Latent Variable Graphical Model Selection Via Convex Optimization

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1109/ALLERTON.2010.5707106">http://dx.doi.org/10.1109/ALLERTON.2010.5707106</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Institute of Electrical and Electronics Engineers (IEEE)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Fri Dec 14 17:34:12 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/72612">http://hdl.handle.net/1721.1/72612</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Latent Variable Graphical Model Selection via Convex Optimization

Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky

Invited paper

Abstract—Suppose we have samples of a subset of a collection of random variables. No additional information is provided about the number of latent variables, nor of the relationship between the latent and observed variables. Is it possible to discover the number of hidden components, and to learn a statistical model over the entire collection of variables? We address this question in the setting in which the latent and observed variables are jointly Gaussian, with the conditional statistics of the observed variables conditioned on the latent variables being specified by a graphical model. As a first step we give natural conditions under which such latent-variable Gaussian graphical models are identifiable given marginal statistics of only the observed variables. Essentially these conditions require that the graphical model is “sparse”, while the effect of the latent variables is “spread out” over most of the observed variables. Next we propose a tractable convex program based on regularized maximum-likelihood for model selection in this latent-variable setting; the regularizer uses both the $\ell_1$ norm and the nuclear norm. Our modeling framework can be viewed as a combination of dimensionality reduction (to identify latent variables) and graphical modeling (to capture remaining statistical structure not attributable to the latent variables), and it consistently estimates both the number of hidden components and the conditional graphical structure among the observed variables. These results are applicable in the high-dimensional setting in which the number of observed variables grows with the number of samples of the observed variables. The geometric properties of the algebraic varieties of sparse matrices and of low-rank matrices play an important role in our analysis.

This paper is a short summary of a longer submission [8]; we refer the reader to the longer version for more details.

Statistical model selection in the high-dimensional regime arises in a number of applications. In many data analysis problems in geophysics, radiology, genetics, climate studies, and image processing, the number of samples available is comparable to or even smaller than the number of variables. However, it is well-known that empirical statistics such as sample covariance matrices are not well-behaved when both the number of samples and the number of variables are large and comparable to each other (see [20]). Model selection in such a setting is therefore both challenging and of great interest. In order for model selection to be well-posed given limited information, a key assumption that is often made is that the underlying model to be estimated only has a few degrees of freedom. Common assumptions are that the data are generated according to a graphical model, or a stationary time-series model, or a simple factor model with a few latent variables. Sometimes geometric assumptions are also made in which the data are viewed as samples drawn according to a distribution supported on a low-dimensional manifold.

A model selection problem that has received considerable attention recently is the estimation of covariance matrices in the high-dimensional setting. As the sample covariance matrix is poorly behaved in such a regime [20], [17], some form of regularization of the sample covariance is adopted based on assumptions about the true underlying covariance matrix. For example approaches based on banding the sample covariance matrix [2] have been proposed for problems in which the variables have a natural ordering (e.g., time-series), while “permutation-invariant” methods that use thresholding are useful when there is no natural variable ordering [13], [3]. These approaches provide consistency guarantees under various sparsity assumptions on the true covariance matrix. Other techniques that have been studied include methods based on shrinkage [19], [26] and factor analysis [14]. A number of papers have studied covariance estimation in the context of Gaussian graphical model selection. In a Gaussian graphical model the inverse of the covariance matrix, also called the concentration matrix, is assumed to be sparse, and the sparsity pattern reveals the conditional independence relations satisfied by the variables. The model selection method usually studied in such a setting is $\ell_1$-regularized maximum-likelihood, with the $\ell_1$ penalty applied to the entries of the inverse covariance matrix to induce sparsity. The consistency properties of such an estimator have been studied [24], [22], [18], and under suitable conditions [18], [22] this estimator is also “sparsistent”, i.e., the estimated concentration matrix has the same sparsity pattern as the true model from which the samples are generated. An alternative approach to $\ell_1$-regularized maximum-likelihood is to estimate the sparsity pattern of the concentration matrix by performing regression separately on each variable [21]; while such a method consistently estimates the sparsity pattern, it does not directly provide estimates of the covariance or concentration matrix.

In many applications throughout science and engineering, a challenge is that one may not have access to observations of all the relevant phenomena, i.e., some of the relevant variables may be hidden or unobserved. Such a scenario arises in data analysis tasks in psychology, computational biology, and economics. In general latent variables pose a significant difficulty for model selection because one may not know the number of relevant latent variables, nor the relationship between these variables and the observed variables. Typical algorithmic methods that try to get around this difficulty...
usually fix the number of latent variables as well as the structural relationship between latent and observed variables (e.g., the graphical model structure between latent and observed variables), and use the EM algorithm to fit parameters [9]. This approach suffers from the problem that one optimizes non-convex functions, and thus one may get stuck in suboptimal local minima. An alternative method that has been suggested is based on a greedy, local, combinatorial heuristic that assigns latent variables to groups of observed variables, based on some form of clustering of the observed variables [12]; however, this approach has no consistency guarantees.

We study the problem of latent-variable graphical model selection in the setting where all the variables, both observed and hidden, are jointly Gaussian. More concretely let the covariance matrix of a finite collection of jointly Gaussian random variables $X_O \cup X_H$ be denoted by $\Sigma_{(O,H)}$, where $X_O$ are the observed variables and $X_H$ are the unobserved, hidden variables. The marginal statistics corresponding to the observed variables $X_O$ are given by the marginal covariance matrix $\Sigma_O$, which is simply a submatrix of the full covariance matrix $\Sigma_{(O,H)}$. However suppose that we parameterize our model by the concentration matrix $K_{(O,H)} = \Sigma_{(O,H)}^{-1}$ which as discussed above reveals the connection to graphical models. In such a parametrization, the marginal concentration matrix $\Sigma_O^{-1}$ corresponding to the observed variables $X_O$ is given by the Schur complement [16] with respect to the block $K_H$:

$$\tilde{K}_O = \Sigma_O^{-1} = K_O - K_{O,H}K_H^{-1}K_{H,O}.$$ 

Thus if we only observe the variables $X_O$, we only have access to $\Sigma_O$ (or $\tilde{K}_O$). The two terms that compose $\tilde{K}_O$ above have interesting properties. The matrix $K_O$ specifies the concentration matrix of the conditional statistics of the observed variables given the latent variables. If these conditional statistics are given by a sparse graphical model then $K_O$ is sparse. On the other hand the matrix $K_{O,H}K_H^{-1}K_{H,O}$ serves as a summary of the effect of marginalization over the hidden variables $H$. This matrix has small rank if the number of latent, unobserved variables $H$ is small relative to the number of observed variables $O$ (the rank is equal to $|H|$). Therefore the marginal concentration matrix $\tilde{K}_O$ of the observed variables $X_O$ is generally not sparse due to the additional low-rank term $K_{O,H}K_H^{-1}K_{H,O}$. Hence standard graphical model selection techniques applied directly to the observed variables $X_O$ are not useful.

A modeling paradigm that infers the effect of the latent variables $X_H$ would be more suitable in order to provide a simple explanation of the underlying statistical structure. Hence we decompose $\tilde{K}_O$ into the sparse and low-rank components, which reveals the conditional graphical model structure in the observed variables as well as the number of and effect due to the unobserved latent variables. Such a method can be viewed as a blend of principal component analysis and graphical modeling. In standard graphical modeling one would directly approximate a concentration matrix by a sparse matrix in order to learn a sparse graphical model. On the other hand in principal component analysis the goal is to explain the statistical structure underlying a set of observations using a small number of latent variables (i.e., approximate a covariance matrix as a low-rank matrix).

In our framework based on decomposing a concentration matrix, we learn a graphical model among the observed variables conditioned on a few (additional) latent variables. Notice that in our setting these latent variables are not principal components, as the conditional statistics (conditioned on these latent variables) are given by a graphical model. Therefore we refer to these latent variables informally as hidden components.

Our first contribution is to address the fundamental question of identifiability of such latent-variable graphical models given the marginal statistics of only the observed variables. The critical point is that we need to tease apart the correlations induced due to marginalization over the latent variables from the conditional graphical model structure among the observed variables. As the identifiability problem is one of uniquely decomposing the sum of a sparse matrix and a low-rank matrix into the individual components, we study the algebraic varieties of sparse matrices and low-rank matrices. An important theme in our work is the connection between the tangent spaces to these algebraic varieties and the question of identifiability. Specifically let $\Omega(K_O)$ denote the tangent space at $K_O$ to the algebraic variety of sparse matrices, and let $T(K_{O,H}K_H^{-1}K_{H,O})$ denote the tangent space at $K_{O,H}K_H^{-1}K_{H,O}$ to the algebraic variety of low-rank matrices. Then the statistical question of identifiability of $K_O$ and $K_{O,H}K_H^{-1}K_{H,O}$ given $\tilde{K}_O$ is determined by the geometric notion of transversality of the tangent spaces $\Omega(K_O)$ and $T(K_{O,H}K_H^{-1}K_{H,O})$. The study of the transversality of these tangent spaces leads us to natural conditions for identifiability. In particular we show that latent-variable models in which (1) the sparse matrix $K_O$ has a small number of nonzeros per row/column, and (2) the low-rank matrix $K_{O,H}K_H^{-1}K_{H,O}$ has row/column spaces that are not closely aligned with the coordinate axes, are identifiable. These two conditions have natural statistical interpretations. The first condition ensures that there are no densely-connected subgraphs in the conditional graphical model structure among the observed variables $X_O$ given the hidden components, i.e., that these conditional statistics are indeed specified by a sparse graphical model. Such statistical relationships may otherwise be mistakenly attributed to the effect of marginalization over some latent variable. The second condition ensures that the effect of marginalization over the latent variables is “spread out” over many observed variables; thus, the effect of marginalization over a latent variable is not confused with the conditional graphical model structure among the observed variables. In fact the first condition is often assumed in some papers on standard graphical model selection without latent variables (see for example [22]). We note here that question of parameter identifiability was recently studied for models with discrete-valued latent variables (i.e., mixture models, hidden Markov models) [1]. However, this work is not applicable to our setting in which both the latent and observed variables are assumed to be jointly Gaussian.

1611
As our next contribution we propose a regularized maximum-likelihood decomposition framework to approximate a given sample covariance matrix by a model in which the concentration matrix decomposes into a sparse matrix and a low-rank matrix. A number of papers over the last several years have suggested that heuristics based on using the $\ell_1$ norm are very effective for recovering sparse models [10], [11], [4]. Indeed such heuristics have been effectively used, as described above, for model selection when the goal is to estimate sparse concentration matrices. In her thesis [15] Fazel suggested a convex heuristic based on the nuclear norm for rank-minimization problems in order to recover low-rank matrices. This method generalized the previously studied trace heuristic for recovering low-rank positive semidefinite matrices. Recently several conditions have been given under which these heuristics provably recover low-rank matrices in various settings [23], [5]. Motivated by the success of these heuristics, we propose the following penalized likelihood method given a sample covariance matrix $\Sigma_n$ formed from $n$ samples of the observed variables:

$$
(\hat{S}_n, \hat{L}_n) = \arg \min_{S,L} -\ell(S - L; \Sigma_n) + \lambda_n (\gamma||S||_1 + \text{tr}(L))
$$

\[\text{s.t. } S - L \succ 0, \quad L \succeq 0.\]

Here $\ell$ represents the Gaussian log-likelihood function and is given by $\ell(K; \Sigma) = \log \det(K) - \text{tr}(K \Sigma)$ for $K \succ 0$, where $\text{tr}$ is the trace of a matrix and $\det$ is the determinant. The matrix $\hat{S}_n$ provides an estimate of $K_{O,\Omega}$, which represents the conditional concentration matrix of the observed variables; the matrix $\hat{L}_n$ provides an estimate of $K_{O,\Omega} K_{H}^{-1} K_{H,\Omega}$, which represents the effect of marginalization over the latent variables. Notice that the regularization function is a combination of the $\ell_1$ norm applied to $S$ and the nuclear norm applied to $L$ (the nuclear norm reduces to the trace over the cone of symmetric, positive-semidefinite matrices), with $\gamma$ providing a tradeoff between the two terms. This variational formulation is a convex optimization problem. In particular it is a regularized max-det problem and can be solved in polynomial time using standard off-the-shelf solvers [25].

Our main result is a proof of the consistency of the estimator (1) in the high-dimensional regime in which both the number of observed variables and the number of hidden components are allowed to grow with the number of samples (of the observed variables). We show that for a suitable choice of the regularization parameter $\lambda_n$, there exists a range of values of $\gamma$ for which the estimates $(\hat{S}_n, \hat{L}_n)$ have the same sparsity (and sign) pattern and rank as $(K_{O,\Omega} K_{H}^{-1} K_{H,\Omega})$ with high probability. The key technical requirement is an identifiability condition for the two components of the marginal concentration matrix $K_{O}$ with respect to the Fisher information (see [8] for more details). We make connections between our condition and the irrepresentability conditions required for support/graphical-model recovery using $\ell_1$ regularization [27], [22]. Our results provide numerous scaling regimes under which consistency holds in latent-variable graphical model selection. For example we show that under suitable identifiability conditions consistent model selection is possible even when the number of samples and the number of latent variables are on the same order as the number of observed variables.

\[a\] Related previous work: The problem of decomposing the sum of a sparse matrix and a low-rank matrix, with no additional noise, into the individual components was initially studied in [7] by a superset of the authors of the present paper. Specifically this work proposed a convex program using a combination of the $\ell_1$ norm and the nuclear norm to recover the sparse and low-rank components, and derived conditions under which the convex program exactly recovers these components. In subsequent work Candès et al. [6] also studied this noise-free sparse-plus-low-rank decomposition problem, and provided guarantees for exact recovery using the convex program proposed in [7]. The problem setup considered in the present paper is quite different and is more challenging because we are only given access to an inexact sample covariance matrix, and we are interested in recovering components that preserve both the sparsity pattern and the rank of the components in the true underlying model. In addition to proving such a consistency result for the estimator (1), we also provide a statistical interpretation of our identifiability conditions and describe natural classes of latent-variable Gaussian graphical models that satisfy these conditions. As such our paper is closer in spirit to the many recent papers on covariance selection, but with the important difference that some of the variables are not observed.

REFERENCES


