Latent Variable Inference with Factor Graphs

Addison J. Hu

December 21, 2016

Abstract

The multivariate normal distribution is a cornerstone of statistical and graphical model theory, lending elegant properties to both fields. Graphical models provide neat abstractions for reasoning about conditional independence statements on sets of variables, while statistical inference provides a toolset for fitting parameters to data. These two objectives are joined beautifully by the multivariate Gaussian distribution, whose inverse covariance parameter may be interpreted as an adjacency matrix for the corresponding graphical model. In this project, we consider the setting in which a factor model is bipartite, conditioned on a small set of global factors influencing one of the parts. As such, we are implicitly considering tripartite graphs, although two of the parts are ever observed. We pose the search for an optimal setting of parameters under this model as an optimization problem, and provide update rules for gradient descent. We also offer an implementation for this solver in NumPy, with which we perform computational experiments on synthetic data. Future work can examine more effective update rules for maintaining the symmetric positive-semidefinite invariant and for encouraging sparsity.

1 Introduction

Graphical models provide a structured, powerful framework encoding condition dependencies between variables for performing statistical inference, with applications ranging from topic modeling to social network analysis. In this project, we propose a model for inferring latent variables, which are assumed provide a lower-rank structure to the diseases in a disease-symptom factor graph.

This report will first provide an introduction to Gaussian graphical models and the sparse plus low-rank matrix decomposition in Section 2. In Section 3 we describe our model, formulate an optimization problem, and provide gradient descent updates. We then briefly describe the data of interest in 4. Finally in Sections 5 and 6, we analyze the performance of our model and propose future work for this model.

$$p(\mathbf{x}) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Figure 1: Probability density function for the multivariate Gaussian.

2 Background

2.1 Gaussian Graphical Models

The estimation of parameters in a multivariate distribution is a classical problem in statistics. When the variables are not assumed to be independent, as is commonly the case, graphical models provide a neat framework in which we may encode conditional independence statements regarding the variables.[5]

For our purposes we consider the multivariate Gaussian distribution, which is fully parameterized by the mean μ and covariance Σ, is shown in Figure 1. This setting is particularly interesting due as its parameterization admits an adjacency matrix of its graph structure in the form of its precision matrix $\Omega = \Sigma^{-1}$.

The multivariate Gaussian is also pleasant to work with due to the fact that the marginal covariance matrix on a subset of variables is given by the principal minor of the covariance Σ with respect to that subset. Furthermore, the conditional precision matrix on a subset of variables is given by the principal minor of the precision matrix Ω with respect to the subset. Suppose we have subsets $A, B \subset V$, $A \cup B = V, A \cap B = \varnothing$. Then, we may relate these two statements by the following equations:

$$(\Omega_A)^{-1} = \Sigma_{A|B}$$
$$(\Sigma_A)^{-1} = \Omega_{A|B}$$

2.1.1 Maximum Likelihood Estimation

Given a data matrix $X \in \mathbf{R}^{n \times p}$ of n observations of p features, we may write the empirical covariance matrix as $\hat{\Sigma} = \frac{1}{n} X^\top X$. We assume the data has been zero-centered. The classical maximum likelihood problem for fitting the parameters of the covariance matrix Σ to the data is given by:

$$\begin{array}{ll} \underset{\Sigma}{\text{maximize}} & -\log\det|\Sigma| - \langle \hat{\Sigma}, \Sigma^{-1} \rangle \\ \text{subject to} & \Sigma \succeq 0 \end{array}$$

In this non-regularized case, the solution is simply the empirical estimate.

2.1.2 Graphical Lasso

To take advantage of the conditional independence property of the inverse covariance matrix Ω, numerous authors proposed techniques for directly producing sparse estimates of the inverse covariance matrix. In this spirit, Tibshirani proposed the graphical lasso[4], in which an ℓ_1 penalty is imposed on the elements of the precision matrix:

$$\begin{aligned} &\underset{\Omega}{\text{maximize}} && \log\det|\Omega| - \langle\hat{\Sigma}, \Omega\rangle - \rho\,\|\Omega\|_1 \\ &\text{subject to} && \Omega \succeq 0 \end{aligned}$$

2.2 Sparse and Low-Rank Matrix Decompositions

In many settings, a complex system can be decomposed into several simpler constituent parts. Recently, Parillo considered an approach to solving problems of this form, in which a matrix is decomposed into the addition of sparse and low-rank matrices.[2] More specifically, given a matrix $C = A^* + B^*$, with A^* an unknown sparse matrix and B^* an unknown low-rank matrix, they seek to recover A^*, B^*, through the following optimization problem:

$$\begin{aligned} &\underset{A,B}{\text{minimize}} && \gamma\,\|A\|_1 + \|B\|_* \\ &\text{subject to} && A + B = C \end{aligned}$$

In the multivariate Gaussian setting, this corresponds to decomposing a dense covariance matrix on the observed variables into a sparse component of dependencies between observed variables V, and low-rank component of latent variables H connected to the observed variables, where $|V| \gg |H|$. More specifically, we consider the decomposition via the Schur complement relation:

$$(\Sigma_V)^{-1} = \underbrace{\Omega_V}_{\text{Sparse}} - \underbrace{\Omega_{VH}\Omega_H^{-1}\Omega_{VH}}_{\text{Low-Rank}}$$

which can be solved under the framework proposed by Parillo.

This decomposition may be interpreted as discovering a relatively small set of global factors, such that given those global factors, the observed factors are largely independent.

3 Methods

3.1 Proposed Model

In this project, we adapt parts of the Tibshirani's graphical lasso and Parillo's sparse plus low-rank matrix decomposition to consider a new problem: can we discover a small set of latent factors to improve our ability to infer disease diagnoses given observed symptoms?

$$\Omega = \begin{bmatrix} \Omega_L & \Omega_{LD} & \Omega_{LS} \\ \Omega_{DL} & \Omega_D & \Omega_{DS} \\ \Omega_{SL} & \Omega_{SD} & \Omega_S \end{bmatrix}$$

Figure 2: Precision matrix in our model. Note that our model assumes $\Omega_{LS} = \Omega_{SL}^\top = 0$.

In our problem, we receive training examples $\{(\mathbf{x}_D, \mathbf{x}_S)\}_{i=1}^n$, and given a new set set of symptoms $\mathbf{x}'_S$, we want to infer $\mathbf{x}'_D$. To encode possible interactions between disease, e.g., co-occurrence, we assume the existence of a set of latent nodes L connected to the disease nodes D.

We then assume a tripartite structure, in which there are connections only between the latent and disease nodes, and between the disease and symptom nodes S.

We denote the blocks of the precision matrix in Figure 2. Note that due to our assumed tripartite structure, Ω_{LS} and its transpose are the zero matrix, and $\Omega_L, \Omega_D, \Omega_S$ are diagonal.

3.2 Regularized Maximum Likelihood Formulation

We begin by establishing notation. Let us define the empirical marginal covariance matrix on the symptoms and diseases:

$$\hat{\Sigma}_{DS} = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \mathbf{x}_S^{(i)} \\ \mathbf{x}_D^{(i)} \end{bmatrix} \begin{bmatrix} \mathbf{x}_S^{(i)} & \mathbf{x}_D^{(i)} \end{bmatrix}$$

Let us further denote for convenient the block matrix for the sparse component:

$$A = \begin{bmatrix} \Omega_D & \Omega_{DS} \\ \Omega_{DS}^\top & \Omega_S \end{bmatrix}$$

and analogously for the low-rank component:

$$B = \begin{bmatrix} \frac{1}{2} LL^\top & 0 \\ 0 & 0 \end{bmatrix}$$

We can thus consider the following optimization problem.

$$\begin{aligned} \underset{\Omega_D, \Omega_S, \Omega_{DS}, L}{\text{minimize}} \quad & -\log\det|A+B| + \langle A+B, \hat{\Sigma} \rangle + \lambda \left\|\Omega_{DS}\right\|_1 + \rho \operatorname{tr}(B) \\ \text{subject to} \quad & A \succeq 0,\ B \succeq 0, \\ & \Omega_D = \operatorname{diag}(\Omega_D),\ \Omega_S = \operatorname{diag}(\Omega_S) \end{aligned}$$

3.2.1 Gradient Descent Updates

We now give the gradient descent updates for the proposed optimization problem.

$$\begin{aligned}
\nabla_{\Omega_S} &= \operatorname{diag}\left((A+B)_S^{-1} + \Sigma_S\right) \\
\nabla_{\Omega_D} &= \operatorname{diag}\left((A+B)_D^{-1} + \Sigma_D\right) \\
\nabla_{\Omega_{DS}} &= (A+B)_{DS}^{-1} + \Sigma_{DS} + \lambda \operatorname{sign}(\Omega_{DS}) \\
\nabla_L &= (A+B)_D^{-1} L + \rho L
\end{aligned}$$

The diagonalization operations represent the projection of the gradient updates onto the set of diagonal matrices. Assuming Ω_S, Ω_D are initialized with diagonal matrices, this maintains the invariant that the graph structure is tripartite.

The $(A+B)^{-1}$ inversion of matrix sum computation can be performed more efficiently through the Woodbury matrix identity, which gives an efficient update to the inverse of a matrix subject to a low-rank perturbation.[1] In our case, assuming we append $|S|$ rows of zeros to L, we may rewrite:

$$(A+B)^{-1} = A^{-1} - A^{-1}L(\mathbf{I}_{|L|} + L^T A^{-1} L)^{-1} L^T A^{-1}$$

providing computational savings as we assume $|L| \ll |D|$.

4 Data

4.1 Synthetic Data

As a proof of concept, we randomly generated data by designing a precision matrix corresponding to our assumed tripartite model, and drawing samples from a zero-centered multivariate Gaussian distribution with the marginal covariance matrix on the symptoms and diseases. In our experiments, we took $n = 10000$ samples, $|L| = 10$ latent variables, $|D| = 50$ diseases, and $|S| = 100$ symptoms. Note that because the optimization routine only relies on the sufficient statistics of the multivariate Gaussian, the complexity of the algorithm does not grow with the number of samples.

5 Results

Using the synthetically generated data described in Section 4, we recovered the parameters of the model through gradient descent, with penalty parameters $\rho = \lambda = 10$. The gradient descent program was written in NumPy. In Figure 3, we depict the loss incurred by our estimated parameters over 2500 iterations of gradient descent, with different numbers of assumed latent variables. We observe that the loss generally decreases as we increase the

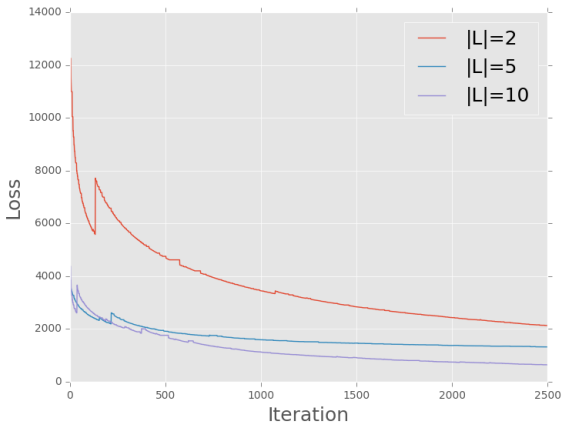
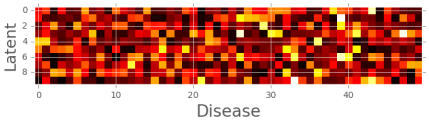


Figure 3: Loss incurred across iterations.

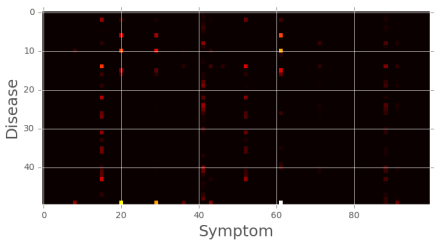
number of latent variables, as one would expect, given the generative model originally had ten latent factors.

One may observe in Figure 3 that there are periods of time during which the loss does not change. This is because the gradient would at times move the estimated parameters outside the cone of symmetric positive semi-definite matrices, and often the loss could not be computed (due to the negative determinant).

In Figure 4 we depict the recovered low-rank and sparse components of our graphical model. Lighter colors correspond to greater absolute value in the block of the precision matrix. We observe in Figure 4a that dense connections between the latent and disease variables and recovered. This is to be expected, as one of the conditions for identifiability in the sparse plus low-rank recovery framework is the assumption that there exist dense connections with the global factors [2]. In Figure 4b we observe that sparse connections between the diseases and symptoms are recovered, as encouraged by our sparsity penalty.



(a) Dense connections between latent and diseases (Ω_{LD}).



(b) Sparse connections between diseases and symptoms (Ω_{DS}).

Figure 4: Heatmaps depicting recovered sparse and low-rank components.

We also set aside a subset of our synthetic data for testing purposes. Having learned the model, we predicted the values for $\mathbf{x}_D$ having only observed a new $\mathbf{x}_S$. We found that we achieved a lower mean squared error with these predictions as opposed to the naive approach of simply using the empirical covariance matrix, which does not have knowledge of the lower dimensional latent structure.

6 Conclusion

In this project we considered extended frameworks for sparse covariance matrix estimation[4] and latent variable inference[2] to develop a model for disease diagnosis, under the assumption that interactions between the diseases may be fully explained by small set of latent variables. We implemented a solver relying on gradient descent updates and were able to recover sparse and low-rank structure from synthetic data.

Future work with respect to this model could primarily involve more effective updating schemes. Of primary concern is the fact that the gradient descent algorithm frequently moves the estimated parameters outside the cone of symmetric positive-semidefinite matrices. An updating scheme that maintains the symmetric positive-semidefinite invariant would improve the algorithm substantially.

Given we desire sparsity in Ω_{DS}, our updating scheme can also be improved by adopting strategy based upon iterative soft-thresholding or coordinate descent, as is done in the Graphical Lasso[4].

Although the Woodbury identity decreases the complexity of the matrix inverse operation required in the gradient update, it would be interesting to consider whether it is possible to maintain a sketch of the matrix inverse and update it iteratively as gradients are received, rather than recomputing it on every step. This consideration stems from the idea that the Woodbury inverse may be interpreted as an update to A^{-1} when A receives a low-rank update; however, the gradient updates considered to our knowledge are not low-rank. Such a solution would decrease the complexity of our algorithm.

Finally, as the ultimate objective of our model is to infer whether a patient has a number of diseases based on observable symptoms and test results, it is reasonable to model to disease layer as binary variables. Without the additional latent structure, this problem is essentially Logistic PCA[3]. It would be interesting to assume latent global factors influencing this binary variables, which would smooth the the output. Unfortunately, under this setting, we no longer can take advantage of the elegant theory afforded by the Gaussian assumption.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004. ISBN: 0521833787.

[2] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. "Latent variable graphical model selection via convex optimization". In: *Ann. Statist.* 40.4 (Aug. 2012), pp. 1935–1967. DOI: 10.1214/11-AOS949. URL: http://dx.doi.org/10.1214/11-AOS949.

[3] Michael Collins, Sanjoy Dasgupta, and Robert E. Schapire. "A generalization of principal component analysis to the exponential family". In: *Advances in Neural Information Processing Systems*. MIT Press, 2001.

[4] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics* 9.3 (2008), pp. 432–441. DOI: http://dx.doi.org/10.1093/biostatistics/kxm045.

[5] Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series. Oxford: Clarendon Press, 1996. ISBN: 0-19-852219-3.