FIML Estimation for Multivariate Normal Data with Newton's Algorithm

This document provides the mathematical building blocks for applying Newton's algorithm to obtain maximum likelihood estimates of a mean vector and covariance matrix, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The accompanying R program shows the computations for a bivariate normal estimation problem.

Derivative Equations

The Matrix Cookbook is good resource that gives details behind some of the expressions and matrix derivatives in this document. The Cookbook is available at www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf. To begin, the first derivatives of the log-likelihood with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are as follows.

$$\frac{\partial LL}{\partial \boldsymbol{\mu}} = -N\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1}\sum_{i=1}^{N}\mathbf{Y}_i \tag{1}$$

$$\frac{\partial LL}{\partial \boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{i=1}^{N}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right) \tag{2}$$

Note that Equation 1 is a vector but Equation 2 is itself a matrix (a 2 by 2 matrix in the bivariate illustration).

Second derivatives quantify the curvature or steepness of the log-likelihood function near its peak (i.e., the rate at which the first-order slopes change across the range of parameter values). Second derivatives are obtained by applying matrix calculus rules to the previous equations, and the Hessian collects these equations in a symmetric matrix with P rows and columns, where P is the number of unique parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The Hessian consists of three unique blocks: a block each for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and a block for the cross-product derivatives involving and element from $\boldsymbol{\mu}$ and an element from $\boldsymbol{\Sigma}$.

$$\boldsymbol{H}_{\mathrm{O}}(\boldsymbol{\theta}) = \begin{pmatrix} \dfrac{\partial^2 LL}{\partial \boldsymbol{\mu}^2} & \dfrac{\partial^2 LL}{\partial \boldsymbol{\mu} \partial \boldsymbol{\Sigma}} \\ \dfrac{\partial^2 LL}{\partial \boldsymbol{\Sigma} \partial \boldsymbol{\mu}} & \dfrac{\partial^2 LL}{\partial \boldsymbol{\Sigma}^2} \end{pmatrix} \tag{3}$$

The second derivative equations below are the building blocks for the observed information matrix, the inverse of which is the variance–covariance matrix of the estimates with squared standard errors on the diagonal. The blocks of the Hessian matrix are as follows.

$$\frac{\partial^2 LL}{\partial \boldsymbol{\mu}^2} = -N\boldsymbol{\Sigma}^{-1} \tag{4}$$

$$\frac{\partial^2 LL}{\partial \boldsymbol{\Sigma}^2} = -\sum_{i=1}^{N} \mathbf{D}_V' \left(\boldsymbol{\Sigma}^{-1} \otimes \left\{ \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu})(\boldsymbol{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} - .5\boldsymbol{\Sigma}^{-1} \right\} \right) \mathbf{D}_V \tag{5}$$

$$\frac{\partial^2 LL}{\partial \boldsymbol{\mu} \partial \boldsymbol{\Sigma}} = -\sum_{i=1}^{N} \left(\boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{Y}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} \right) \mathbf{D}_V \tag{6}$$

The ⊗ symbol is a Kronecker product that multiplies one matrix by each element of another matrix, and $\mathbf{D}_V$ is the duplication matrix (Magnus & Neudecker, 1999). Each covariance term appears only once in the Hessian (and similarly, only once in the variance–covariance matrix of the estimates). The duplication matrix combines redundant off-diagonal terms into a single value. Substituting the maximum likelihood estimates into the derivative expressions, multiplying $\boldsymbol{H}_{\mathrm{O}}(\hat{\boldsymbol{\theta}})$ by −1, then taking its inverse gives the variance–covariance matrix of the estimates.

Newton's Algorithm

With Newton's algorithm, the jump from the current to the updated parameter value was as follows.

$$\theta^{(t+1)} = \theta^{(t)} - \left(\frac{\partial^2 LL}{\partial \theta^2} \right)^{-1} \left(\frac{\partial LL}{\partial \theta} \right) \text{current estimate} \ + \text{step size} \tag{7}$$

The step size, computed as the ratio of the first and second derivatives at the current parameter value θ^t, corresponds to the horizontal distance between the current estimate and the peak of the projected quadratic curve (the point at which the first derivative of the quadratic approximation equals 0). The updating step extends to more complex models with multiple parameters. In this case, the multivariate updating equation is

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \boldsymbol{H}_{\mathrm{O}}^{-1}\left(\boldsymbol{\theta}^{(t)}\right)\nabla LL^{(t)} \tag{8}$$

where $\boldsymbol{\theta}$ is a vector of parameter values, t indexes the iterations, $\boldsymbol{H}_{\mathrm{O}}^{-1}\left(\boldsymbol{\theta}^{(t)}\right)$ is the inverse of the Hessian (the matrix of second derivatives) evaluated at the current parameter estimates at iteration t, and $\nabla LL^{(t)}$ is a vector of first derivatives (the gradient vector) evaluated at the current estimates.

$$\nabla LL^{(t)} = \begin{pmatrix} \dfrac{\partial LL}{\partial \mu_X^{(t)}} \\ \dfrac{\partial LL}{\partial \mu_Y^{(t)}} \\ \vdots \\ \dfrac{\partial LL}{\partial \sigma_Y^{2(t)}} \end{pmatrix} \tag{9}$$

Recall that Equation 1 (the first derivatives with respect to $\boldsymbol{\mu}$) returned a vector, and these quantities appear as the first two elements of the gradient vector. However, the first derivative of the log-likelihood with respect to $\boldsymbol{\Sigma}$ in Equation 2 returns matrix (a 2 by 2 matrix in the bivariate example). The vec function in the equation below stacks all matrix elements into a vector (e.g., a 4-element vector if $\boldsymbol{\Sigma}$ is a 2 by 2 matrix), and the duplication matrix combines the two identifical off-diagonal elements into a single quantity.

$$\nabla LL_{\boldsymbol{\Sigma}} = -\frac{1}{2}\sum_{i=1}^{N} \mathbf{D}_V' \mathrm{vec}\left(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu})(\boldsymbol{Y}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}\right) \tag{10}$$

For example, vector if **Σ** is a 2 by 2 matrix, Equation 10 returns a 3-element vector, one element for each unique estimate in **Σ**.