1 Abstract

In this paper we revisit one of the classic questions in empirical finance: which factors in combination are useful for explaining the time-series and cross-section behavior of equity and portfolio returns. The contribution of this paper is to evaluate this question from a Bayesian perspective recognizing that proper evaluation of the worth of a factors has to be in the context of models with and without other factors. Answering such a question therefore requires the consideration of all possible subset models, where the subset models are essentially special cases of a seemingly unrelated regression (SUR) model with the same subset of factors on the right-hand side but with different asset-specific factor coefficients, and a jointly distributed vector error with an unknown precision matrix. In the case of a leading set of 12 factors (along with the intercept which can be present or absent in each possible case), this leads to $2^1 3$ possible SUR models, which along with 6 different assumptions about the error distribution, amounts to the comparison of 49152 SUR models. We carefully compare these models with the help of objectively constructed priors (one for each of our models) using a training sample, and with the calculation of marginal likelihoods (computed by the method of (Chib, 1995)). Marginal likelihoods are proportional to the posterior probability of each model and have recognized finite sample and asymptotic properties. In particular, marginal likelihoods include a penalty for complexity (in other words models with more factors do not necessarily gather greater support) and asymptotically pick either the true model (if it is in the class being considered) or find the model that is closet to the true model (if it is not in the class being considered). Our comparison focuses on test assets from the current literature on this topic: a collection of 10 asset portfolios and a collection of 10 equities. Our results show ..

2 Introduction

3 Model and Framework

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

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

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

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

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 $p = D(K+1)$.

3.1 Gibbs Sampler with Normal Errors

3.1.1 Prior and Posterior Distribution

To complete the model we specify the following prior distribution

$$\pi(\boldsymbol{\gamma}, \Omega^{-1}) = \mathcal{N}_p(\boldsymbol{\gamma} | \boldsymbol{\gamma}_{\mathbf{0}}, G_0) \, \mathcal{W}_D\left(\Omega^{-1} | \rho_0, R_0\right)$$

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

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

3.1.2 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.

3.1.3 Calculating the Marginal likelihood

We calculate the marginal likelihood using the method of Chib (1995). Let θ denote the full collection of parameters. By Bayes' Rule

$$\pi\left(\theta|Y_T\right) = \frac{\pi(\theta) f(Y_T|\theta)}{f(Y_T)}$$

where $f(Y_T)$ is the marginal likelihood, aka the marginal data density, aka the evidence. This identity holds true for *any* value of θ. In particular it holds at the posterior mean θ^*. Solving for $f(Y_T)$ and evaluating the result at θ^*, we have

$$f\left(Y_T\right) = \frac{\pi(\theta^*) f(Y_T|\theta^*)}{\pi(\theta^*|Y_T)}$$

Thus, we can express the *log* marginal likelihood as

$$\log f(Y_T) = \log \pi(\theta^*) + \log f\left(Y_T|\theta^*\right) - \log \pi\left(\theta^*|Y_T\right)$$

Specializing this to the SUR model considered above,

$$\log f(Y_T) = \log \pi(\boldsymbol{\gamma}^*) + \log \pi\left(\Omega^{-1*}\right) + \log f\left(Y_T|\boldsymbol{\gamma}^*, \Omega^{-1*}\right) - \log \pi\left(\boldsymbol{\gamma}^*, \Omega^{-1*}|Y_T\right)$$

since our priors over $\boldsymbol{\gamma}$ and Ω^{-1} are independent. The Chib (1995) method approximates $\log f(Y_T)$ by evaluating each of the terms on the right-hand-side of the preceding expression using the output of the Gibbs sampler.

The Contribution of the Prior Evaluating the first two terms, $\log \pi(\boldsymbol{\gamma}^*)$ and $\log \pi\left(\Omega^{-1*}\right)$, is easy: these are simply the priors for $\boldsymbol{\gamma}$ and Ω^{-1} evaluated at the posterior means. We take the sample average of the Gibbs draws to approximate $\boldsymbol{\gamma}^*$ and Ω^{-1*} and evaluate the Normal and Wishart distributions at these points, with parameters given by the prior:

$$\begin{aligned} \pi\left(\boldsymbol{\gamma}^*\right) &= \mathcal{N}_p\left(\boldsymbol{\gamma}^*|\boldsymbol{\gamma}_0, G_0\right) \\ \pi\left(\Omega^{-1*}\right) &= \mathcal{W}_D\left(\Omega^{-1*}|\rho_0, R_0\right) \end{aligned}$$

The Contribution of the Likelihood Above we assumed a normal distribution for the regression errors, specifically, $\boldsymbol{\varepsilon}_t|\mathbf{x}_t \sim$ iid $\mathcal{N}_D(0, \Omega)$. From the regression specification it follows that $\mathbf{y}_t \sim$ iid $\mathcal{N}_D\left(X_t\boldsymbol{\gamma}, \Omega\right)$ and thus the log likelihood evaluated at the posterior mean is

$$\log f(Y_T|\boldsymbol{\gamma}^*, \Omega^{-1*}) = \sum_{t=1}^{T} \log \mathcal{N}_D\left(\mathbf{y}_t|X_t\boldsymbol{\gamma}^*, \Omega^{-1*}\right)$$

parameterized in terms of the precision matrix rather than the covariance matrix. Equivalently, but more conveniently, we may write

$$\log f(Y_T|\boldsymbol{\gamma}^*, \Omega^{-1*}) = \sum_{t=1}^{T} \log \mathcal{N}_D\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^*|\mathbf{0}, \Omega^{-1*}\right)$$

The advantage of this version of the likelihood is that the parameters of the normal density are constant over t, allowing us to exploit the efficient algorithm for repeatedly

evaluating a MV normal density with fixed parameters, described above. Note that we can simultaneously calculate all of the arguments for the normal density as follows:

$$(\widetilde{Y} - \widetilde{X}\Gamma^*)' = (\boldsymbol{\varepsilon}^*)' = \left[\begin{array}{ccc} \boldsymbol{\varepsilon}_1^* & \ldots & \boldsymbol{\varepsilon}_T^* \end{array} \right]$$

where $\boldsymbol{\varepsilon}_t^* = \mathbf{y}_t - X_t\boldsymbol{\gamma}^*$ and $\Gamma^* = (\boldsymbol{\gamma}_1^*, \ldots, \boldsymbol{\gamma}_D^*)$.

The Contribution of the Posterior To evaluate the third term, we factorize the joint posterior as the product of a conditional and marginal, namely:

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

so that we have

$$\log \pi\left(\boldsymbol{\gamma}^*, \Omega^{-1*}|Y_T\right) = \log \pi\left(\boldsymbol{\gamma}^*|\Omega^{-1*}, Y_T\right) + \log \pi\left(\Omega^{-1*}|Y_T\right)$$

Because we have *analytical expressions* for the conditional posteriors in this model we can evaluate the first term in the product immediately. We have $\boldsymbol{\gamma}|\Omega^{-1} \sim \mathcal{N}_p\left(\boldsymbol{\gamma}|\bar{\boldsymbol{\gamma}}, G_T\right)$ where G_T and $\bar{\gamma}$ depend only on the prior, the data, and Ω^{-1}. To perform the required calculation, we simply evaluate the normal density at $\boldsymbol{\gamma}^*$ and evaluate G_T and $\bar{\gamma}$ at Ω^{-1*}, that is:

$$\pi\left(\boldsymbol{\gamma}^*|\Omega^{-1*}, Y_T\right) = \mathcal{N}_p\left(\boldsymbol{\gamma}^*|\bar{\gamma}^*, G_T^{-1*}\right)$$

where

$$\begin{aligned} G_T^{-1*} &= \left[G_0^{-1} + \sum_{t=1}^{T} X_t'\Omega^{-1*}X_t\right] = \left[G_0^{-1} + \Omega^{-1*} \otimes \widetilde{X}'\widetilde{X}\right] \\ \bar{\boldsymbol{\gamma}}^* &= G_T^*\left[G_0^{-1}\boldsymbol{\gamma}_0 + \sum_{t=1}^{T} X_t'\Omega^{-1*}\mathbf{y}_t\right] = \text{solve}\left[G_T^{-1*},\ G_0^{-1}\gamma_0 + \text{vec}\left(\widetilde{X}'\widetilde{Y}\Omega^{-1*}\right)\right] \end{aligned}$$

The evaluation of the second term in the product that gives the contribution of the posterior to the marginal likelihood is a bit more involved. We write

$$\begin{aligned} \pi\left(\Omega^{-1*}|Y_T\right) &= \int \pi\left(\boldsymbol{\gamma}, \Omega^{-1*}|Y_T\right)\ d\boldsymbol{\gamma} \\ &= \int \pi\left(\Omega^{-1*}|\boldsymbol{\gamma}, Y_T\right)\pi\left(\boldsymbol{\gamma}|Y_T\right)\ d\boldsymbol{\gamma} \end{aligned}$$

and approximate the second integral using the draws from the Gibbs sampler:

$$\begin{aligned} \pi\left(\Omega^{-1*}|Y_T\right) &\approx \frac{1}{G}\sum_{g=1}^{G} \pi\left(\Omega^{-1*}|\boldsymbol{\gamma}^{(g)}, Y_T\right) \\ &= \frac{1}{G}\sum_{g=1}^{G} \mathcal{W}_D\left(\Omega^{-1*}\left|\rho_0 + T, R_T^{(g)}\right.\right) \end{aligned}$$

where

$$\begin{aligned} R_T^{(g)} &= \left[R_0^{-1} + \sum_{t=1}^{T} \left(\mathbf{y}_t - X_t \boldsymbol{\gamma}^{(g)} \right) \left(\mathbf{y}_t - X_t \boldsymbol{\gamma}^{(g)} \right)' \right]^{-1} \\ &= \left[R_0^{-1} + \left(\widetilde{Y} - \widetilde{X} \Gamma^{(g)} \right)' \left(\widetilde{Y} - \widetilde{X} \Gamma^{(g)} \right) \right]^{-1} \end{aligned}$$

3.2 Gibbs Sampler with Student-t Errors

3.2.1 A Hierarchical Representation

Suppose now that the errors follow a multivariate Student-t distribution rather than a normal distribution:

$$\boldsymbol{\varepsilon}_t \sim t_{D,\nu} (0, \Omega)$$

where ν denotes the degrees of freedom of the distribution, the location parameter is zero and the scale matrix is Ω. If $\nu > 1$ then $E(\boldsymbol{\varepsilon}) = 0$. If $\nu > 2$ then $Var(\boldsymbol{\varepsilon}) = \nu\sigma/(n-2)$. Replacing the normal likelihood from above with the Student-t likelihood, however, breaks the conditional conjugacy that we exploited above to construct an MCMC algorithm based on the Gibbs sampler. The solution to this problem is to work with a hierarchical representation in which the Student-t likelihood is introduced as a scale mixtrure of normal distributions ((Chib & Greenberg, 1995)), in particular

$$\begin{aligned} \boldsymbol{\varepsilon}_t | \lambda_t &\sim N \left(0, \lambda_t^{-1} \Omega \right) \\ \lambda_t &\sim G \left(\frac{\nu}{2}, \frac{\nu}{2} \right) \end{aligned}$$

where $G(\alpha, \beta)$ denotes the Gamma distribution with shape parameter α and rate parameter β. (See below for more discussion on the parameterization of the gamma distribution.) Using this representation, after conditioning on $(\nu, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_T)'$, we are essentially back in the familiar normal case from above. In any inference that we carry out, as well as in the calculation of the marginal likelihood, we will marginalize over $\boldsymbol{\lambda}$ by simply ignoring these draws. We will continue to place a normal prior on $\boldsymbol{\gamma}$ and a Wishart prior on Ω^{-1}, exactly as we did when working with the normal likelihood above.

3.2.2 The Sampler

The sampler proceeds by fixing the degrees of freedom parameter ν. If ν is to be chosen from the data, this can be accomplished using the marginal likelihood, as described below. Holding ν fixed, the full set of conditional posteriors is as follows:

Regression Coefficients: $\boldsymbol{\gamma}|\Omega^{-1}, Y_T \sim \mathcal{N}_p\left(\bar{\boldsymbol{\gamma}}_\lambda, G_{T,\lambda}\right)$

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

Inverse Scale Matrix: $\Omega^{-1}|Y_T \sim \mathcal{W}_D\left(\rho_0 + T, R_{T,\lambda}\right)$

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

Auxiliary Parameter: $\lambda_t|\boldsymbol{\gamma}, \nu, Y_T \sim G\left(\frac{\nu + D}{2}, \frac{\nu + \boldsymbol{\varepsilon}_t'\boldsymbol{\varepsilon}_t}{2}\right)$, $\boldsymbol{\varepsilon}_t = \mathbf{y}_t - X_t\boldsymbol{\gamma}$

To implement the Gibbs sampler, we simply need to draw sequentially from these distributions, in the order given above. We will require, however, starting values for both Ω^{-1} and each of the λ_t parameters. A reasonable starting value for λ_t is one, which makes the initial draws for the regression coefficients and the inverse scale matrix the same as if were work withing with the normal model.

3.2.3 Marginal Likelihood for Student-t Model

First we describe the method for an arbitrary three-block Gibbs sampler. We will use the idea of a "reduced run" introduced here in our calculation of the marginal likelihood of the Student-t model.

General Three-Block Algorithm Re-arranging Bayes' Rule we have the identity

$$f(y) = \frac{f(y|\theta_1^*, \theta_2^*, \theta_3^*)\pi\left(\theta_1^*, \theta_2^*, \theta_3^*\right)}{\pi\left(\theta_1^*, \theta_2^*, \theta_3^*|y\right)}$$

for any specified values $(\theta_1^*, \theta_2^*, \theta_3^*)$ of the parameters. In particular this holds at the *posterior mean* which is where we will evaluate the expression. Hence, the *log* marginal likelihood is given by

$$\log f(y) = \log \pi\left(\theta_1^*, \theta_2^*, \theta_3^*\right) + \log f(y|\theta_1^*, \theta_2^*, \theta_3^*) - \pi\left(\theta_1^*, \theta_2^*, \theta_3^*|y\right)$$

The first two terms are easy: we simply evaluate the log of the prior and likelihood at the posterior mean for the three parameters. The third one, however, is more complicated. To calculate it we use the factorization:

$$\pi\left(\theta_1^*, \theta_2^*, \theta_3^*|y\right) = \pi(\theta_1^*|y)\pi\left(\theta_2^*|\theta_1^*, y\right)\pi\left(\theta_3^*|\theta_2^*, \theta_1^*, y\right)$$

This leaves us with three a product of new terms that we need to calculate. The last of the terms in the product, $\pi(\theta_3^*|\theta_2^*, \theta_1^*, y)$ is immediately available: this conditional density is known since we used it as a step in the Gibbs sampler. All we need to do is substitute in the appropriate values for θ_2^*, θ_1^* and evaluate the density at θ_3^*. To calculate the first term in the product we need to marginalize over θ_2, θ_3. A Monte-Carlo approximation to the appropriate integral can be computed directly from the Gibbs sampler output:

$$\widehat{\pi}\left(\theta_1^*|y\right) = \frac{1}{G}\sum_{g=1}^{G} \pi\left(\theta_1^*|\theta_2^{(g)}, \theta_3^{(g)}, y\right)$$

To do this we rely on the fact that $\pi(\theta_1|\theta_2, \theta_3, y)$ is a known density – we use it in the Gibbs sampler. The middle term in the product is the most difficult one to calculate. To begin, notice that

$$\pi\left(\theta_2^*|\theta_1^*, y\right) = \int \pi(\theta_2^*, \theta_3|\theta_1^*, y)\, d\theta_3 = \int \pi(\theta_2^*|\theta_1^*, \theta_3^*, y)\pi(\theta_3|\theta_1^*, y)\, d\theta_3$$

The idea is to construct a Monte-Carlo approximation of the integral on the right-hand-side of the preceding expression. The approximation we use is

$$\hat{\pi}\left(\theta_2^*|\theta_1^*, y\right) = \frac{1}{G}\sum_{g=1}^{G} \pi(\theta_2^*|\theta_1^*, \theta_3^{(g)}, y)$$

but the draws $\left\{\theta_3^{(g)}\right\}$ come *not* from the original run of the Gibbs sampler but from a so-called "reduced run" in which we sample $\theta_2^{(g)}$ and $\theta_3^{(g)}$ from $\pi(\theta_2|\theta_1^*, \theta_3, y)$ and $\pi(\theta_3|\theta_1^*, \theta_2, y)$. In other words, the reduced run holds θ_1 *fixed* at θ_1^*, the posterior mean calculated from the draws of the *usual* Gibbs sampler. We can carry out the reduced run using the exact same algorithm as we use for the full Gibbs sampler: we just need to keep θ_1^* fixed and make sure that we store the draws $\theta_3^{(g)}$ that we will need to calculate $\hat{\pi}(\theta_2^*|\theta_1^*, y)$.

Calculations for the Student-t Model For the SUR model we have

$$\log f(Y_T) = \log \pi(\boldsymbol{\gamma}^*) + \log \pi(\Omega^{-1*}) + \log f(Y_T|\boldsymbol{\gamma}^*, \Omega^{-1*}) - \log \pi(\boldsymbol{\gamma}^*, \Omega^{-1*}|Y_T)$$

exactly as in the case with normal errors. This is because we still use the *same prior*, under which $\boldsymbol{\gamma}$ and Ω^{-1} are independent. Moreover, note that λ_t does not make a direct appearance. This is because we use it only as a way of obtaining conditional conjugacy: it is not a parameter over which we place a prior. The first two terms of this expression give the contribution of the prior to the marginal likelihood. These are computed in exactly the same way as above in the normal case. The third term gives the contribution

of the likelihood. In the Student-t model we have $\varepsilon_t|\mathbf{x}_t \sim$ iid $t_{D,\nu}(0,\Omega)$ and hence, from the regression specification, it follows that $\mathbf{y}_t \sim$ iid $t_{D,\nu}(X_t\boldsymbol{\gamma},\Omega)$. Hence, evaluating the log-likelihood at the posterior mean gives

$$\log f(Y_T|\boldsymbol{\gamma}^*,\Omega^{-1*}) = \sum_{t=1}^{T} \log t_{D,\nu}\left(\mathbf{y}_t|X_t\boldsymbol{\gamma}^*,\Omega^{-1*}\right)$$

where we parameterize in terms of Ω^{-1} instead of Ω since this is how our C++ function for the Student-t density is specified. From here, the calculation is identical to the case for the normal model except with a Student-t density in place of a normal.[1]

It is the *final* term in the log marginal likelihood expression, the contribution of the posterior, whose computation is substantially different for the case of the Student-t model. We factorize the contribution of the posterio according to:

$$\log \pi(\boldsymbol{\gamma}^*,\Omega^{-1*}|Y_T) = \log \pi\left(\Omega^{-1*}|Y_T\right) + \log \pi\left(\boldsymbol{\gamma}^*|\Omega^{-1*},Y_T\right)$$

To approximate the $\log \pi(\Omega^{-1*}|Y_T)$ term we evaluate the density $\pi(\Omega^{-1}|\boldsymbol{\gamma},\boldsymbol{\lambda},Y_T)$ at Ω^{-1*} and marginalize over the original Gibbs sampler draws $\left\{\boldsymbol{\gamma}^{(g)},\boldsymbol{\lambda}^{(g)}\right\}$, that is

$$\begin{aligned}
\widehat{\pi}\left(\Omega^{-1*}|Y_T\right) &= \frac{1}{G}\sum_{g=1}^{G} \pi\left(\Omega^{-1*}|\boldsymbol{\gamma}^{(g)},\boldsymbol{\lambda}^{(g)},Y_T\right) \\
&= \frac{1}{G}\sum_{g=1}^{G} \mathcal{W}_D\left(\Omega^{-1*}\left|\rho_0+T,R^{(g)}_{T,\lambda^{(g)}}\right.\right)
\end{aligned}$$

where

$$\begin{aligned}
R^{(g)}_{T,\lambda^{(g)}} &= \left[R_0^{-1} + \sum_{t=1}^{T} \lambda_t\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(g)}\right)\left(\mathbf{y}_t - X_t\boldsymbol{\gamma}^{(g)}\right)'\right]^{-1} \\
&= \left[R_0^{-1} + \left(\widetilde{Y} - \widetilde{X}\Gamma^{(g)}\right)'\Lambda^{(g)}\left(\widetilde{Y} - \widetilde{X}\Gamma^{(g)}\right)\right]^{-1}
\end{aligned}$$

All that remains is to approximate $\pi\left(\boldsymbol{\gamma}^*|\Omega^{-1*},Y_T\right)$. To do this, we will need to use a reduced run similar to the one used in the calculations for the general three-block Gibbs-sampler described above. We know $\pi\left(\boldsymbol{\gamma}|\Omega^{-1},\boldsymbol{\lambda},Y_T\right)$ so we will evaluate this expression at $\boldsymbol{\gamma}^*$ and Ω^{-1*} and integrate out $\boldsymbol{\lambda}$ using a set of draws $\left\{\boldsymbol{\lambda}^{(j)}\right\}$ that were generated holding

[1] In particular, it's still convenient to work with the likelihood in terms of ε_t rather than $\mathbf{y}_t$ and we have a way to simultaneously evaluate the likelihood contribution for each observation using our C++ routine for the Student-t density. See above for more details.

Ω^{-1} *fixed* at Ω^{-1*}. Accordingly, we use

$$\begin{aligned}\widehat{\pi}(\gamma^*|\Omega^{-1*}, Y_T) &= \frac{1}{J}\sum_{j=1}^{J}\pi\left(\boldsymbol{\gamma}^*|\Omega^{-1*}, \boldsymbol{\lambda}^{(j)}, Y_T\right) \\ &= \frac{1}{J}\sum_{j=1}^{J}\mathcal{N}_p\left(\boldsymbol{\gamma}^*|\bar{\boldsymbol{\gamma}}^*_{\lambda^{(g)}}, G^*_{n,\lambda^{(j)}}\right)\end{aligned}$$

where

$$\begin{aligned}\bar{\boldsymbol{\gamma}}^*_{\lambda^{(j)}} &= G^*_{T,\lambda^{(j)}}\left(G_0^{-1}\boldsymbol{\gamma}_0 + \sum_{t=1}^{T}\lambda_t^{(j)}X_t'\Omega^{-1*}\mathbf{y}_t\right) \\ G^*_{T,\lambda^{(j)}} &= \left(G_0^{-1} + \sum_{t=1}^{T}\lambda_t^{(j)}X_t'\Omega^{-1*}X_t\right)^{-1}\end{aligned}$$

3.3 Priors

We use a small part of our sample to form reasonable priors. The procedure consists of two stages:

1. Estimate the model for the training sample
2. Estimate the model for the remainder of the sample using the draws from the first step to form a prior

Both for Student-t and normal errors we specify two prior distributions: for the coefficients $\gamma \sim \mathcal{N}_p\left(\boldsymbol{\gamma}|\boldsymbol{\gamma_0}, G_0\right)$ and for the precision matrix $\Omega^{-1} \sim \mathcal{W}_D\left(\Omega^{-1}|\rho_0, R_0\right)$. Parameters for the two distributions are specified differently for each stage.

3.3.1 Stage 1: Training Sample

Parameters of the prior for the regression coefficients are set as follows:

$$\begin{aligned}\gamma_0 &= 0 \\ G_0 &= C_1^2 I_p\end{aligned}$$

The prior distribution of the coefficient vector is centered around zero. The covariance matrix is assumed to be diagonal.
Precision matrix is assumed to follow the inverse Wishart distribution with the following parameters:

$$\begin{aligned}\rho_0 &= d + C_2 \\ R_0 &= \frac{1}{C_3^2(\rho_0 - d - 1)}I_d\end{aligned}$$

The mean value of Ω implied by the prior is a diagonal matrix $C_3^2 I_d$.
For the rest of the paper we set $C_1 = 2$, $C_2 = 6$ and $C_3 = 0.05$.

3.3.2 Stage 2

Here we use the draws based on the training sample to construct priors. Denote posterior means of draws of γ as $\overline{\gamma}$ and the sample covariance matrix as $\widehat{G}$. We also calculate a posterior mean of the Ω^{-1} draws: $\overline{\Omega}^{-1}$.
The regression coefficients prior is:

$$\begin{aligned} \gamma_0 &= \overline{\gamma} \\ G_0 &= C_4^2 \widehat{G} \end{aligned}$$

The prior is centered around the posterior mean of the first step draws. The standard deviation is based on the sample standard deviation of the first stage draws adjusted by the factor of C_4 to reflect uncertainty.
The prior of the precision matrix is set:

$$\begin{aligned} \rho_0 &= d + C_5 \\ R_0 &= \frac{1}{\rho} \overline{\Omega}^{-1} \end{aligned}$$

The prior is constructed to set the mean of the precision matrix Ω^{-1} equal to the posterior mean based on the training sample. One way of widening the prior would be to decrease number of degrees of freedom ρ_0 by adjusting the value of C_5.
For all the applications we use $C_4 = 3$ and $C_5 = 6$.

4 Motivational Example

In order to motivate our research we simulate asset returns and demonstrate that the true factors are selected as the result of the procedure described above.
As asset returns we take 10 value-weighted Fama-French industry portfolios. The returns are assumed to be follow the famous Fama-French 3 factor structure without intercept (Mkt.RF, HML and SMB). The errors follow Student-t distribution with 2.5 degrees of freedom. The simulation is based on posterior means obtained when fitting the model to the real data. Other parameters are the same as in the original sample.

We assume that the researcher does not know the true distribution. Considered distributions include normal and Student-t with 4, 6, 8, 10 and 12 degrees of freedom. The pool of candidate models includes all combinations of Fama-French 5 factors(Mkt.RF, HML, SMB, RMW and CMA), a constant and a non-factor asset - Microsoft stock (MSFT). We fit in total $6 \times 2^7 = 768$ model. The simulation is based on the sample range Apr 1986 - Dec 2014 (345 observations). The training sample includes observations Apr 1986 - Dec 1990 (57 observations).
The simulation setup is described below:

1. Fit the Fama-French 3 factor model without an intercept to 10 value-weighted industry portfolios using the full sample. The errors are assumed to follow Student-t distribution with 4 degrees of freedom.

2. Simulate a dataset assuming the true parameters γ and Ω^{-1} to be equal to the posterior means:

   - Simulate errors:
   $$\boldsymbol{\varepsilon}_t^s \sim t_{10,2.5}(0, \Omega)$$
   - Simulate returns using values of Fama-French 3 factors observed in the data:
   $$\mathbf{y}_t^s = X_t \boldsymbol{\gamma} + \boldsymbol{\varepsilon}_t^s$$

3. Estimate all candidate models and evaluate the likelihood. Run the usual two step estimation procedure using the training sample to construct priors for each model.

		NoDur	Durbl	Manuf	Enrgy	HiTec	Telcm	Shops	Hlth	Utils
Data	Apr 1986 - Dec 2014	0.0080	0.0056	0.0078	0.0080	0.0074	0.0062	0.0073	0.0084	0.0061
	Apr 1986 - Dec 1990	0.0099	-0.0033	0.0034	0.0092	-0.0037	0.0084	0.0036	0.0095	0.0033
	Jan 1991 - Dec 2014	0.0076	0.0073	0.0086	0.0078	0.0096	0.0057	0.0080	0.0082	0.0066
Simulated	Apr 1986 - Dec 2014	0.0035	0.0104	0.0063	0.0036	0.0062	0.0089	0.0045	0.0052	0.0043
	Apr 1986 - Dec 1990	-0.0016	0.0039	0.0020	0.0005	0.0032	0.0106	-0.0016	0.0025	0.0047
	Jan 1991 - Dec 2014	0.0045	0.0117	0.0071	0.0042	0.0067	0.0086	0.0057	0.0057	0.0043

Table 1: Average Portfolio Returns for Simulated and Real Data

The best model selects the true factors. Even though the true error distribution (Student-t with 2.5 degrees of freedom) was not considered, the distribution of the best model (Student-t with 4 degrees of freedom) is the closest to the truth.
As can be seen from the results, the procedure correctly identified the factors. Models including non-relevant factors or the non-factor (Microsoft stock) are significantly worse on the log scale.

	Mkt.RF	SMB	HML	RMW	CMA	MSFT
Apr 1986 - Dec 2014	0.0062	0.0011	0.0023	0.0036	0.0033	0.0102
Apr 1986 - Dec 1990	0.0028	-0.0068	0.0003	0.0047	0.0054	0.0327
Jan 1991 - Dec 2014	0.0069	0.0027	0.0027	0.0034	0.0029	0.0058

Table 2: Average Factor Returns

Model	DF	log margLike
Mkt.RF + SMB + HML	4	7187.32
Mkt.RF + SMB + HML + CMA	4	7177.05
Mkt.RF + SMB + HML + MSFT	4	7170.99
Mkt.RF + SMB + HML + RMW	4	7170.24
Mkt.RF + SMB + HML	6	7169.59
Mkt.RF + SMB + HML + CMA + MSFT	4	7161.34
constant + Mkt.RF + SMB + HML	4	7160.41
Mkt.RF + SMB + HML + CMA	6	7159.73
constant + Mkt.RF + SMB + HML + CMA	4	7157.97
Mkt.RF + SMB + HML + RMW + MSFT	4	7152.92
Mkt.RF + SMB + HML	8	7152.24
constant + Mkt.RF + SMB + HML + MSFT	4	7151.56
Mkt.RF + SMB + HML + RMW	6	7151.19
constant + Mkt.RF + SMB + HML	6	7149.93
constant + Mkt.RF + SMB + HML + RMW	4	7147.47
Mkt.RF + SMB + HML + MSFT	6	7147.06
Mkt.RF + SMB + HML + RMW + CMA	4	7146.92
Mkt.RF + SMB + HML + CMA + MSFT	6	7146.04
constant + Mkt.RF + SMB + HML + CMA + MSFT	4	7145.13
constant + Mkt.RF + SMB + HML + MSFT	6	7139.5

Table 3: Motivational Simulation: 20 Best Models

5 Application

5.1 Data

We apply our method to 10 value-weighted industry portfolios available at the Kenneth French's website. The 12 candidate factors include:

- Five

6 Conclusion

References

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, *90*(432), 1313–1321.

Chib, S., & Greenberg, E. (1995). Hierarchical analysis of sur models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics*, *68*(2), 339–360.