distributions $P := P^{(d)}, Q := Q^{(d)}$ are undirected graphical models [8]. More specifically, P, Q are *Markov on a common tree* $T = (V(T), E(T))$, where $V(T) = \{1, \ldots, d\}$ is the *vertex set* and $E(T) \subset \binom{V}{2}$ is the *edge set*. That is, P, Q admit the factorization:

$$P(\mathbf{x}) = \prod_{i \in V(T)} P_i(x_i) \prod_{(i,j) \in E(T)} \frac{P_{i,j}(x_i, x_j)}{P_i(x_i) P_j(x_j)}. \qquad (26)$$

A5: (*Subtree*) The salient set $S := S_d$ is such that P_S, Q_S are Markov on a *common (connected) subtree* $T_S = (V(T_S), E(T_S))$ in T.

Note that $T_S \subset T$ has to be connected so that the marginals P_S and Q_S remain Markov on trees. Otherwise, additional edges may be introduced when the variables in S^c are marginalized out [8]. Under A4 and A5, the KL-divergence decomposes as:

$$D(P \,||\, Q) = \sum_{i \in V(T)} D_i + \sum_{(i,j) \in E(T)} W_{i,j}, \qquad (27)$$

where $D_i := D(P_i \,||\, Q_i)$ is the KL-divergence of the marginals and the *weights* $W_{i,j} := D_{i,j} - D_i - D_j$. A similar decomposition holds for $D(P_S \,||\, Q_S)$ with $V(T_S), E(T_S)$ in (27) in place of $V(T), E(T)$. Let $\mathcal{T}_k(T)$ be the set of subtrees with $k < d$ vertices in T, a tree with d vertices. We now describe an efficient algorithm to learn S when T is unknown.

Firstly, using the samples $(\mathbf{x}^n, \mathbf{y}^n)$, learn a *single* Chow-Liu [9] tree model T_{ML} using the sum of the empirical mutual information quantities $\{I(\widehat{P}_{i,j}) + I(\widehat{Q}_{i,j})\}$ as the edge weights. It is known that the Chow-Liu max-weight spanning tree algorithm is consistent and large deviations rates have also been studied [10]. Secondly, solve the following optimization:

$$T_k^* = \operatorname*{argmax}_{T_k' \in \mathcal{T}_k(T_{\mathrm{ML}})} \sum_{i \in V(T_k')} \widehat{D}_i + \sum_{(i,j) \in E(T_k')} \widehat{W}_{i,j}, \qquad (28)$$

where $\widehat{D}_i$ and $\widehat{W}_{i,j}$ are the empirical versions of D_i and $W_{i,j}$ respectively. In (28), the sum of the node and edge weights over all size-k subtrees in T_{ML} is maximized. The problem in (28) is known as the k-CARD TREE problem [11] and it runs in time $O(dk^2)$ using a dynamic programming procedure on trees. Thirdly, let the estimate of the salient set be the vertex set of T_k^*, i.e, $\psi_n(\mathbf{x}^n, \mathbf{y}^n) := V(T_k^*)$.

Proposition 8 (Complexity Reduction for Trees): Assume that A4 and A5 hold. Then if k, d are constant, the algorithm described above to estimate S is consistent. Moreover, the time complexity is $O(dk^2 + nd^2|\mathcal{X}|^2)$.

Hence, there are significant savings in computational complexity if the probability models P and Q are trees (26).

V. Conclusion and Further Work

In this paper, we defined the notion of saliency and provided necessary and sufficient conditions for the asymptotic recovery of salient subsets in the high-dimensional regime. In future work, we seek to strengthen these results by reducing the gap between the achievability and converse theorems. In addition, we would like to derive similar types of results for the scenario when k is unknown to the decoder. We have developed thresholding rules for discovering the number of edges in the context of learning Markov forests [12] and we believe that similar techniques apply here. We also plan to analyze error rates for the algorithm introduced in Section IV.

Acknowledgments: The authors acknowledge Prof. V. Saligrama and B. Nakiboğlu for helpful discussions.

References

[1] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Research*, vol. 3, pp. 1157–1182, 2003.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[3] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on PAMI*, vol. 27, no. 8, pp. 1226–1238, 2005.

[4] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[5] A. Y. Ng, "On feature selection: learning with exponentially many irrelevant features as training examples," in *Proc. 15th ICML*. Morgan Kaufmann, 1998, pp. 404–412.

[6] M. J. Wainwright, "Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting," *IEEE Trans. on Info. Th.*, pp. 5728–5741, Dec 2009.

[7] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer, 1998.

[8] S. Lauritzen, *Graphical Models*. Oxford University Press, USA, 1996.

[9] C. K. Chow and C. N. Liu, "Approximating discrete probability distributions with dependence trees." *IEEE Trans. on Info. Th.*, vol. 14, no. 3, pp. 462–467, May 1968.

[10] V. Y. F. Tan, A. Anandkumar, L. Tong, and A. S. Willsky, "A Large-Deviation Analysis for the ML Learning of Markov Tree Structures," *submitted to IEEE Trans. on Info. Th., Arxiv 0905.0940*, May 2009.

[11] C. Blum, "Revisiting dynamic programming for finding optimal subtrees in trees," *European J. of Ops. Research*, vol. 177, no. 1, 2007.

[12] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning High-Dimensional Markov Forest Distributions: Analysis of Error Rates," *submitted to J. Mach. Learn. Research, on Arxiv*, May 2010.