Linear Algebra Tricks for Statistics

Stephen Bates

August 6, 2019

Introduction

This document walks through a small set of linear algebra manipulations of multivariate Gaussians that are widely useful in statistics. A very helpful complement to this document is the Matrix Cookbook, which has an expansive set of matrix identities. I highly recommend you have this accessible during the qualifying exam.

Decomposing the Covariance Matrix

Suppose we have a multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$ where $X \in \mathbb{R}^p$. Since Σ is a covariance matrix, it must be positive definite, so it has a real eigendecomposition with nonnegative eigenvalues:

$$\Sigma = UDU^T$$

where U is orthonormal and D is diagonal with nonnegative entries. Equivalently, we could write this as

$$\Sigma = \sum_{j=1}^{p} d_j u_j u_j^T,$$

where d_j is the jth element of the diagonal of D, and u_j is column j of U. A consequence of this representation is that when Σ is full rank the inverse covariance has the form

$$\Sigma^{-1} = UD^{-1}U^T = \sum_{j=1}^{p} \frac{1}{d_j} u_j u_j^T,$$

and similarly we have

$$\Sigma^{1/2} = UD^{1/2}U^T = \sum_{j=1}^{p} \sqrt{d_j} u_j u_j^T,$$

Distribution of Norms

As an application of the above decomposition, we can compute the distribution of $\|X - \mu\|^2$. Let $Z \sim \mathcal{N}(0, I)$. Then

$$\begin{aligned}
\|X - \mu\|^2 &\overset{d}{=} \left\|\Sigma^{1/2} Z\right\|^2 \\
&= Z^T \Sigma Z \\
&= Z^T UDU^T Z \\
&\overset{d}{=} Z^T D Z \\
&= d_1 Z_1^2 + \cdots + d_p Z_p^2,
\end{aligned}$$

so $\|X-\mu\|^2$ is distributed as the convolution of p scaled χ_1^2 random variables.

Whitening

It is frequently useful to transform a Gaussian random variable so that the covariance matrix is the identity, i.e. we transform the problem by $\Sigma^{1/2}$. As an example, suppose we have an estimation problem where the errors are correlated:

$$Y = \mu\vec{1} + \epsilon$$

where $\epsilon \sim N(0,\Sigma)$. Suppose we wish to estimate μ and give a confidence interval. Then letting $Y' = \Sigma^{-1/2}Y$, we have

$$Y' = \mu\Sigma^{-1/2}\vec{1} + \epsilon'$$

where $\epsilon' \sim \mathcal{N}(0,I)$. We can then use standard estimation and testing techniques to estimate μ and give a confidence interval.

Rotating to a Convenient Coordinate System

Another frequently used manipulation is rotating the coordinate system (i.e. multiplying by an orthogonal matrix) so that the coordinate directions are the components of interest.

As an example suppose we are in the standard linear regression setting with $Y = X\beta + \epsilon$ where $\epsilon \sim N(0,\sigma^2 I)$, $X \in \mathbb{R}^{n\times p}$, and β, σ are unknown parameters. Suppose we wished to get an exact confidence interval for $\beta_2 - \beta_1$. Let U be an orthogonal matrix such that the first column $U_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, \ldots, 0)$. Then

$$\begin{aligned}\beta' := U^T\hat{\beta} &= U^T(X^TX)^{-1}X^TY \\ &\sim \mathcal{N}(U^T\beta, U^T(X^TX)^{-1}U\sigma^2).\end{aligned}$$

We can then form a confidence interval for the first coordinate of β' in the usual way using the t-statistic, which gives a confidence interval for $\beta_2 - \beta_1$, after rescaling.

Blockwise Matrix Inversion

For any invertible matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

the inverse can be written as

$$\Sigma^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1}\Sigma_{21}\Sigma_{11}^{-1} & (\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})^{-1} \end{bmatrix}.$$

This identity comes in handy when you actually need an expression for Σ^{-1}, for instance, if you wanted to whited a covariance matrix by hand.

A second place where this formula comes up is when you are doing estimation with nuisance parameters. As an example, consider $\theta = (\theta_1, \theta_2) \in \mathbb{R}^{p_1+p_2}$. Let I be the Fisher information, and suppose

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathcal{N}(0, I^{-1}).$$

The writing out the Fisher Information blockwise, as above, we have that

$$I_{11}^{-1} = (I_{11} - I_{12}I_{22}I_{21})^{-1}$$

is the limiting variance of $\hat{\theta}_1$. In particular, this limiting variance is larger in PSD order than I_{11}^{-1}, which would be the limiting variance if θ_2 were known. The statistical interpretation is that estimation of θ_1 is harder when the nuisance parameters is unknown, and the amount of additional variance is

$$I_{11}^{-1} - (I_{11} - I_{12} I_{22} I_{21})^{-1}.$$

In the special case where $I_{21} = 0$ (i.e. $\hat{\theta}_1$ and $\hat{\theta}_2$ are asymptotically independent), we have that the variances are the same: $(I_{11})^{-1} = (I^{-1})_{11}$, so in this case estimating the nuisance parameters does not incur extra asymptotic variance.