1.5 GRAPHICAL MODELS: NODEWISE REGRESSION

Gaussian graphical models explore and display conditional dependence relationships among Gaussian random variables. A connection between conditional independence of variables in a Gaussian random vector and graphs is made by identifying the graph's nodes with random variables and translating the graph's edges into a parametrization that relates to the precision matrix of a multivariate normal distribution.

In this section, an overview of a regression-based approach to inducing sparsity in the precision matrix is given. Sparsity of the precision matrix is of great interest in the literature of Gaussian graphical models (cf. Whittaker, 1990; Hastie et al., 2009, Chap. 17). The key idea is that in a multivariate normal random vector $\boldsymbol{Y}$, the conditional dependencies among its entries are related to the off-diagonal entries of its precision matrix $\boldsymbol{\Sigma}^{-1} = (\sigma^{ij})$. More precisely, the variables i and j are conditionally independent given all other variables, if and only if $\sigma^{ij} = 0$, so that the problem of estimating a Gaussian graphical model is equivalent to estimating a precision matrix.

To cast the problem in the language of linear regression, let $\widehat{Y}_i = \sum_{j \neq i} \beta_{ij} Y_j$ be the linear least-squares predictor of Y_i based on the rest of the components of $\boldsymbol{Y}$ and $\varepsilon_i = Y_i - \widehat{Y}_i$ be its prediction error. Then, writing

$$Y_i = \sum_{j \neq i} \beta_{ij} Y_j + \varepsilon_i, \tag{1.16}$$

we show in Section 5.2 that the coefficients $\beta_{i,j}$ of Y_i based on the remaining variables and the corresponding prediction error variances are given by:

$$\beta_{i,j} = -\frac{\sigma^{ij}}{\sigma^{ii}},\ j \neq i, \quad \mathrm{Var}(Y_i | Y_j,\ j \neq i) = \frac{1}{\sigma^{ii}},\ i = 1, \cdots, p. \tag{1.17}$$

This confirms that σ^{ij}, the (i, j)th entry of the precision matrix is, up to a scalar, the coefficient of variable j in the multiple regression of the node or variable i on the rest. As such, each $\beta_{i,j}$ is an unconstrained real number with $\beta_{j,j} = 0$, but note that $\beta_{i,j}$ is not necessarily symmetric in (i, j).

Now, for each node i one may impose an ℓ_1 penalty on the regression coefficients in (1.16). The idea of nodewise Lasso regression, proposed first in Meinshausen and Bühlmann (2006), has been the source of considerable research in sparse estimation of precision matrices leading to the penalized normal likelihood estimation and the graphical Lasso (Glasso) algorithm discussed in Chapter 5.

1.6 CHOLESKY DECOMPOSITION AND REGRESSION

In this section, the close connection between the modified Cholesky decomposition of a covariance matrix and the idea of regression is reviewed when the variables are ordered or there is a notion of distance between variables, as in time series, longitudinal and functional data, and spectroscopic data.

In what follows, it is assumed that $\boldsymbol{Y}$ is a time-ordered random vector with mean zero and a positive-definite covariance matrix $\boldsymbol{\Sigma}$. Let $\widehat{Y}_t$ be the linear least-squares predictor of Y_t based on its predecessors $Y_{t-1}, \ldots, Y_1$ and $\varepsilon_t = Y_t - \widehat{Y}_t$ be its prediction error with variance $\sigma_t^2 = \text{var}(\varepsilon_t)$. Then, there are unique scalars ϕ_{tj} so that

$$Y_t = \sum_{j=1}^{t-1} \phi_{tj} Y_j + \varepsilon_t, \ t = 1, \cdots, p, \tag{1.18}$$

and the regression coefficients ϕ_{tj} can be computed using the covariance matrix as shown in Section 3.6. Let $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_p)'$ be the vector of successive uncorrelated prediction errors with

$$\text{cov}(\varepsilon) = \text{diag}\left(\sigma_1^2, \cdots, \sigma_p^2\right) = D.$$

Then, (1.18) can be rewritten in matrix form as $T\boldsymbol{Y} = \varepsilon$, where T is the following unit lower triangular matrix with $-\phi_{tj}$ as its (t, j)th entry:

$$T = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{p1} & -\phi_{p2} & \cdots & -\phi_{p,p-1} & 1 \end{pmatrix}. \tag{1.19}$$

Now, computing the covariance

$$\text{cov}(\varepsilon) = \text{cov}(T\boldsymbol{Y}) = T\boldsymbol{\Sigma}T',$$

one arrives at the decompositions

$$T\boldsymbol{\Sigma}T' = D, \boldsymbol{\Sigma}^{-1} = T'D^{-1}T. \tag{1.20}$$

Thus, both $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ are diagonalized by unit lower triangular matrices; we refer to (1.20) as the modified Cholesky decompositions of the covariance and its precision matrix, respectively.

Since the ϕ_{ij}'s in (1.18) are simply the regression coefficients computed from an unstructured covariance matrix, these coefficients along with $\log \sigma_t^2$ are unconstrained (Pourahmadi, 1999). From (1.18) and the subsequent results, it is evident that the task of modeling a covariance matrix is reduced to that of a sequence of p varying-coefficient and varying-order regression models. Thus, one can bring the whole regression analysis machinery to the service of the unintuitive and challenging task of modeling covariance matrices. In particular, the Lasso penalty can be imposed on the regression coefficients in (1.18) as in Huang et al. (2006). An important

consequence of (1.20) is that for any estimate $(\widehat{T}, \widehat{D})$ of the Cholesky factors, the estimated covariance and precision matrix, $\widehat{\boldsymbol{\Sigma}}^{-1} = \widehat{T}'\widehat{D}^{-1}\widehat{T}$, are guaranteed to be positive-definite.

1.7 THE BIGGER PICTURE: LATENT FACTOR MODELS

A broader framework for covariance matrix reparameterization and modeling is the class of *latent factor models* (Anderson, 2003) or *linear mixed models* (Searle et al., 1992). It has the goal of finding a few common (latent, random) factors that may explain the covariance matrix of the data.

For a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_p)'$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, the classical factor model postulates that entries of $\boldsymbol{Y}$ are linearly dependent on a few *unobservable random variables* $F = (f_1, \ldots, f_q)'$, $q \ll p$, and p additional noises $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$. The f_i's are called *the common factors* and the ε_i's are the *idiosyncratic errors*. More precisely, a *factor model* for $\boldsymbol{Y}$ is

$$\begin{aligned}
Y_1 - \mu_1 &= \ell_{11} f_1 + \cdots + \ell_{1q} f_q + \varepsilon_1, \\
Y_2 - \mu_2 &= \ell_{21} f_1 + \cdots + \ell_{2q} f_q + \varepsilon_2, \\
&\vdots \\
Y_p - \mu_p &= \ell_{p1} f_1 + \cdots + \ell_{pq} f_q + \varepsilon_p,
\end{aligned}$$

or in matrix notation

$$\boldsymbol{Y} - \boldsymbol{\mu} = LF + \varepsilon, \tag{1.21}$$

where $L = (\ell_{ij})$ is the $p \times q$ matrix of *factor loadings* (see Section 3.7 for more details).

Algebraically, the factor model (1.21) amounts to decomposing a $p \times p$ covariance matrix as

$$\boldsymbol{\Sigma} = L\Lambda L' + \Psi, \tag{1.22}$$

where L is a $p \times q$ matrix, Λ and Ψ are usually $q \times q$ and $p \times p$ diagonal matrices, respectively. Note that the factor model swaps $\boldsymbol{\Sigma}$ by the triplet (L, Λ, Ψ) which leads to a considerable reduction in the number of covariance parameters if q is small relative to p. The generality of (1.22) becomes more evident by choosing the components of the quadruple (q, L, Λ, Ψ) in various ways as indicated below:

1. *Spectral decomposition or PCA* amounts to choosing $q = p$, $\Psi = 0$, L as the orthogonal matrix of eigenvectors, and Λ as the diagonal matrix of eigenvalues of the covariance matrix.
2. *Spiked covariance models* (Paul, 2007; Johnstone and Lu, 2009) are special orthogonal factor models obtained by choosing $q < p$, Λ the identity matrix,

and $\Psi = \sigma^2 I$. In this case, writing $L = (L_1, \cdots, L_q)$ where the columns L_j's are orthogonal with decreasing norms

$$||L_1||_2 > ||L_2||_2 > \cdots > ||L_q||_2,$$

leads to the decomposition

$$\mathbf{\Sigma} = \sum_{j=1}^{q} L_j L_j' + \sigma^2 I, \tag{1.23}$$

where the L_j's are the ordered eigenvectors of the population covariance matrix. The spiked covariance models are ideal for studying consistency of high-dimensional standard PCA. For example, it will be shown in Chapter 4 that the sparse PCA is consistent for the spiked models if the leading L_j's are sparse or concentrated vectors.

3. *Modified Cholesky decomposition* amounts to taking L to be a unit lower triangular matrix, Λ the diagonal matrix of innovation variances, and $\Psi = 0$.

A possible drawback of the standard latent factor models is that the decomposition (1.22) is not unique, see (3.40) and the subsequent discussions. A common approach to ensure uniqueness is to constrain the factor loadings matrix L to be block lower triangular matrix with strictly positive diagonal elements (Geweke and Zhou, 1996). Unfortunately, such a constraint induces order dependence among the responses or the variables in $\boldsymbol{Y}$ as in the Cholesky decomposition. However, for certain tasks such as prediction, identifiability of a unique decomposition may not be necessary.

The *precision matrix* is needed in variety of statistical applications (Anderson, 2003) and in portfolio management (Fan and Lv, 2008). For the latent factor model (1.22), it can be shown that

$$\mathbf{\Sigma}^{-1} = \Psi^{-1} - \Psi^{-1} L(\Lambda + L'\Psi^{-1}L)^{-1} L'\Psi^{-1}, \tag{1.24}$$

which involves inversion of smaller $q \times q$ matrices as Ψ is assumed to be diagonal in the standard factor models. This assumption may not be valid in some economics and financial studies (Chamberlain and Rothschild, 1983) in which case it is more appropriate to assume that Ψ is sparse.

When working with a collection of covariance matrices $\mathbf{\Sigma}_1, \cdots, \mathbf{\Sigma}_K$ sharing some similarities, the factor model framework can be used to model them as

$$\mathbf{\Sigma}_k = \mathbf{\Sigma} + L\Lambda_k L', k = 1, \cdots, K, \tag{1.25}$$

where $\mathbf{\Sigma}$ is a common covariance matrix, L is a $p \times q$ full rank matrix, and Λ_k is a $q \times q$ nonnegative definite matrix which is not necessarily diagonal. The maximum likelihood estimation of such a family of reduced covariance matrices is developed in Schott (2012). An important example of a collection of covariance matrices arises in

multivariate time-varying volatility $\mathbf{\Sigma}_t, t = 1, \cdots, T$ for p assets over T time periods. It has many applications in finance including asset allocation and risk management. Here, using a standard factor model one writes

$$\mathbf{\Sigma}_t = L_t \Lambda_t L_t' + \Psi_t, t = 1, \cdots, T, \tag{1.26}$$

where the $p \times q$ matrix of factor loadings L_t is block lower triangular matrix with diagonal elements equal to one, Λ_t, Ψ_t are the diagonal covariance matrices of the common and specific factors, respectively. In the literature of finance, (1.26) is referred to as the *factor stochastic volatility* model.

For the versatility of the latent factor models in dealing with large covariances arising from spatial data, financial data, etc., see Fox and Dunson (2011). They propose a Bayesian nonparametric approach to multivariate covariance regression where L in (1.22) and hence the covariance matrix are allowed to change flexibly with covariates or predictors. The approach readily scales to high-dimensional datasets. The JRSS B discussion paper by Fan et al. (2013) offers the most recent and diverse perspectives of research on the roles of factor models in covariance estimation.

1.8 FURTHER READING

An excellent unified view of regularization methods in statistics is given by Bickel and Li (2006). All regularization methods require a tuning parameter. Choosing it too small keeps many variables in the model and retains too much variance in estimation and prediction, but choosing it too large knocks out many variables and introduces too much bias. Thus, controlling the bias–variance tradeoff is the key in selecting the tuning parameter. Cross-validation is the most popular method which requires data splitting and have been only justified for low-dimensional settings. In the following, an outline of some recent progress and relevant references are provided for interested readers.

Theoretical justifications of most methods for tuning parameter selection are usually based on oracle choices of some related parameters (like the noise variance) that cannot be implemented in practice. Choosing the regularization parameter in a data-dependent way remains a challenging problem in a high-dimensional setup. Since the introduction of the degrees of freedom for Lasso regression (Zou et al., 2007), it seems some older techniques such as the C_p-statistic, Akaike information criterion (AIC), Bayesian information criterion (BIC) are becoming popular in the high-dimensional context. Significant progress has been made recently on developing likelihood-free regularization selection techniques like subsampling (Meinshausen and Bühlmann, 2010), which are computationally expensive and still may lack theoretical guarantees.

There is a growing desire and tendency to avoid data-based methods of selecting the tuning parameter by exploiting certain optimal value of the asymptotic thresholding parameter from the Gaussian sequence models (Johnstone, 2011; Yang et al., 2011;

Cai and Liu, 2011). A new procedure (Liu and Wang, 2012) for estimating high-dimensional Gaussian graphical models, called TIGER (tuning-insensitive graph estimation and regression), enjoys a *tuning-insensitive property*, in the sense that it automatically adapts to the unknown sparsity pattern and is asymptotically tuning-free. In finite sample settings, one only needs to pay minimal attention to tuning the regularization parameter. Its main idea, like that in Glasso, is to estimate the precision matrix one column at a time. For each column, the computation is reduced to a sparse regression problem where unlike the existing methods, the TIGER solves the sparse regression problem using the recent SQRT-Lasso method proposed by Belloni et al. (2012). The main advantage of the TIGER over the existing methods is its asymptotic tuning-free property, which allows one to use the entire data to efficiently estimate the model.

PROBLEMS

1. For $y \in R, \lambda > 0$, consider the objective function

$$f_\lambda(\beta) = \beta^2 - 2y\beta + p_\lambda(|\beta|),$$

where $p_\lambda(\cdot)$ is a penalty function. Let $\hat{\beta}$ be a minimizer of the objective function.

(a) For the ridge penalty $p_\lambda(|\beta|) = \lambda\beta^2$, show that the minimizer of the objective function is

$$R(y, \lambda) = \frac{y}{1+\lambda}.$$

(b) For $p_\lambda(|\beta|) = \lambda^2 I(|\beta| \neq 0)$, plot $f_2(\beta)$ when $y = 3$. Is the function continuous? Differentiable? Convex?

(c) Show that the minimizer of $f_\lambda(y)$ is the hard-thresholding function

$$H(y, \lambda) = yI(|y| > \lambda).$$

(d) Repeat (c) for the penalty function $p_\lambda(|\beta|) = 2\lambda|\beta|$ and show that the minimizer of the objective function is the soft-thresholding function $S(y, \lambda)$.

(e) Plot $H(y, \lambda)$ and $S(y, \lambda)$ as functions of the data y for $\lambda = 1, 5, 7$. What are the similarities and differences between these two penalized least-squares estimators?

2. *Soft–hard-thresholding function*: Plot the function

$$SH(y, \lambda_1, \lambda_2) = \begin{cases} 0 & \text{if} \quad |y| \leq \lambda_1, \\ \text{sign}(y)\frac{\lambda_2(|y|-\lambda_1)}{\lambda_2-\lambda_1} & \text{if} \quad \lambda_1 \leq |y| \leq \lambda_2, \\ y & \text{if} \quad |y| > \lambda_2, \end{cases}$$