Covariance Estimation for Multivariate Linear Models

Charles Zheng

October 30, 2015

Abstract

Consider the problem of estimating the covariance of the errors, Σ, in a multivariate linear regression model. In the classical low-dimensional setting, one can estimate Σ using the empirical covariance of the ordinary-least squares residuals. In high dimensions, one can either estimate Σ by the empirical covariance of the response, or through the residuals obtained from regularized regression, but either approach introduces bias into the estimator. At the same time, even if an unbiased estimator were available, the dimensionality of the problem makes it desirable to apply shrinkage to the unbiased estimator. We study the problem of estimating Σ in the high-dimensional case, and note that one can form *debiased* estimators of Σ by estimators of the form $(y - X\hat{B})^T \Xi (y - X\hat{B})$, where Ξ is an $n \times n$ debiasing matrix. Furthermore, we consider optimal shrinkage of these debiased estimators (TODO!) Simulation results demonstrate the superiority of our approach under a variety of loss functions, especially when the covariate vectors in X exhibit clustering or contain repeats.

1 Introduction

High-dimensional covariance estimation is the problem of estimating the covariance matrix Σ of a q-dimensional random vector ϵ, from independent observations $\epsilon_1, \ldots, \epsilon_n$, when n is comparable or smaller than q. In such a regime, the MLE

$$S = \epsilon^T (I/n) \epsilon$$

performs very poorly under a variety of loss functions. Since the seminal work by Stein, numerous researchers have developed methods for optimally estimating Σ from the data, under various modelling assumptions. When no assumptions are made on the eigenvectors of Σ, it is natural to consider *shrinkage* procedures, which construct the estimator $\hat{\Sigma}$ by keeping the eigenvectors of S and transforming the eigenvalues; these procedures are therefore equivariant to orthogonal transformations. But in network inference applications, it is also common to assume *sparsity* of the precision matrix, leading to the *graphical lasso* class of procedures.

The problem of covariance estimation becomes even more challenging if E is not directly observed. In this paper we consider estimation of Σ in the context of the multivariate linear model

$$Y = XB + E$$

where Y and E are $n \times q$ matrices, X is $n \times p$ matrix, and where

$$E = \begin{pmatrix} \epsilon_1^T \\ \vdots \\ \epsilon_n^T \end{pmatrix}$$

for $\epsilon_1, \ldots, \epsilon_n$ identically and independently distributed $N(0, \Sigma)$. The usual problem of covariance estimation, where E is directly observed, is therefore a special case of the model when $XB = 0$.

There is an extensive literature on the multivariate linear model, but most of the focus is on the problem of estimating the coefficient matrix B, which for our purposes is just a nuisance parameter! Of course, knowledge of Σ helps obtain a better estimate of B, hence many works have considered estimating Σ in the multivariate linear model for this purpose. Notably, Witten et al. consider simultaneous estimation of B and Σ, under assumption of L1-sparsity for B and also L1-sparsity of Σ^{-1}. However, in many problems, the "graphical lasso" assumption of L1-sparsity of Σ^{-1} is not appropriate; hence estimation of Σ in the multivariate linear model for general Σ remains an important open problem. Therefore our focus in this work is estimation in the general case.

In the classical low-dimensional setting where $n \leq p$, we can use

$$\hat{\Sigma} = (Y - X\hat{B})^T (I/n)(Y - \hat{B})$$

where $\hat{B}$ is the OLS estimator. However, in high-dimensional settings, one cannot obtain an unbiased estimate of B. Instead, one could take use shrinkage estimator for $\hat{B}$, and estimate Σ using the residuals of the model, as before. However, due to the bias in $\hat{B}$, the empirical covariance of the residuals becomes biased as well, since

$$\mathbf{E}[(Y - X\hat{B})^T(I/n)(Y - \hat{B})] = \Sigma + \delta^T X^T X \delta$$

where $\delta = B - \hat{B}$. The bias term $\delta^T X^T X \delta$ can be substantial, especially if the true signal XB is large.

Yet even if n is much smaller than p, there are situations where it possible to construct an unbiased estimator of $\hat{\Sigma}$.

Example 1. Suppose the design matrix takes the form

$$X = \begin{pmatrix} 0 \\ \tilde{X} \end{pmatrix} \tag{1}$$

for some $(n - n_h) \times p$ matrix $\tilde{X}$. Then it is clear that the first n_h rows of Y are equal to $\epsilon_1, \ldots, \epsilon_n$, and hence the empirical covariance of the first n_h rows of Y provides an unbiased estimator. In other words, we can form an unbiased estimator

$$\hat{\Sigma} = Y^T \begin{pmatrix} \frac{1}{n_h} I_{n_h} & 0 \\ 0 & 0 \end{pmatrix} Y$$

Example 2. Suppose the design matrix takes the form

$$X = \begin{pmatrix} \tilde{X} \\ \tilde{X} \end{pmatrix}$$

i.e. that $x_i = x_{n/2+i}$ for $i = 1, \ldots, n/2$. It then follows that

$$y_i - y_{i+n/2} = B^T x_i + \epsilon_i - B^T x_{i+n/2} - \epsilon_{i+n/2} = \epsilon_i - \epsilon_{i+n/2}, \tag{2}$$

where the difference $\epsilon_i - \epsilon_{i+n/2}$ is distributed $N(0, 2\Sigma)$. Therefore, we can form an unbiased estimator as

$$\hat{\Sigma} = Y^T \begin{pmatrix} \frac{1}{n} I_{n/2} & \frac{-1}{n} I_{n/2} \\ \frac{-1}{n} I_{n/2} & \frac{1}{n} I_{n/2} \end{pmatrix}.$$

Motivated by the preceding examples, we now consider the problem of finding *debiased* estimators for general X.

2 Debiased estimator

In this section we consider two classes of covariance estimators. The first class takes the form

$$\hat{\Sigma} = Y^T \Xi Y,$$

where Ξ is some $n \times n$ matrix, which can be dependent on the design matrix X. The second class takes the form

$$\hat{\Sigma} = (Y - X\hat{B})^T \Xi (Y - X\hat{B}),$$

where $\hat{B}$ can be an arbitrary estimator of B.

2.1 First class: $Y^T \Xi Y$

It it clear that (1) and (2), the two examples of unbiased estimation introduced in the previous section, fall into the first class of estimators. It is also worth noting that in the low-dimensional case, the classical estimator

$$\hat{\Sigma}_{OLS} = \frac{1}{n}(Y - X\hat{B}_{OLS})^T (Y - X\hat{B}_{OLS})$$

falls into the first class as well. Observe that $Y - X\hat{B}_{OLS} = (I - X(X^TX)^{-1}X^T)Y$, hence $\hat{\Sigma}_{OLS} = Y^T \Xi Y$ for

$$\Xi = \frac{1}{n}(I - X(X^TX)^{-1}X^T).$$

In general, for $n << p$, it is not possible to construct an unbiased estimator for Σ. Hence we consider *debiased* estimators which minimize some combination of bias and variance. By studying the bias and variance of the random matrix $S_\Xi = Y^T \Xi Y$, we arrive at a heuristic for constructing Ξ based on the design matrix X. We have

$$\mathbf{E}[S_\Xi] = \mathbf{E}[Y^T \Xi Y] = \mathbf{E}[E^T \Xi E] + \mathbf{E}[B^T X^T \Xi X B]$$

since the cross-term has zero expectation. Let us deal with the first term. Letting $\Xi = V^T \Lambda V$, We have

$$\mathbf{E}[E^T \Xi E] = \mathbf{E}[\Sigma^{1/2} Z^T \Lambda Z \Sigma^{1/2}] = \Sigma^{1/2} \mathbf{E}[Z^T \Lambda Z] \Sigma^{1/2}$$

where Z is a $q \times n$ matrix with iid normal elements. We have

$$\mathbf{E}[Z^T \Lambda Z] = \sum_{i=1}^n \lambda_i \mathbf{E}[Z_i Z_i^T] = \sum_{i=1}^n \lambda_i I = \mathrm{tr}[\Lambda] I$$

where Z_i is the ith column of Z. Since $\mathrm{tr}[\Lambda] = \mathrm{tr}[\Xi]$ we therefore get

$$\mathbf{E}[E^T \Xi E] = \mathrm{tr}[\Xi]\Sigma$$

Hence it makes sense to require $\mathrm{tr}[\Xi] = 1$ so that $\mathbf{E}[E^T \Xi E] = \Sigma$ and therefore S_Ξ is unbiased in the case where $B = 0$.

Meanwhile, let us evaluate the trace of the bias term

$$\mathrm{bias} = \mathrm{tr}\mathbf{E}[B^T X^T \Xi X B] = \mathrm{tr}[\Xi X X^T \mathbf{E}[B B^T]]$$

If we are willing to assume a generative model for B where $\mathbf{E}[BB^T]$ is some multiple of the identity, then we get

$$\mathrm{bias} \propto \mathrm{tr}[\Xi X X^T]$$

Finally, let us consider the variance of S_Ξ. The full expression of the variance is quite complex. Hence, we instead look only at the variance

$$\mathrm{Var}[E^T \Xi E] = \mathrm{Var}[\Sigma^{1/2} Z^T \Lambda Z \Sigma^{1/2}]$$

This is still difficult to analyze, so we neglect the $\Sigma^{1/2}$ terms and consider

$$\mathrm{tr}\mathrm{Var}[Z^T \Lambda Z] = 2\mathrm{tr}[\Lambda^2] = 2\mathrm{tr}[\Xi^2].$$

Now we propose the method for choosing Ξ. As noted previously, we would like to require $\mathrm{tr}[\Xi] = 1$, so that S_Ξ is unbiased in the case of perfect signal estimation, and on top of that we would like to minimize some combination of the bias and variance. This leads to the convex program

$$\text{minimize } \mathrm{tr}[\Xi X X^T] + \frac{\eta}{2}\mathrm{tr}[\Xi^2] \text{ subject to } \mathrm{tr}[\Xi] = 1.$$

where η is a tuning parameter which controls the bias-variance tradeoff of $\hat{\Sigma}_\Xi$. Let $XX^T = \Gamma W \Gamma^T$ where W is diagonal. We claim that the minimizing Ξ also takes the form of $\Gamma D \Gamma^T$ for some diagonal D. If we accept the claim, then the above program is rewritten as

$$\text{minimize } \sum_{i=1}^n d_i \lambda_i + \frac{\eta}{2}\sum_{i=1}^n \lambda_i^2 \text{ subject to } \sum_{i=1}^n \lambda_i = 1.$$

We can solve it by forming the Lagrangian

$$\sum_{i=1}^{n} d_i\lambda_i + \frac{\eta}{2}\sum_{i=1}^{n}\lambda_i^2 + \gamma\left(1 - \sum_{i=1}^{n}\lambda_i\right) - \sum_i \mu_i\lambda_i$$

and from the KKT conditions we get

$$\begin{cases} 0 = \lambda_i = \frac{\mu+\gamma-d_i}{\eta} & \text{or} \\ 0 < \lambda_i = \frac{\gamma-d_i}{\eta} \end{cases}$$

Hence, we obtain $\lambda_i = [\frac{\gamma-d_i}{\eta}]_+$, and we can determine γ by the condition

$$\sum_{i=1}^{n}\lambda_i = \sum_{i=1}^{n}\left[\frac{\gamma-d_i}{\eta}\right]_+ = 1$$

Let $\gamma(\eta)$ denote the solution to the above. We therefore form

$$\Xi(\eta) = V\mathrm{diag}\left(\left[\frac{\gamma-d_i}{\eta}\right]_+\right)V^T$$

The above derivation does not shed light on how to choose the parameter η which controls the bias-variance tradeoff. In later sections we will outline a cross-validation approach for choosing η.

3 Shrinkage

Due to high dimensionality, it is not optimal to use an estimator S_Ξ even if it is unbiased or nearly unbiased. Building on the covariance estimation literature, we consider shrinkage estimators of the form

$$\hat{\Sigma} = V^T\mathrm{diag}(f_i(\Lambda))V$$

where f_i is some function of $\Lambda = (\lambda_1, \ldots, \lambda_q)$, and where $V^T\Lambda V$ is the eigendecomposition of S_Ξ.

TO BE CONTD!

4 Results

I have nice simulation results; will include later.