1 Model and Likelihood

Consider a linear K-factor model for D assets of the form

$$y_{it} = \alpha_d + \mathbf{f}_t'\boldsymbol{\beta}_d + \varepsilon_{it}$$

where $d = 1, \ldots, D$ and $t = 1, \ldots, T$ and $\mathbf{f}_t' = (f_{t1}, \ldots, f_{tK})$ is a $K \times 1$ vector. This is a special case of the seemingly unrelated regression (SUR) model in which the regressors are *identical* across equations. Stacking observations for a given time period across assets, define $\mathbf{y}_t' = (y_{1t}, \ldots, y_{Dt})$ and analogously $\boldsymbol{\varepsilon}_t' = (\varepsilon_{t1}, \ldots \varepsilon_{tD})$. Now let $\mathbf{x}_t' = (1, \mathbf{f}_t')$ and $\boldsymbol{\gamma}_d' = (\alpha_d, \boldsymbol{\beta}_d')$ so we have

$$\mathbf{y}_t = X_t\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t$$

where $X_t = I_D \otimes \mathbf{x}_t'$ and $\boldsymbol{\gamma}' = (\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_D')$. Now, suppose that

$$\boldsymbol{\varepsilon}_t|\mathbf{x}_t \sim \text{ iid } \mathcal{N}_D(0, \Omega^{-1})$$

Let Y_T denote the full data sample, i.e. $\{\mathbf{y}_t, \mathbf{x}_t\}_{t=1}^T$. Then the likelihood is

$$\pi(\boldsymbol{\gamma}, \Omega^{-1}|Y_T) \propto |\Omega^{-1}|^{T/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}(\mathbf{y}_t - X_t\boldsymbol{\gamma})'\,\Omega^{-1}\,(\mathbf{y}_t - X_t\boldsymbol{\gamma})\right]$$

where we parameterize this problem in terms of the $D \times D$ *precision* matrix Ω^{-1} and the $p \times 1$ vector of regression coefficients $\boldsymbol{\gamma}$, where $p = D(K+1)$.

2 Prior and Posterior Distribution

To complete the model we specify the following prior distribution

$$\pi(\boldsymbol{\gamma}, \Omega^{-1}) = \mathcal{N}_p\left(\boldsymbol{\gamma}|\boldsymbol{\gamma_0}, G_0\right)\mathcal{W}_D\left(\Omega^{-1}|\rho_0, R_0\right)$$

This prior is conditionally conjugate with the normal likelihood. In particular, we have $\boldsymbol{\gamma}|\Omega^{-1}, Y_T \sim \mathcal{N}_p\left(\bar{\boldsymbol{\gamma}}, G_T\right)$ where

$$\begin{aligned} G_T &= \left[G_0^{-1} + \sum_{t=1}^{T} X_t'\Omega^{-1}X_t\right]^{-1} \\ \bar{\boldsymbol{\gamma}} &= G_T\left[G_0^{-1}\boldsymbol{\gamma}_0 + \sum_{t=1}^{T} X_t'\Omega^{-1}\mathbf{y}_t\right] \end{aligned}$$

and $\Omega^{-1}|Y_T \sim \mathcal{W}_D(\rho_0 + T, R_T)$ where

$$R_T = \left[R_0^{-1} + \sum_{t=1}^{T} (\mathbf{y}_t - X_t\boldsymbol{\gamma})(\mathbf{y}_t - X_t\boldsymbol{\gamma})'\right]^{-1}$$

3 MCMC

Using the full set of conditional posteriors, given in the preceding section, we can simulate from the joint posterior for this model using a Gibbs sampler:

1. Select a starting value $\Omega^{-1(0)}$ for the precision matrix.
2. Draw $\boldsymbol{\gamma}^{(1)} \sim \mathcal{N}\left(\bar{\boldsymbol{\gamma}}^{(1)}, G_T^{(1)}\right)$ where

$$\begin{aligned} G_T^{(1)} &= \left[G_0^{-1} + \sum_{t=1}^{T} X_t'\Omega^{-1(0)}X_t\right]^{-1} \\ \bar{\boldsymbol{\gamma}}^{(1)} &= G_T^{(1)}\left[G_0^{-1}\boldsymbol{\gamma}_0 + \sum_{t=1}^{T} X_t'\Omega^{-1(0)}\mathbf{y}_t\right] \end{aligned}$$

3. Draw $\Omega^{-1(1)} \sim \mathcal{W}_D\left(\rho_T, R_T^{(1)}\right)$ where

$$R_T^{(1)} = \left[R_0^{-1} + \sum_{t=1}^{T} \left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(1)}\right)\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(1)}\right)'\right]^{-1}$$

4. Repeat the preceding two steps a total of G times. In the gth iteration:

   (i) Draw $\boldsymbol{\gamma}^{(g)} \sim \mathcal{N}\left(\bar{\boldsymbol{\gamma}}^{(g)}, G_T^{(g)}\right)$ where

$$\begin{aligned} G_T^{(g)} &= \left[G_0^{-1} + \sum_{t=1}^{T} X_t'\Omega^{-1(g-1)}X_t\right]^{-1} \\ \bar{\boldsymbol{\gamma}}^{(g)} &= G_T^{(g)}\left[G_0^{-1}\boldsymbol{\gamma}_0 + \sum_{t=1}^{T} X_t'\Omega^{-1(g-1)}\mathbf{y}_t\right] \end{aligned}$$

(ii) Draw $\Omega^{-1(g)} \sim \mathcal{W}_D\left(\rho_T, R_T^{(g)}\right)$ where

$$R_T^{(g)} = \left[R_0^{-1} + \sum_{t=1}^{T}\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(g)}\right)\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(g)}\right)'\right]^{-1}$$

5. Discard the first B draws.

Note that in iteration g, $G_T^{(g)}$ and $\bar{\boldsymbol{\gamma}}^{(g)}$ are calculated using $\Omega^{-1(g-1)}$ while $R_T^{(g)}$ is calculated using $\boldsymbol{\gamma}^{(0)}$. This is because we choose to initialize the sample with a starting value $\Omega^{-1(0)}$ for the precision matrix rather than for the vector of regression coefficients.

4 Numerical Details for the Gibbs Sampler

4.1 Drawing from a Normal Distribution

Parameterize in terms of the precision matrix.

4.2 Drawing from a Wishart Distribution

Use the Bartlett Decomposition.

4.3 Efficient Calculation of R_T

In the second step of each iteration we compute $\left(R_0^{-1} + \sum_{t=1}^{T}\hat{\boldsymbol{\varepsilon}}_t\hat{\boldsymbol{\varepsilon}}_t'\right)^{-1}$ where $\hat{\boldsymbol{\varepsilon}}_t = \mathbf{y}_t - X_t\boldsymbol{\gamma}$. Since R_0 is simply the prior scale matrix for Ω^{-1} and hence remains unchanged during the iterations, we can pre-compute it and store the result before starting the sampler. Since X_t is a sparse matrix, there is a much more efficient and compact way to compute the sum of outer products of residuals. Define:

$$\widetilde{Y} = \begin{bmatrix} \mathbf{y}_1' \\ \vdots \\ \mathbf{y}_T' \end{bmatrix}, \quad \widetilde{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix}, \quad \Gamma = \begin{bmatrix} \boldsymbol{\gamma}_1 & \cdots & \boldsymbol{\gamma}_D \end{bmatrix}, \quad \hat{\boldsymbol{\varepsilon}} = \begin{bmatrix} \hat{\boldsymbol{\varepsilon}}_1' \\ \vdots \\ \hat{\boldsymbol{\varepsilon}}_T' \end{bmatrix}$$

so that $\hat{\boldsymbol{\varepsilon}} = \widetilde{Y} - \widetilde{X}\Gamma$. Note that the vector of regression coefficients $\boldsymbol{\gamma}$ is the vec of the *matrix* of regression coefficients Γ. Thus, expressed in terms of dense matrix operations

$$R_T^{-1} = R_0^{-1} + \left(\widetilde{Y} - \widetilde{X}\Gamma\right)' \left(\widetilde{Y} - \widetilde{X}\Gamma\right)$$

The final step is to invert this sum (which is positive definite) to calculate R_T. Note that the Matrix Inversion Lemma (Sherman-Morrison-Woodbury Formula) does *not* simplify this calculation unless $D > T$.

4.4 Efficient Calculation of G_T

Because we parameterize our multivariate normal sampler in terms of the *precision* matrix rather than the covariance matrix, we work with the *inverse* of G_T, namely

$$G_T^{-1} = G_0^{-1} + \sum_{t=1}^{T} X_t' \Omega^{-1} X_t$$

Since it is simply the prior precision matrix for the vector $\boldsymbol{\gamma}$ of regression coefficients we can pre-compute G_0 (assuming that we elicit a prior in terms of the covariance matrix). Now, the sum over $X_t' \Omega X_t$ can in fact be simplified using the properties of the Kronecker product.[1] Recall that $X_t = I_D \otimes \mathbf{x}_t$. Since $(A \otimes B)' = A' \otimes B'$,

$$X_t' \Omega^{-1} X_t = (I_D \otimes \mathbf{x}_t)\, \Omega^{-1} X_t$$

Since $\Omega^{-1} X_t = (\Omega^{-1} X_t) \otimes 1$, $\Omega^{-1} = \Omega^{-1} \otimes 1$, and $(A \otimes B)(C \otimes D) = AC \otimes BD$, provided that everything is conformable, we have

$$\begin{aligned}
(I_D \otimes \mathbf{x}_t)\, \Omega^{-1} X_t &= (I_D \otimes \mathbf{x}_t)\, (\Omega^{-1} X_t \otimes 1) \\
&= \Omega^{-1} X_t \otimes \mathbf{x}_t = \left[\Omega^{-1}(I_D \otimes \mathbf{x}_t')\right] \otimes \mathbf{x}_t \\
&= \left[(\Omega^{-1} \otimes 1)(I_D \otimes \mathbf{x}_t')\right] \otimes \mathbf{x}_t \\
&= \Omega^{-1} \otimes \mathbf{x}_t' \otimes \mathbf{x}_t
\end{aligned}$$

[1] See, e.g., Horn and Johnson (1994) Chapter 4.2.

Finally, since $A \otimes (B + C) = A \otimes B + A \otimes C$,

$$\sum_{t=1}^{T} X_t' \Omega^{-1} X_t = \sum_{t=1}^{T} \Omega^{-1} \otimes \mathbf{x}_t' \otimes \mathbf{x}_t = \Omega^{-1} \otimes \left(\sum_{t=1}^{T} \mathbf{x}_t' \otimes \mathbf{x}_t \right)$$

This is an extremely useful simplification: because $\sum_{t=1}^{T} \mathbf{x}_t' \otimes \mathbf{x}_t$ involves neither Ω^{-1} nor $\boldsymbol{\gamma}$, only the data, we can pre-compute this quantity. In fact, there is one final simplification that makes this quantity even simpler. By writing out the definition of the Kronecker Product, we see that $\mathbf{x}_t' \otimes \mathbf{x}_t = \mathbf{x}_t \mathbf{x}_t'$ and hence

$$\sum_{t=1}^{T} X_t' \Omega^{-1} X_t = \Omega^{-1} \otimes \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right) = \Omega^{-1} \otimes \widetilde{X}' \widetilde{X}$$

where $\widetilde{X}' = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_T \end{bmatrix}$. Thus, we have $G_T^{-1} = G_0^{-1} + \Omega^{-1} \otimes \widetilde{X}' \widetilde{X}$.

4.5 Efficient Calculation of $\bar{\gamma}$

5 Calculating the Marginal likelihood

The marginal likelihood is available by the method of Chib (1995). From the Chib (1995) identity, we have

$$\log m\left(\{y_t\}\right) = \log \pi\left(\gamma^*\right) + \log \pi\left(\Omega^{-1*}\right) + \sum_{t=1}^{n} \log p\left(y_t | X_t \gamma^*, \Omega^*\right) - \log \pi\left(\gamma^*, \Omega^{-1*} | Y_n\right)$$

where the last term is calculated as

$$\pi\left(\Omega^{-1*} | Y_n\right) \times \pi\left(\gamma^* | Y_n, \Omega^{-1*}\right)$$

in which the first term is estimated by averaging the full-conditional Wishart density over the draws $\left\{\gamma^{(g)}\right\}_{g=1}^{G}$ from the main MCMC run

$$\pi\left(\Omega^{-1*} | Y_n\right) = \frac{1}{G} \sum_{g=1}^{G} \mathcal{W}_d \left(\Omega^{-1*} | \rho_0 + n, \left(R_0^{-1} + \sum_{t=1}^{n} \left(y_t - X_t \gamma^{(g)}\right) \left(y_t - X_t \gamma^{(g)}\right)' \right)^{-1} \right)$$

and the second term $\pi\left(\gamma^*|Y_n,\Omega^{-1*}\right)$ is available directly as

$$\pi(\gamma^*|Y_n,\Omega^{-1*}) = \mathcal{N}_{d+p}\left(\gamma^*|\hat{\gamma}^*, G_n^*\right)$$

where

$$\hat{\gamma}^* = G_n^*\left(G_0^{-1}\gamma_0 + \sum_{t=1}^{n} X_t'\Omega^{-1*}y_t\right)$$
$$G_n^* = \left(G_0^{-1} + \sum_{t=1}^{n} X_t'\Omega^{-1*}X_t\right)^{-1}$$

6 Prediction

Suppose we are interested in predicting the cross-section of returns y_{n+1} at time $(n+1)$. The Bayes prediction density of these returns, conditioned on the data Y_{n+1} and the factors f_{n+1}, is given by

$$p(y_{n+1}|Y_n, f_{n+1}) = \int_{\gamma,\Omega^{-1}} \mathcal{N}_d\left(y_{n+1}|X_{n+1}\gamma,\Omega\right) d\pi\left(\gamma,\Omega^{-1}|Y_n\right)$$

which is estimated by the ergodic Monte Carlo average

$$p(y_{n+1}|Y_n, f_{n+1}) = \frac{1}{G}\sum_{g=1}^{G}\mathcal{N}_d\left(y_{n+1}|X_{n+1}\gamma^{(g)},\Omega^{(g)}\right)$$

with the MCMC draws $\left\{\gamma^{(g)},\Omega^{(g)}\right\}$ from the posterior distribution.

7 Student-t errors

Suppose now that the errors are distributed as multivariate-t

$$\varepsilon_t \sim t_{d,\nu}\left(0,\Omega\right)$$

so that

$$E(\varepsilon_t) = 0\,,\ \nu > 1$$

$$Var\left(\varepsilon_t\right) = \frac{\nu}{\nu-2}\Omega,\ \nu > 2$$

The analysis of this model utilizes the hierarchical reprentation

$$\varepsilon_t|\lambda_t \sim N\left(0, \lambda_t^{-1}\Omega\right)$$
$$\lambda_t \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$$

which means that conditioned on $(\nu, \{\lambda_t\})$, the results presented for the Gaussian model can be applied with minor modifications. The MCMC sampling is completed with the sampling of $(\nu, \{\lambda_t\})$.

Following Albert and Chib (1993), let us assume that the support of ν is the set of values $\{\nu_j\}_{j=1}^{J}$, for example, $\{4, 6, 8, 10, 12, 14, 16\}$ and that a priori

$$\Pr\left(\nu = \nu_j\right) = q_j$$

Then, simple calculations show that

$$\gamma|Y_n, \Omega^{-1}, \nu, \{\lambda_t\} \sim \mathcal{N}_{d+p}\left(\hat{\gamma}_\lambda, G_{n,\lambda}\right)$$

where

$$\hat{\gamma}_\lambda = G_{n,\lambda}\left(G_0^{-1}\gamma_0 + \sum_{t=1}^{n}\lambda_t X_t'\Omega^{-1}y_t\right)$$
$$G_{n,\lambda} = \left(G_0^{-1} + \sum_{t=1}^{n}\lambda_t X_t'\Omega^{-1}X_t\right)^{-1}$$

and

$$\Omega^{-1}|Y_n, \gamma, \nu, \{\lambda_t\} \sim \mathcal{W}_d\left(\rho_0 + n, \left(R_0^{-1} + \sum_{t=1}^{n}\lambda_t\left(y_t - X_t\gamma\right)\left(y_t - X_t\gamma\right)'\right)^{-1}\right)$$

Moreover,

$$\Pr\left(\nu = \nu_j|Y_n, \gamma, \Omega^{-1}\right) \propto q_j\prod_{t=1}^{n} t_{d,\nu_j}\left(y_t|X_t\gamma, \Omega\right)$$

and

$$\lambda_t|Y_n, \gamma, \nu \sim G\left(\frac{\nu+d}{2}, \frac{\nu + \left(y_t - X_t\gamma\right)'\left(y_t - X_t\gamma\right)}{2}\right)$$

One sweep of the MCMC sampling is completed by sampling these four distributions in this order.

7.1 Marginal likelihood

The Chib (1995) method can again be applied to find the log marginal likelihood as

$$\log \Pr(\nu^*) + \log \pi(\gamma^*) + \log \pi\left(\Omega^{-1*}\right) + \sum_{t=1}^{n} \log t_{d,\nu^*}(y_t|X_t\gamma^*, \Omega^*) - \log \pi\left(\nu^*, \gamma^*, \Omega^{-1*}|Y_n\right)$$

where ν^* is the posterior mode (which is easily computed from the sampled values), the last term is calculated as

$$\Pr(\nu^*|Y_n) \times \pi\left(\Omega^{-1*}|Y_n, \nu^*\right) \times \pi\left(\gamma^*|Y_n, \Omega^{-1*}, \nu^*\right)$$

in which the first term is obtained from the posterior frequency distribution of ν, the second term is obtained from a reduced run in which ν is fixed at ν^* and the remaining three distributions are sampled and the draws

$$\left\{\gamma^{(g)}, \lambda_t^{(g)}\right\}_{g=1}^{G}$$

from this reduced MCMC run are used to calculate $\pi(\Omega^{-1*}|Y_n, \nu^*)$ as

$$\frac{1}{G}\sum_{g=1}^{G} \mathcal{W}_d\left(\Omega^{-1*}|\rho_0 + n, \left(R_0^{-1} + \sum_{t=1}^{n} \lambda_t^{(g)}\left(y_t - X_t\gamma^{(g)}\right)\left(y_t - X_t\gamma^{(g)}\right)'\right)^{-1}\right)$$

and the final term $\pi(\gamma^*|Y_n, \Omega^{-1*})$ is obtained from a second reduced run in which ν is fixed at ν^* and Ω^{-1} is fixed at Ω^{-1*} and the draws

$$\left\{\lambda_t^{(g)}\right\}$$

from this reduced run are used to give

$$\pi(\gamma^*|Y_n, \Omega^{-1*}, \nu^*) = \frac{1}{G}\sum_{g=1}^{G} \mathcal{N}_{d+p}\left(\gamma^*|\hat{\gamma}^*_{\lambda^{(g)}}, G^*_{n,\lambda^{(g)}}\right)$$

where

$$\hat{\gamma}^*_{\lambda^{(g)}} = G^*_{n,\lambda^{(g)}}\left(G_0^{-1}\gamma_0 + \sum_{t=1}^{n} \lambda_t^{(g)} X_t'\Omega^{-1*} y_t\right)$$

$$G^*_{n,\lambda^{(g)}} = \left(G_0^{-1} + \sum_{t=1}^{n} \lambda_t^{(g)} X_t'\Omega^{-1*} X_t\right)^{-1}$$