Bayesian Analysis (2010) **5**, Number 2, pp. 237–262

Shrinkage Regression for Multivariate Inference with Missing Data, and an Application to Portfolio Balancing

Robert B. Gramacy* and Ester Pantaleo†

Abstract. Portfolio balancing requires estimates of covariance between asset returns. Returns data have histories which greatly vary in length, since assets begin public trading at different times. This can lead to a huge amount of missing data—too much for the conventional imputation-based approach. Fortunately, a well-known factorization of the MVN likelihood under the prevailing historical missingness pattern leads to a simple algorithm of OLS regressions that is much more reliable. When there are more assets than returns, however, OLS becomes unstable. Gramacy et al. (2008) showed how classical shrinkage regression may be used instead, thus extending the state of the art to much bigger asset collections, with further accuracy and interpretation advantages. In this paper, we detail a fully Bayesian hierarchical formulation that extends the framework further by allowing for heavy-tailed errors, relaxing the historical missingness assumption, and accounting for estimation risk. We illustrate how this approach compares favorably to the classical one using synthetic data and an investment exercise with real returns. An accompanying R package is on CRAN.

Keywords: multivariate, monotone missing data, data augmentation, ridge regression, double-exponential, heavy tails, factor model, portfolio balancing

1 Introduction

Mean–variance portfolio allocation (e.g., Markowitz 1959) requires the accurate and tractable estimation of the mean return, and the covariance between the returns, of a large number of assets. Assets become publicly tradeable at different times, so their return histories can greatly vary in length. Aside from a few "gaps", the histories of assets which are publicly tradeable at purchase time will exhibit a *monotone missingness pattern*. For example, Figure 1 shows the monthly return availability for 1,200-odd stocks on NYSE & AMEX over 29 years, some with as little as 1 years worth of data. The assets along the columns have been sorted by the number of missing entries. Besides some "gaps" the boundary between the dark and light regions is monotone.

Generally speaking, inference in the presence of missing data is notoriously difficult, usually requiring hill-climbing iterative techniques like *expectation maximization* (Little and Rubin 2002; Schafer 1997)) which rapidly lose stability as the level of missingness increases. The Bayesian alternative of *data augmentation* is similarly unsatisfac-

*Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK, mailto:bobby@statslab.cam.ac.uk

†Dipartimento di Fisica, Università di Bari, Bari, Italy, mailto:ester.pantaleo@ba.infn.it

 DOI:10.1214/10-BA602

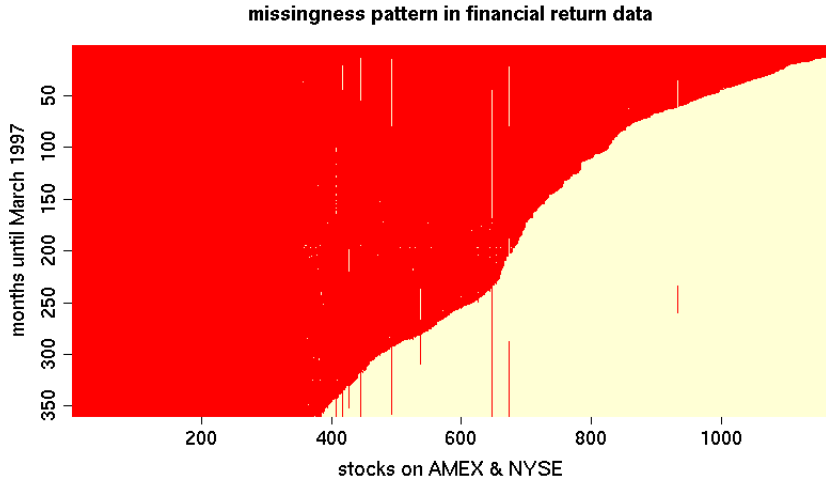

Figure 1: Missingness pattern in stock returns on the NYSE & AMEX. The assets (columns) are put into increasing order of the number of missing entries.

tory. Software packages implementing such algorithms come with prominently displayed warnings of failure when the missingness level is above 15% (see Gramacy et al. 2008).

The nice thing about a monotone missingness pattern is that the likelihood has a convenient factorization which makes inference tractable without imputation. Under a multivariate normal (MVN) assumption, a simple algorithm (Andersen 1957; Stambaugh 1997) of ordinary least squares (OLS) regressions, one for each asset, yields a maximum likelihood estimator (MLE). Unfortunately, there must be fewer stocks than the length of the shortest return history, so that the design matrices of the OLS regressions are of full rank. In particular, you cannot have more stocks than historical returns. Gramacy et al. (2008) showed that by replacing the OLS with "parsimonious regressions", e.g., principal components (PCR), ridge, lasso, etc., the above algorithm can be applied when there are more assets than historical returns. This extended the reach of Stambaugh's (1997) methods from dozens to thousands of assets, accommodating an essentially arbitrary level of historical missingness.

In this paper we shall further extend the above parsimonious methodology in several directions by taking a fully Bayesian approach. Section 2 recalls the monotone decomposition and (MLE/Bayesian) inference algorithm. Section 3 reviews approaches to Bayesian shrinkage regression that are particularly convenient in this context, and which allow model averaging and heavy-tailed errors as minor embellishments. Section 4 details how the benefits of the Bayesian shrinkage posteriors filter through to inference about the mean vector and covariance matrix. It features extensions for data augmentation to deal with "gaps", and allows estimation risk to influence the balanced portfolios—both of which were unavailable previously. In Section 5 we apply our methods in a Monte Carlo investment exercise and show how they compare favorably to the

classical alternatives on real financial returns data. The paper concludes with a brief discussion in Section 6.

The methods that are core to this paper are implemented in a fully documented R (R Development Core Team 2007) package called `monomvn` (Gramacy 2009), which is available for download on the Comprehensive R Archive Network (CRAN).

2 Multivariate normal monotone missing data

We assume that the missingness mechanism is *missing completely at random* (MCAR). In the case of historical asset returns this may be a tenuous assumption, but it is convenient and common (e.g., Stambaugh 1997). We work with a $n \times m$ data matrix $\mathbf{Y}$ that collects the historical returns of the assets. Denote $y_{i,j} = \texttt{NA}$ if the i^{th} sample (historical return) of the j^{th} covariate (asset) is missing; otherwise $y_{i,j} \in \Re$. Formally speaking, the missingness pattern in $\mathbf{Y}$ is said to be *monotone* [e.g., (Schafer 1997, Section 6.5.1) or (Little and Rubin 2002, Section 7.4)] if its columns can be re-arranged so that $y_{i,j} \neq \texttt{NA}$ whenever $y_{i,j+1} \neq \texttt{NA}$.

We assume throughout that they are indeed arranged in this way, so that when we define $n_j = \sum_{i=1}^{n} I_{\{y_{i,j} \neq \texttt{NA}\}}$ as the number of observed entries in column j, for $j = 1, \ldots, m$, we have that $n \equiv n_1$ and $n_j \geq n_{j+1}$. Furthermore, the rows may be arranged according to the same property ($y_{i,j} \neq \texttt{NA}$ whenever $y_{i+1,j} \neq \texttt{NA}$) without loss of generality, so that when we define $\mathbf{y}_j \equiv y_{1:n_j,j}$ we are collecting the entirety of the observed entries in the j^{th} column. The pattern that results is illustrated pictorially in Figure 2. For now we will assume that there are no "gaps". It is also helpful to define $\mathbf{Y}_j \equiv \mathbf{Y}_{0:(j-1)}^{(n_j)}$ as the $n_j \times j$ design matrix

$$\mathbf{Y}_j \equiv \mathbf{Y}_{0:(j-1)}^{(n_j)} = \begin{pmatrix} 1 & y_{1,1} & \cdots & y_{1,(j-1)} \\ 1 & y_{2,1} & \cdots & y_{2,(j-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{n_j,1} & \cdots & y_{n_j,(j-1)} \end{pmatrix}$$

containing an intercept, and the first n_j rows of the first $j-1$ columns of $\mathbf{Y}$; see Figure 2.

Under the monotone pattern the likelihood $f(\mathbf{Y}|\boldsymbol{\theta})$ emits a convenient factorization in terms of an auxiliary parameterization $\boldsymbol{\phi} = \Phi(\boldsymbol{\theta})$. If the rows of $\mathbf{Y}$ are i.i.d. MVN, this factorization leads to an iterative algorithm for inferring the MLE $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$. These assumptions may not be appropriate for financial returns but they are common to keep inference tractable (e.g., Stambaugh 1997; Chan et al. 1999; Jagannathan and Ma 2003).

The algorithm, due originally to Andersen (1957), begins by calculating $\hat{\mu}_1$ and $\hat{\Sigma}_{1,1} \equiv \hat{\sigma}_1^2$ in the usual way: $\hat{\mu}_1 = n_1^{-1} \sum_{i=1}^{n_1} y_{i,1}$ and $\hat{\sigma}_1^2 = n_1^{-1} \sum_{i=1}^{n_1} (y_{i,1} - \hat{\mu}_1)^2$. Then, for $j = 2, \ldots, m$ the MLEs of $\boldsymbol{\theta}_j = (\mu_j, \boldsymbol{\Sigma}_{1:j,j})$, $j = 2, \ldots, m$, can then be obtained via a regression on the complete data in columns $1, \ldots, j-1$, i.e., using the model

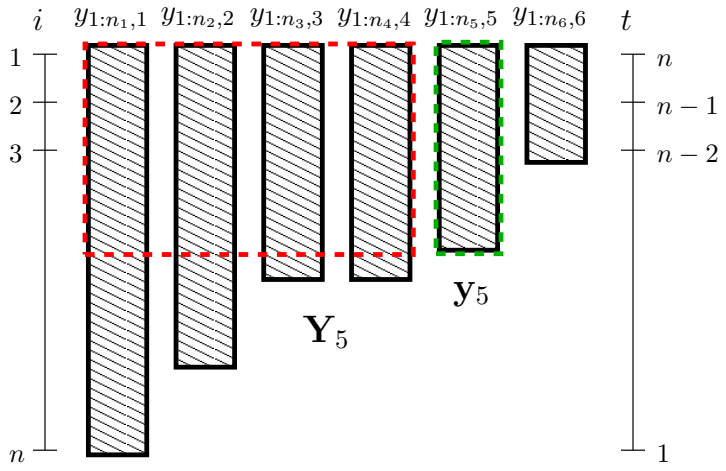

Figure 2: Diagram of a monotone missingness pattern with $m = 6$ covariates, and n completely observed samples in $\mathbf{y}_1 = y_{:,1}$. The design matrix $\mathbf{Y}_5$ (without an intercept term) and the response vector $\mathbf{y}_5$ for the fifth regression involved in maximizing the likelihood of MVN data under a monotone missingness pattern is also shown. Time (t) runs counter to the index i so that the most recent historical return is at time $t = n$, indexed by $i = 1$.

$\mathbf{y}_j = \mathbf{Y}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$, where $\{\epsilon_{i,j}\}_{i=1}^{n_j} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_j^2)$. Here $\boldsymbol{\beta}_j^\top = (\beta_{0,j}, \beta_{1,j}, \ldots, \beta_{(j-1),j})$, and σ_j^2 are the auxiliary parameters $\boldsymbol{\phi}_j$. When $\text{rank}(\mathbf{Y}_j) = j$, and particularly when $n_j > j$, MLEs $\hat{\boldsymbol{\phi}}_j$ may be obtained in the usual way: $\hat{\boldsymbol{\beta}}_j = (\mathbf{Y}_j^\top\mathbf{Y}_j)^{-1}\mathbf{Y}_j^\top\mathbf{y}_j$ and $\hat{\sigma}_j^2 = \frac{1}{n_j}||\mathbf{y}_j - \mathbf{Y}_j\hat{\boldsymbol{\beta}}_j||^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(y_{i,j} - (\mathbf{y}_i^\top)_{1:n_j}\hat{\boldsymbol{\beta}}_j)^2$. The components of $\hat{\boldsymbol{\theta}}_j$ given $\hat{\boldsymbol{\theta}}_{1:(j-1)} = (\hat{\boldsymbol{\mu}}_{1:(j-1)}^\top, \hat{\boldsymbol{\Sigma}}_{1:(j-1),1:(j-1)})$ and $\hat{\boldsymbol{\phi}}_j$ are then

$$\begin{aligned} \hat{\mu}_j &= \hat{\beta}_{0,j} + \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top\hat{\boldsymbol{\mu}}_{1:(j-1)}, \quad \text{and} \\ \hat{\boldsymbol{\Sigma}}_{1:j,j} &= \begin{pmatrix} \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top\hat{\boldsymbol{\Sigma}}_{1:(j-1),1:(j-1)} \\ \hat{\sigma}_j^2 + \hat{\boldsymbol{\beta}}_{1:(j-1),j}^\top\hat{\boldsymbol{\Sigma}}_{1:(j-1),1:(j-1)}\hat{\boldsymbol{\beta}}_{1:(j-1),j}, \end{pmatrix}. \end{aligned} \tag{1}$$

The $\hat{\boldsymbol{\Sigma}}$ thereby obtained will be positive-definite, as long as $n_j > j$ for all $j = 1, \ldots, m$ so that $\mathbf{Y}_j$ is of full rank, and $\mathbf{Y}_j^\top\mathbf{Y}_j$ invertible.

2.1 Bayesian inference

The Bayesian approach follows naturally from priors on the auxiliary parameters $\boldsymbol{\beta}_j$ and σ_j^2, for $j = 1, \ldots, m$. Samples from the implied posterior of μ_j and $\boldsymbol{\Sigma}_{1:j,j}$ are then obtained via Φ^{-1} in Eq. (1). It may be more desirable to choose priors directly in the natural parameter space $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, but it can be difficult to analytically derive the implied priors for $\boldsymbol{\beta}_j$ and σ_j^2. However, a popular non-informative prior used for MVN

data, $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\left(\frac{m+1}{2}\right)}$, can be shown (Schafer 1997, Section 6.5.3) to imply the (independent) prior(s) $p(\boldsymbol{\beta}_j, \sigma_j^2) \propto (\sigma_j^2)^{-\left(\frac{m+1}{2}-m+j\right)}$, giving the posterior conditionals:

$$\begin{aligned} \boldsymbol{\beta}_j|\sigma_j^2, \boldsymbol{y}_j, \boldsymbol{Y}_j &\sim \mathcal{N}_j(\hat{\boldsymbol{\beta}}_j, \sigma_j^2(\boldsymbol{Y}_j^\top \boldsymbol{Y}_j)^{-1}) \\ \sigma_j^2|\mathbf{y}_j, \mathbf{Y}_j &\sim \text{IG}((n_j - m + j - 1)/2, (||\mathbf{y}_j - \mathbf{Y}_j\hat{\boldsymbol{\beta}}_j||^2)/2). \end{aligned}$$

Stambaugh (1997) showed that, under this non-informative prior, it is possible to derive the moments of the Bayesian posterior *predictive* distribution in terms of the MLEs $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ in closed form. When these are used in the mean–variance framework to construct portfolios, they are said to take *estimation risk* (Klein and Bawa 1976; Brown 1979) into account. However, these calculations similarly break down when $n_j \leq j$.

Inverted Wishart priors for $\boldsymbol{\Sigma}$ are amenable to tractable posterior inference under the MVN with monotone missingness (Liu 1993). One example is a *ridge prior* (Schafer 1997, Section 5.2.3), which is helpful when $m > n$ and is closely related to *ridge regression* [see Section 3] as used by Gramacy et al. (2008) in $\boldsymbol{\phi}$-space to good effect. This motivates a more pragmatic approach to prior selection: a deliberate search for appropriate shrinkage priors for the "big p small n" regression problem in $\boldsymbol{\phi}$-space, where the low rank $\mathbf{Y}_j$ problem is manifest. Then, Φ^{-1} completes the description implicitly in $\boldsymbol{\theta}$-space.

3 Bayesian shrinkage regression

Here we focus on appropriate regression models for the "big p small n" problem that employ shrinkage. The customary formulation is

$$\mathbf{y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \text{where} \quad \epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n). \tag{2}$$

One typically assumes a standardized $n \times p$ design matrix $\mathbf{X}$ where the columns are individually adjusted to have zero-mean and unit L2-norm. This causes β_0 and $\boldsymbol{\beta}$ to be independent *a posteriori* and recognizes that regularized posterior summaries for $\boldsymbol{\beta}$ are not equivariant under a re-scaling of $\mathbf{X}$. Any such pre-processing must be undone before evaluating $\boldsymbol{\theta} = \Phi^{-1}(\boldsymbol{\phi})$ at samples of $\boldsymbol{\phi} = (\beta_0, \boldsymbol{\beta}, \sigma^2)$.

Ridge regression and the *lasso* (e.g., Hastie et al. 2001, Section 3.4.3) are classical approaches to shrinkage regression that penalize large coefficients:

$$\hat{\boldsymbol{\beta}}^{(q)} = \text{argmin}_{\boldsymbol{\beta}} \left\{ (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{3}$$

for some $\lambda \geq 0$, where the intercept is excluded from penalization via $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_n$. Choosing $q = 2$ yields *ridge regression* (Hoerl and Kennard 1970) where $\hat{\boldsymbol{\beta}}^{(2)} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$. The *lasso* (Tibshirani 1996) corresponds to $q = 1$. There is no closed form solution for $\hat{\boldsymbol{\beta}}^{(1)}$, but the entire path of solutions for all λ can be obtained iteratively via the LARS algorithm (Efron et al. 2004). Both estimators may be interpreted as

the posterior mode under a particular prior. For ridge regression the prior is $\boldsymbol{\beta}^{(2)}|\sigma^2 \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\lambda\mathbf{I}_p)$; for the lasso it is i.i.d. Laplace (i.e., double-exponential) $\pi(\boldsymbol{\beta}^{(1)}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j^{(1)}|/\sqrt{\sigma^2}}$.

Large values of the penalty parameter λ cause the coefficients of $\hat{\boldsymbol{\beta}}^{(q)}$ to be shrunk towards zero. The lasso estimator $\hat{\boldsymbol{\beta}}^{(1)}$ may have many coefficients shrunk to exactly zero, which is convenient for variable selection. Often, λ is chosen via cross validation (CV). As a $\boldsymbol{\phi}$-space regression for obtaining monotone MVN estimators. Gramacy et al. (2008) chose λ by applying the "one-standard-error" rule (Hastie et al. 2001, Section 7.10) with CV.

3.1 Hierarchical models for Bayesian shrinkage regression

For a fully Bayesian lasso we use the latent variable formulation of Park and Casella (2008) and Carlin and Polson (1991) by representing the Laplace as a scale mixture of normals:

$$\begin{aligned}
\mathbf{y}|\beta_0, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}_n(\beta_0\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n), \\
\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{D}_\tau), \quad \mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2), \quad \beta_0 \propto 1 \\
\sigma^2 &\sim \mathrm{IG}(a_\sigma/2, b_\sigma/2) \\
\tau_j^2|\lambda &\overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda^2/2) \\
\lambda^2 &\sim G(a_\lambda, b_\lambda).
\end{aligned} \tag{4}$$

IG and G are the rate- and scale-parameterized inverse-gamma and gamma distributions, respectively. The default prior $\pi(\sigma^2) \propto \sigma^{-2}$ is obtained with $a_\sigma = b_\sigma = 0$, and ridge regression is the special case where $\tau^2 \equiv \tau_1^2 = \cdots = \tau_p^2$, using $\tau^2 \sim \mathrm{IG}(a_\tau/2, b_\tau/2)$ (i.e., dropping λ^2), possibly with $a_\tau = b_\tau = 0$. Fixing $\tau^2 = \infty$ yields the standard family of (improper) priors for linear regression.

Choosing the prior parameterization for λ^2 can be difficult. Park and Casella (2008) note that choosing $a_\lambda = b_\lambda = 0$ leads to an improper posterior, and suggest some automatic alternatives. Another option is to further expand the hierarchy using a so-called normal-gamma (NG) prior for β (e.g., Griffin and Brown 2010) by specifying

$$\begin{aligned}
\lambda^2|\gamma &\sim G(a_\lambda, b_\lambda/\gamma), \quad \text{where} \quad \gamma \sim \mathrm{Exp}(1) \\
\text{and} \quad \tau_j^2|\lambda^2, \gamma &\overset{\text{iid}}{\sim} G(\gamma, \lambda^2/2).
\end{aligned} \tag{5}$$

Griffin and Brown (2010) suggest $a_\lambda = 2$, and $b_\lambda = M/2$, where M is chosen via empirical Bayes considerations. Observe that fixing $\gamma = 1$ encodes the specific Laplace prior case. So the NG prior is more adaptive than the lasso. This may come in handy when $p \gg n$, i.e., when our prior plays a more important role, or when the posterior drives many β_j's to zero.

The availability of full conditionals for all of the parameters makes for efficient Gibbs sampling (GS). For the baseline Bayesian lasso model (Park and Casella 2008) these

are:

$$\begin{aligned}
\beta_0|\sigma^2, \mathbf{y} &\sim \mathcal{N}(\bar{y}, \sigma^2/n) \\
\boldsymbol{\beta}|\sigma^2, \{\tau_j^2\}_{j=1}^p, \mathbf{y} &\sim \mathcal{N}_p(\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{A}^{-1}), \quad \mathbf{A} = \mathbf{X}^\top\mathbf{X} + \mathbf{D}_\tau^{-1}, \;\; \tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{X}^\top\tilde{\mathbf{y}} \\
\sigma^2|\boldsymbol{\beta}, \{\tau_j^2\}_{j=1}^p, \mathbf{y} &\sim \text{IG}((a_\sigma + n - 1 + p)/2, (b_\sigma + \psi_\beta)/2), \\
\psi_\beta &= ||\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}||^2 + \boldsymbol{\beta}^\top\mathbf{D}_\tau^{-1}\boldsymbol{\beta} \\
\tau_j^{-2}|\beta_j, \sigma^2, \lambda &\overset{\text{iid}}{\sim} \text{Inv-Gauss}(\sqrt{\lambda^2\sigma^2/\beta_j^2}, \lambda^2) \\
\lambda^2|\tau_1^2, \ldots, \tau_p^2 &\sim \text{G}(a_\lambda + p\gamma, b_\lambda/\gamma + \textstyle\sum_{j=1}^p \tau_j^2/2). \quad \text{[assuming } \gamma = 1\text{]}
\end{aligned} \tag{6}$$

Using a marginal posterior conditional for σ^2 instead can help reduce autocorrelation in the Markov chain. Integrating over the posterior conditional for $\boldsymbol{\beta}$ gives that $\sigma^2|\tau_1^2, \ldots, \tau_p^2, \mathbf{y} \sim \text{IG}((a_\sigma + n - 1)/2, (b_\sigma + \psi_{\tilde{\beta}})/2)$, where $\psi_{\tilde{\beta}} = ||\tilde{\mathbf{y}} - \mathbf{X}\tilde{\boldsymbol{\beta}}||^2 + \tilde{\boldsymbol{\beta}}^\top\mathbf{D}_\tau^{-1}\tilde{\boldsymbol{\beta}} = \tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} - \tilde{\boldsymbol{\beta}}^\top\mathbf{A}\tilde{\boldsymbol{\beta}}$.

Under the ridge regression model the posterior conditionals are the same (6) except that we ignore λ^2 and take $\tau^2 \sim \text{IG}((a_\tau + p)/2, (b_\tau + \sigma^{-2}\boldsymbol{\beta}^\top\boldsymbol{\beta})/2)$. Upon fixing $\tau^2 = \infty$ we must use $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$, subtract $p/2$ to the rate parameter to the IG conditional(s) for σ^2, and ensure a proper posterior with $a_\sigma > p - n - 1$. This may pose a non-trivial restriction on the prior when $p \geq n$. Under the NG prior γ may vary, leading to the conditionals

$$\begin{aligned}
\tau_j^2|\beta_j, \sigma^2, \lambda^2, \gamma &\overset{\text{iid}}{\sim} \text{GIG}(\gamma - 1/2, \beta_j^2/\sigma^2, \lambda^2) \\
\gamma|\{\tau_j^2\}_{j=1}^p, \lambda^2 &\propto \left(\frac{\lambda^2}{2}\right)^{p\gamma} \frac{\pi(\gamma)}{(\Gamma(\gamma))^p} \left(\prod_{i=1}^p \tau_j^2\right)^\gamma,
\end{aligned} \tag{7}$$

where $\text{GIG}(\lambda, \chi, \psi)$ is the generalized inverse Gaussian distribution. Griffin and Brown (2010) suggest a random walk Metropolis update for the γ using proposals $\gamma' = \exp\{\sigma_\gamma z\}$, for $z \sim \mathcal{N}(0, 1)$. These proposals are accepted with probability

$$\min\left\{1, \frac{\pi(\gamma')}{\pi(\gamma)}\left(\frac{\Gamma(\gamma)}{\Gamma(\gamma')}\right)^p \left(\left(\frac{2}{\lambda^2}\right)^{-p} \prod_{i=1}^p \tau_j^2\right)^{\gamma' - \gamma}\right\}, \tag{8}$$

where $\pi(\gamma) = \gamma^{-2}\exp(-\gamma - \frac{M}{2\gamma}\lambda^2)$ and σ_γ is chosen to give an acceptance rate of 20-30%.

An alternative hierarchical modeling framework for the Bayesian lasso is provided by Hans (2008). While it does not require p latent τ_j^2 variables, the resulting GS procedure is not fully blocked, and rejection sampling is required for σ^2. "Orthogonalizing" the sampler helps mitigate slow mixing of the un-blocked conditionals. However, we prefer the simpler approach of Park and Casella (2008) as it is more readily adaptable to the $p \gg n$ case, to model selection, and our extensions to the heavy-tailed errors.

Whereas the classical lasso has the property that the estimate $\hat{\boldsymbol{\beta}}^{(1)}$ may have components which are zero—in fact, it would never have more than $\min\{p, n-1\}$ nonzero components— samples of $\boldsymbol{\beta}$ from the posterior would never have zeros. So the Bayesian

lasso is less useful for variable selection. We also note that when $p \geq n$—and without the ability to explicitly restrict $\boldsymbol{\beta}$ to having at most $\min\{p, n-1\}$ nonzero components—a proper prior must be used for σ^2 or the posterior will be improper. An empirical Bayes remedy that works well in this case is to take a small a_σ, say $a_\sigma = 3/2$, and then set b_σ so that the $(1-\alpha)$ part of the $\text{IG}(a_\sigma, b_\sigma)$ distribution lies at the point $\tilde{\mathbf{y}}^\top \tilde{\mathbf{y}}$ (i.e., the MLE under the intercept model) via the incomplete gamma inverse function. Another remedy is Bayesian model averaging.

3.2 Bayesian model selection and averaging

Although the MAP lasso fit may indeed set some of the coordinates of $\hat{\boldsymbol{\beta}}^{(1)}$ to zero, this is more of a side effect of the solution space of the quadratic program (3) than the result of a deliberate prior modeling choice (Hans 2008). Bayesians rarely base inference on the MAP; it is more natural to select variables by inspecting the posterior model probabilities.

There are several standard ways of performing Bayesian variable selection in regression models that are amenable to GS. They essentially fall into two camps. Loosely, the first camp (e.g., Geweke 1996; George and McCulloch 1993) uses a product-space wherein the prior for each β_j is augmented to include a point-mass at zero. Inference proceeds by GS on each of the conditionals $\beta_j | \boldsymbol{\beta}_{-j}, \mathbf{y}, \ldots$, $j = 1, \ldots, p$, which may flop between zero and nonzero values. Hans (2008) augmented this product space approach to variable selection under the Laplace prior by further conditioning on λ.

The second camp (e.g., Troughton and Godsill 1997) is transdimensional in that the $\boldsymbol{\beta}$-vector may vary in length while model space is traversed via Reversible Jump (RJ) MCMC (Green 1995). We prefer this approach for our monotone inference application in the Park and Casella (2008) setup. When $p \gg n$ it is implementationally more compact, only requiring memory for the (nonzero) $\boldsymbol{\beta}$-components. This represents a big savings when simultaneously storing m regression model parameter sets under a prior that restricts the model to have at most $\min\{j, n_j - 1\}$ (nonzero) coefficients. Also, we like the fully blocked samples for the nonzero components of $\boldsymbol{\beta}$ for the within-model moves.

Suppose that the transdimensional Markov chain is currently visiting some model with k nonzero regression coefficients $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_k)$ using design matrix $\mathbf{X}_k$. The columns of $\mathbf{X}_k$ should come from a two-way partition (of k and $p-k$ elements) of the p columns of $\mathbf{X}$, but they need not coincide with the first k of the p columns. Now consider proposing to add a column to $\mathbf{X}_k$, a so-called "birth" move. Choose one of the $p-k$ columns of $\mathbf{X}$ not present in $\mathbf{X}_k$ for addition, thus creating $\mathbf{X}_{k+1}$. By considering the ratio of the marginal posterior distributions (integrating out $\boldsymbol{\beta}_k$ and $\boldsymbol{\beta}_{k+1}$) conditional on $\sigma^2, \tau_1^2, \ldots, \tau_k^2$ and a new proposed τ_{k+1}^2 (which we take from the prior), it can be shown that the transdimensional move may be accepted with probability

$\min\{1, A_{k\to k+1}\}$, where

$$A_{k\to k+1} = \frac{(\tau_{k+1}^{-2}|\mathbf{A}_{k+1}^{-1}|)^{\frac{1}{2}} \exp\left\{\frac{1}{2\sigma^2}\tilde{\boldsymbol{\beta}}_{k+1}^\top \mathbf{A}_{k+1}\tilde{\boldsymbol{\beta}}_{k+1}\right\}}{|\mathbf{A}_k^{-1}|^{\frac{1}{2}} \exp\left\{\frac{1}{2\sigma^2}\tilde{\boldsymbol{\beta}}_k^\top \mathbf{A}_k\tilde{\boldsymbol{\beta}}_k\right\} q(\tau_{k+1}^2)} \times \frac{\pi(k+1)q(k+1\to k)}{\pi(k)q(k\to k+1)}, \quad (9)$$

and $\mathbf{A}_k = \mathbf{X}_k^\top\mathbf{X}_k + \mathbf{D}_{\tau_k}^{-1}$, $\tilde{\boldsymbol{\beta}}_k = \mathbf{A}_k^{-1}\mathbf{X}_k^\top\tilde{\mathbf{y}}$, with $\mathbf{D}_{\tau_k} = \mathrm{diag}(\tau_1^2, \ldots, \tau_k^2)$. The reverse "death" move, of proposing to remove one of the columns of $\mathbf{X}_k$, may be accepted with probability $\min\{1, A_{k-1\to k}^{-1}\}$. Under the ridge prior, $\tau_{k+1}^2 = \tau^2$ can be dropped from the expression unless $k = 0$; for standard regression it may be ignored so long as a proper prior is used for $\boldsymbol{\beta}_k$.

A uniform prior over all models with k nonzero components is typical. Often, $\pi(k) \propto 1, \forall k \in \{1, \ldots, p^*\}$. However, we prefer to take $k \sim \mathrm{Bin}(p^*, \pi)$, with $\pi \in (0, 1)$ where π controls the "sparsity", and p^* denotes p or $\min\{p, n-1\}$ for compactness. Prior information on π may be interjected either by fixing a particular value, or by taking a hierarchical approach with $\pi \sim \mathrm{Beta}(g, h)$ (e.g., George and McCulloch 1993). Hans (2008) used $g = h = 1$ for with a Laplace prior in the product space. The posterior conditional for GS is $\pi|k \sim \mathrm{Beta}(g+k, g+p^*-k)$. In our transdimensional approach, we choose a uniform proposal for the valid jumps. For a "birth" we take $q(0 \to 1) = 1/p$, and $q(k \to k+1) = 1/2(p-k)$ for $k = 1, \ldots, p^*-1$. Conversely for a "death" we take $q(p^* \to p^*-1) = 1/p^*$ and $q(k \to k-1) = 1/2k$ for $k = p^*-1, \ldots, 1$. Otherwise $q(k \to k') = 0$.

Movement throughout the 2^p sized space will be slow for large p, so a certain amount of thinning of the RJ-MCMC chain is appropriate. Collecting a sample from the posterior after p transdimensional moves approximates the model-level mixing (and computation burden) of the product-space approach. Throughout the RJ-MCMC the length of $\boldsymbol{\beta}$ varies, and the components shift to represent the partition of $\mathbf{X}$ stored in the columns of $\mathbf{X}_k$. Therefore, post-processing is necessary if samples of $\boldsymbol{\beta}$ are to be used elsewhere, e.g., in $\boldsymbol{\theta}$-space via Φ^{-1} (1), where a full p-vector having zero and nonzero entries in the correct positions is needed. This may be facilitated by maintaining a k-vector of column indicators. The posterior probability that variable j, $j = 1, \ldots, p$, is relevant for predicting $\mathbf{y}$ is then proportional to $\sum_{t=0}^{T} I_{\{\beta_j^{(t)} \neq 0\}}$, where T is the number of samples saved from the Markov chain.

3.3 Student-t errors via scale-mixtures

The MVN assumption is not always appropriate. We may wish to consider the possibility that errors in $\mathbf{y}$ have a Student-t distribution with an unknown degrees of freedom ν:

$$\mathbf{y} = \beta_0\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \{\epsilon_i\}_{i=1}^n \overset{\text{iid}}{\sim} \mathrm{St}(0, \sigma^2; \nu). \quad (10)$$

Following Carlin et al. (1992) and Geweke (1993) we shall represent the Student-t distribution as a scale mixture of normals with an $\mathrm{IG}(\nu/2, \nu/2)$ mixing density.

We must redefine $\mathbf{X} = (\mathbf{1}_n, \mathbf{X})$ as a $n \times (p+1)$ matrix, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}^\top)^\top = \{\beta_j\}_{j=0}^p$ so that the model in Eq. (10) becomes $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ since the posterior intercept β_0 is no

longer independent of the other components of $\boldsymbol{\beta}$ in the presence of heavy-tailed errors. The setup is otherwise unchanged from Section 3.1. Upon assuming an exponential prior for the degrees of freedom parameter, ν, the modifications to the hierarchical model in Eq. (4) are:

$$\begin{aligned} \mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2,\{\omega_i^2\}_{i=1}^n &\sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{D}_\omega), \quad \mathbf{D}_\omega = \mathrm{diag}(\omega_1^2,\ldots,\omega_n^2) \\ \boldsymbol{\beta}|\sigma^2,\{\tau_j^2\}_{j=1}^p &\sim \mathcal{N}_{p+1}(\mathbf{0},\sigma^2\mathbf{D}_\tau), \quad \mathbf{D}_\tau = \mathrm{diag}(\infty,\tau_1^2,\ldots,\tau_p^2) \\ \omega_i^2|\nu &\overset{\text{iid}}{\sim} \mathrm{IG}(\nu/2,\nu/2) \\ \nu|\theta &\sim \mathrm{Exp}(\theta). \end{aligned} \tag{11}$$

Note that $\mathbf{D}_\tau$ is a $p+1$ diagonal matrix, and that the first component insures that β_0 is given a flat prior as before. After redefining $\mathbf{A} = \mathbf{X}^\top\mathbf{D}_\omega^{-1}\mathbf{X} + \mathbf{D}_\tau^{-1}$, $\tilde{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{X}^\top\mathbf{D}_\omega^{-1}\mathbf{y}$ and $\psi_\beta = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top\mathbf{D}_\omega^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^\top\mathbf{D}_\tau^{-1}\boldsymbol{\beta}$, the modified full posterior conditionals follow:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2,\{\tau_j^2\}_{j=1}^p,\{\omega_i^2\}_{i=1}^n,\mathbf{y} &\sim \mathcal{N}_{p+1}(\tilde{\boldsymbol{\beta}},\sigma^2\mathbf{A}^{-1}) \\ \sigma^2|\boldsymbol{\beta},\{\tau_j^2\}_{j=1}^p,\{\omega_i^2\}_{i=1}^n,\mathbf{y} &\sim \mathrm{IG}\left(\frac{a_\sigma+n+p}{2}, \frac{b_\sigma+\psi_\beta}{2}\right) \\ \omega_i^2|\boldsymbol{\beta},\sigma^2,\nu,\mathbf{y} &\overset{\text{iid}}{\sim} \mathrm{IG}\left(\frac{\nu+1}{2}, \frac{\nu+\sigma^{-2}((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})_i)^2}{2}\right) \\ p(\nu|\{\omega_i^2\}_{i=1}^n,\theta) &\propto \left(\frac{\nu}{2}\right)^{\frac{n\nu}{2}}\left(\Gamma\left(\frac{\nu}{2}\right)\right)^{-n}\exp(-\eta\nu) \end{aligned} \tag{12}$$

where $\eta = \frac{1}{2}\sum_{i=1}^n(\log(\omega_i^2)+\omega_i^{-2}) + \theta$.[1]

The conditional posterior of ν does not correspond to a standard distribution, however a convenient rejection sampling method (with low rejection rate) is available (Geweke 1992) using an exponential envelope. The optimal scale parameter ν^* can be chosen to minimize the unconditional rejection rate by finding the root of $(n/2)[\log(\nu/2)+1-\Psi(\nu/2)]+\nu^{-1}-\eta$, where Ψ is the digamma function. Standard Newton-like methods work well. A draw from $\nu \sim \mathrm{Exp}(\nu^*)$ may then be retained with probability[2]

$$\min\left\{1, \left[\frac{\Gamma(\nu^*/2)}{\Gamma(\nu/2)}\right]^n\left[\frac{(\nu/2)^\nu}{(\nu^*/2)^{\nu^*}}\right]^{n/2}\exp[(\nu-\nu^*)((\nu^*)^{-1}-\eta)]\right\}.$$

As before we may integrate out $\boldsymbol{\beta}$ obtaining $\sigma^2|\{\tau_j^2\}_{j=1}^p,\{\omega_i^2\}_{i=1}^n,\mathbf{y} \sim \mathrm{IG}((a_\sigma+n-1)/2,(b_\sigma+\psi_{\tilde{\beta}})/2)$ by redefining $\psi_{\tilde{\beta}} = (\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}})^\top\mathbf{D}_\omega^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\beta}}) + \tilde{\boldsymbol{\beta}}^\top\mathbf{D}_\tau^{-1}\tilde{\boldsymbol{\beta}} = \mathbf{y}^\top\mathbf{D}_\omega^{-1}\mathbf{y} - \tilde{\boldsymbol{\beta}}^\top\mathbf{A}\tilde{\boldsymbol{\beta}}$. Finally, the Bayesian model selection and averaging method of Section 3.2, via Eq. (9), may be used with $\mathbf{X}_k = (\mathbf{1}_n,\mathbf{X}_k)$, $\boldsymbol{\beta}_k = (\beta_0,\boldsymbol{\beta}_k^\top)^\top$, $\tilde{\boldsymbol{\beta}}_k = \mathbf{A}_k^{-1}\mathbf{X}_k^\top\mathbf{D}_\omega^{-1}\mathbf{y}$ and $\mathbf{A}_k = \mathbf{X}_k^\top\mathbf{D}_\omega^{-1}\mathbf{X}_k + \mathbf{D}_{\tau_k}^{-1}$ and $\mathbf{D}_{\tau_k} = \mathrm{diag}(\infty,\tau_1^2,\ldots,\tau_k^2)$. The number of latent variables now grows with the sample size, so automatic $O(n)$ thinning from the Markov Chain is sensible.

[1] Note that there is a typo in the conditional for ν provided by Geweke (1993).
[2] There is also a typo in the acceptance probability provided by Geweke (1992).

3.4 Empirical results on detecting fat tails

Hans (2008) and Griffin and Brown (2010) offer a plethora of insights about the Bayesian lasso and NG with comparison to the classical lasso. There is no need to re-produce these results here. Instead we offer a demonstration of the Student-t extensions that are unique to our setup and relevant in light of the recent criticism of MVN in the financial press. Basically, we explore the extent to which deviations from normality may be detected by testing the null hypothesis (model $\mathcal{M}_{\mathrm{N}}$) of normal errors versus the alternative (model $\mathcal{M}_{\mathrm{St}}$) that they follow a Student-t with ν unknown. One way to do this is via a posterior odds ratio (POR):

$$\frac{p(\mathcal{M}_{\mathrm{N}}|\mathbf{y})}{p(\mathcal{M}_{\mathrm{St}}|\mathbf{y})} = \frac{\pi(\mathcal{M}_{\mathrm{N}})}{\pi(\mathcal{M}_{\mathrm{St}})} \times \frac{p(\mathbf{y}|\mathcal{M}_{\mathrm{N}})}{p(\mathbf{y}|\mathcal{M}_{\mathrm{St}})} \equiv [\text{prior ratio}] \times [\text{Bayes factor}]$$

where $\pi(\mathcal{M}_*)$ is the prior on $\mathcal{M}_*$, and $p(\mathbf{y}|\mathcal{M}_*)$ is the marginal likelihood for $\mathcal{M}_*$. By taking equal priors we may concentrate on the Bayes factor (BF).

Calculating PORs and BFs can be difficult in generality; for a review of related methods see Godsill (2001). However, we may exploit that the Student-t and normal models differ by just one parameter in the likelihood, ν. Jacquier et al. (2004, Section 2.5.1) show that this BF may be calculated by writing it as the expectation of the ratio of un-normalized posteriors with respect to the posterior under the Student-t model. That is, we may calculate

$$E\left\{\frac{p(\mathbf{y}|\boldsymbol{\psi}, \mathcal{M}_{\mathrm{N}})}{p(\mathbf{y}|\boldsymbol{\psi}, \nu, \mathcal{M}_{\mathrm{St}})}\right\} \approx \frac{1}{T}\sum_{t=1}^{T} \frac{p(\mathbf{y}|\boldsymbol{\psi}^{(t)}, \mathcal{M}_{\mathrm{N}})}{p(\mathbf{y}|\boldsymbol{\psi}^{(t)}, \nu^{(t)}, \mathcal{M}_{\mathrm{St}})},$$

where $(\boldsymbol{\psi}^{(t)}, \nu^{(t)}) \sim p(\boldsymbol{\psi}, \nu|\mathbf{y}, \mathcal{M}_{\mathrm{St}})$, and $\boldsymbol{\psi}$ collects the parameters shared by both models.

To shed light on the "selectability" of the Student-t model (10), consider synthetic data where $\boldsymbol{\beta} = (2, -3, 0, 0.75, 0, 0, -0.9)^\top$, $\mu \equiv \beta_0 = 1$, the rows of the $n \times 7$ design matrix $\mathbf{X}$ are uniformly distributed in $[0,1]^7$, and $\epsilon_i \sim \mathrm{St}(0, \sigma^2 = 1; \nu)$, for $i = 1, \ldots, n$. We perform a Monte Carlo experiment where n and ν vary, with $n \in \{30, 75, 100, 200, 500, 1000\}$ and $\nu \in \{3, 5, 7, 10, \infty\}$, and consider the frequency of times that the BF indicated "strong" preference for the correct model in repeated trials. In each trial, GS (12) was used to obtain 1200 samples from the posterior by thinning every $7n$ rounds, with the first 200 discarded as burn-in. For $n \leq 200$ we repeated the experiment with random data 300 times; when $n = 500$ we used 50 replications; and when $n = 1000$ we used 20.

Figure 3 shows the relationships between n, ν and the frequency of correct model determinations (higher frequencies are better). In the case of normal errors, and Student-t errors with $\nu = 3$, the correct model can be determined with high accuracy when $n \geq 200$. When $\nu = 5$ a sample size of $n = 1000$ is needed; when $\nu = 7, 10$ we need $n \gg 1000$. Clearly for $10 \leq \nu < \infty$ the situation is hopeless unless n is very large. These results have implications in the context of our motivating financial returns data in Sec-

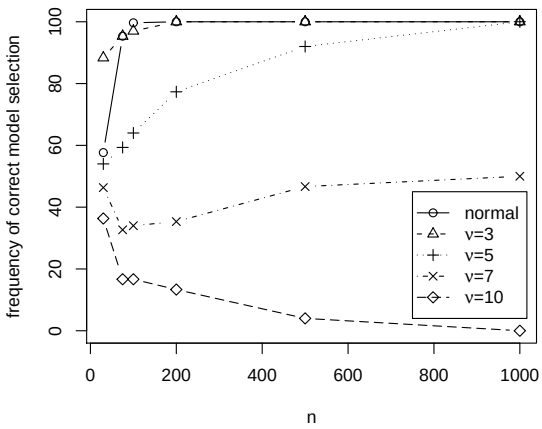

Figure 3: Frequency of correct model determinations as a function of the sample size, n, and the degrees of freedom parameter, ν, where "normal" is interpreted as $\nu = \infty$.

tion 5, indicating that long return histories may be required to benefit from relaxing the MVN assumption.

4 Bayesian inference under monotone missingness

Here we collect ideas from the previous sections in order to sample from the joint posterior distribution of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\boldsymbol{\phi}_j = (\beta_{0,j}, \boldsymbol{\beta}_j, \sigma_j^2) \sim \mathcal{BR}_j \equiv \mathcal{BR}(\mathbf{Y}_j, \mathbf{y}_j)$ represent samples collected from the posterior of the chosen $\boldsymbol{\phi}$-space (shrinkage) regression models—one of the ones from Section 3. Then, following Eq. (1) from Section 2, samples from the posterior distribution of $\boldsymbol{\theta}$ may be obtained by repeating the following steps.

1. Sample $(\mu_1, \Sigma_{1,1}) \equiv (\beta_{0,1}, \sigma_1^2) \sim \mathcal{BR}_1$. See below for details of this special case.

2. For $j = 2, \ldots, m$:

   (a) Sample $(\beta_{0,j}, \boldsymbol{\beta}_j, \sigma_j^2) \sim \mathcal{BR}_j$.

   (b) Convert $(\mu_j, \boldsymbol{\Sigma}_{1:j,j}) = \Phi^{-1}(\beta_{0,j}, \boldsymbol{\beta}_j, \sigma_j^2, \boldsymbol{\mu}_{1:(j-1)}, \boldsymbol{\Sigma}_{1:(j-1),1:(j-1)})$, following Eq. (1).

Since the Bayesian regressions $(\mathcal{BR}_j)$ are mutually independent, step 1 and the $j-1$ steps of 2(a) may be performed in parallel, and the conversion via Φ^{-1} in 2(b) may be performed offline. Zeros in $\boldsymbol{\beta}_j$ may translate into zeros in $\boldsymbol{\Sigma}_{1:j,j}$ which may be used to test hypotheses about the marginal and conditional independence between assets.

In the default formulation of $\mathcal{BR}_j$ with standard normal errors (4) we use $\mathbf{Y}_j \equiv \mathbf{Y}^{(n_j)}_{1:(j-1)}$. In this case the first step above ($j = 1$) simplifies to:

$$\Sigma_{1,1} \sim \text{IG}\left(\frac{a_\sigma + n_1 - 1}{2}, \frac{b_\sigma + ||\tilde{\mathbf{y}}_1||^2}{2}\right), \quad \text{then} \quad \mu_1 \sim \mathcal{N}(\bar{y}_1, \Sigma_{1,1}/n). \tag{13}$$

If heavy-tailed errors are modeled [Section 3.3] for the ϕ-space regressions ($\mathcal{BR}_j$), then take $\mathbf{Y}_j \equiv \mathbf{Y}^{(n_j)}_{0:(j-1)}$. In this case the first step above ($j = 1$) requires integrating over $\{\omega^2_{i1}\}^{n_1}_{i=1}$. Conditional on a particular $\mathbf{D}_{\omega_1} = \text{diag}(\omega^2_{11}, \ldots, \omega^2_{n_1 1})$ sampled from their full conditional under $\mathcal{BR}_1$ (conditional on ν_1 in Eq. (12) which must also be integrated out), we may sample

$$\Sigma_{1,1} \sim \text{IG}\left(\frac{a_\sigma + n - 1}{2}, \frac{b_\sigma + \mathbf{y}_1^\top \mathbf{D}^{-1}_{\omega_1} \mathbf{y}_1}{2}\right), \quad \text{then} \quad \mu_1 \sim \mathcal{N}\left(\frac{\mathbf{D}^{-1}_{\omega_1}\mathbf{y}_1}{n_{\omega_1}}, \frac{\Sigma_{1,1}}{n_{\omega_1}}\right), \tag{14}$$

where $n_{\omega_1} = \sum_{i=1}^{n_1} \omega_{i1}^{-2}$. Carrying this through in the second step, above, for $j = 2, \ldots, m$, etc., integrating over independent $(\nu_2, \{\omega^2_{i2}\}^{n_2}_{i=1}), \ldots (\nu_m, \{\omega^2_{im}\}^{n_m}_{i=1})$, is similarly parallelizable. The marginal Student-t error structure for $\mathbf{y}_j$ carries over into $\boldsymbol{\theta}_j$-space giving a distinct degrees of freedom parameter ν_j for each (marginal) $\mathbf{y}_j$, for $j = 1, \ldots, m$. So the resulting model in $\boldsymbol{\theta}$-space is not the typical multivariate Student-t which has a single ν.

This more standard multivariate Student-t model may be obtained by modifying the prior so that all ω^2_{ij} depend on a common ν. I.e., for $i = 1, \ldots, n_j$ and $j = 1, \ldots, m$

$$\omega^2_{ij} | \boldsymbol{\beta}_j, \sigma^2_j, \nu \overset{\text{iid}}{\sim} \text{IG}\left(\frac{\nu + 1}{2}, \frac{\nu + \sigma_j^{-2}((\mathbf{y}_j - \mathbf{Y}_j \boldsymbol{\beta}_j)_i)^2}{2}\right)$$

(using intercept-extended $\mathbf{Y}_j$ and β_j). Then, the full conditional of ν becomes

$$p(\nu | \{\{\omega^2_{ij}\}^{n_j}_{i=1}\}^m_{j=1}, \theta) \propto \left(\frac{\nu}{2}\right)^{\frac{\nu}{2}\sum_{j=1}^m n_j} \left(\Gamma\left(\frac{\nu}{2}\right)\right)^{-\sum_{j=1}^m n_j} \exp(-\eta\nu),$$

where $\eta = \frac{1}{2}\sum_{j=1}^m \sum_{i=1}^{n_j} (\log(\omega^2_{ij}) + \omega^{-2}_{ij}) + \theta$. The same rejection sampling method (i.e., using $n = \sum_{j=1}^m n_j$) applies. Although we may proceed with sampling from $\mathcal{BR}_j$ using $\nu_j^{(t)} \equiv \nu^{(t)}$ ignoring the full conditional for ν_j, the independence between the $\mathcal{BR}_j$ is now broken: parallelization of the MCMC is bottlenecked by sampling the common ν.

The remainder of the section covers the extensions particular to the portfolio balancing problem: handling "gaps", incorporating known factors, and accounting for estimation risk.

4.1 Dealing with "gaps" by monotone data augmentation

Data augmentation (DA) is an established (Bayesian) technique for dealing with missing data. In short, it involves treating the unknown portion of the data as latent variables

and updating, or *imputing*, their values jointly with the other unknown (model) parameters via the posterior predictive. For a high level overview see Schafer (1997, Section 3.4.2). Rather than treat *all* of the missing data as latent in this way, it is sufficient to impute a small portion to achieve a monotone missingness pattern. Then, inference may proceed as already described. This is known as *monotone data augmentation* (MDA) (Li 1988; Schafer 1997, Section 6.5.4). In the case of the financial returns data $\mathbf{Y}$ at hand, with sorted columns and rows, the appropriate candidates for imputation are easily spotted.

Consider each $y_{i,j} = \texttt{NA}$ such that there exists a $y_{i,j'} \neq \texttt{NA}$ where $j' > j$. Specially mark these with $y_{i,j} = \texttt{NaN}$, say, as these are the entries that must be treated as latent. There will not be any if the pattern is monotone. Then sort the rows of $\mathbf{Y}$ by the number entries which are (still) NA so that those with more NAs appear towards the bottom of $\mathbf{Y}$. Define $n_j = \sum_{i=1}^{n} I_{\{y_{i,j} \neq \texttt{NA}\}}$ as before, but now its interpretation is as the number of observed entries in the j^{th} column, plus the number which are treated as latent. Some entries of $\mathbf{Y}_j$ and $\mathbf{y}_j$, defined as before, may contain NaNs. However, note that $y_{n_j,j} \neq \texttt{NaN}$ by construction.

Let $\mathbf{r}_j$ index the rows of column j of $\mathbf{Y}$ such that $\mathbf{y}_j[\mathbf{r}_j] = \texttt{NaN}$. When sampling from $\mathcal{BR}_j$ in step 2(a), above, ignore the rows of $\mathbf{Y}_j$ and $\mathbf{y}_j$ in $\mathbf{r}_j$, but otherwise proceed as usual. Then, add a step, 2(c), wherein for $i \in \mathbf{r}_j$, take $y_{i,j} \sim \mathcal{N}(\boldsymbol{\beta}_j \mathbf{Y}_{i,1:j}, \sigma_j^2)$, where $\mathbf{Y}_{i,1:j}$ is the row vector containing the i^{th} row of $\mathbf{Y}_j$. Observe that the mutual independence of the $\mathcal{BR}_j$ is broken by this MDA. They must be processed in serial so that $\mathcal{BR}_{j+1}(\mathbf{Y}_{j+1}, \mathbf{y}_{j+1})$ may proceed with an up to date copy of $\mathbf{Y}$.

4.2 Incorporating known factors

A popular way of developing an estimator of the covariance matrix of financial asset returns is via *factor models*. The idea is that certain market-level indices, like the value-weighted market index, the size of the firm associated with the asset, and the book-to-market factor (e.g., Chan et al. 1999; Fama and French 1993) provide a good basis for describing individual returns. Importantly, these factors are easy to calculate as a function of readily available "fundamentals" (characteristics of the listed assets and companies) and the stock returns. Through covariances calculated between the factors and individual asset returns we may infer covariances between each of the assets. For a $n \times K$ matrix of (known) factors $\mathbf{F}$, where $K \ll \min\{n, p\}$ and where the factors have covariance $\boldsymbol{\Omega}$, the factor model approach poses the following model for the returns $\mathbf{Y}$ via columns $j = 1, \ldots, m$:

$$\mathbf{y}_j = \lambda_{0,j} + \mathbf{F}_j \boldsymbol{\lambda}_j + \boldsymbol{\epsilon}_j, \quad \text{where} \quad \boldsymbol{\epsilon}_j \sim \mathcal{N}_{n_j}(\mathbf{0}, \sigma_j^2 \mathbf{I}_{n_j}). \tag{15}$$

Take $\mathbf{F}_j \equiv \mathbf{F}_{1:(j-1)}^{(n_j)}$, i.e., without a column of ones, and treat the regression coefficients $\lambda_{0,j}, \boldsymbol{\lambda}_j$ as unknown. If $\boldsymbol{\Lambda}$ is the $K \times m$ matrix defined by collecting the $\boldsymbol{\lambda}_j$ column-wise, then a covariance matrix on the returns $\mathbf{Y}$ may be obtained as

$$\boldsymbol{\Sigma}^{(f)} = \boldsymbol{\Lambda}^\top \boldsymbol{\Omega} \boldsymbol{\Lambda} + \mathbf{D}_\sigma, \quad \text{where} \quad \mathbf{D}_\sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2).$$

The MLE $\hat{\boldsymbol{\Sigma}}^{(f)}$ may be obtained via the standard estimate of $\hat{\boldsymbol{\Omega}}$ and the MLEs $\{\hat{\boldsymbol{\lambda}}_j\}_{j=1}^m$. Similarly, one may sample from the Bayesian posterior with suitable (non-informative) independent priors on $\boldsymbol{\Omega}$, and $\{\lambda_{0,j}, \boldsymbol{\lambda}_j, \sigma_j^2\}_{j=1}^m$.

The estimators of $\boldsymbol{\Sigma}^{(f)}$ obtained in this way tend to have low variance, which is a desirable property. However, they also have a strong bias, which may be undesirable. The bias stems from an implicit assumption that the returns are mutually independent when conditioned on the factors. Not only might this not be a reasonable assumption, but it also makes the quality of the resulting estimator(s) extremely sensitive to the choice of factors. This bias may be mitigated to some extent by further involving the returns in the estimation process, i.e., in a more direct way. One such approach, considered by Ledoit and Wolf (2002), is to take a convex combination of a factor-based estimator ($\hat{\boldsymbol{\Sigma}}^{(f)}$) and a standard (possibly non-positive definite) complete data estimator ($\hat{\boldsymbol{\Sigma}}^{(c)}$):

$$\hat{\boldsymbol{\Sigma}}^{(\ell)} = \alpha\hat{\boldsymbol{\Sigma}}^{(f)} + (1-\alpha)\hat{\boldsymbol{\Sigma}}^{(c)}, \quad \text{for } \alpha \in [0,1]. \tag{16}$$

The mixing proportion, α, may be determined by CV. That this approach works well is a testament to the importance of combining a factor model with a more direct approach.

Gramacy et al. (2008) described hybrid method of incorporating the factors into the $\boldsymbol{\phi}$-space procedure of the monotone factorization MLE so the data may inform on which independence assumptions are adequate. Consider the *combined* regression model:

$$\mathbf{y}_j = \beta_{0,j} + \mathbf{Y}_j\boldsymbol{\beta}_j + \mathbf{F}_j\boldsymbol{\lambda}_j + \epsilon_j. \tag{17}$$

Observe that the $\lambda_{0,j}$ term present in Eq. (15) has been dropped because it is not identifiable in the presence of $\beta_{0,j}$. With some bookkeeping, the model described in Eq. (17) can be used to obtain a joint estimator for a $m+K$ element mean vector and $(m+K)\times(m+K)$ covariance matrix from which $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may be extracted. To fix ideas, suppose that the factors are completely observed, which is usually the case. Then we may sample from the regression models in $\mathcal{BR}([\mathbf{F}_j\ \mathbf{Y}_j], \mathbf{y}_j)$ in $\boldsymbol{\phi}$-space, and after transformation to $\boldsymbol{\theta}$-space via Φ^{-1} in Eq. (1) the latter m components $\boldsymbol{\mu}_{(K+1):(m+K)}$, and m rows/cols $\boldsymbol{\Sigma}_{(K+1):(m+K),(K+1):(m+K)}$, may be extracted as a sample of the mean and covariance of the returns. Shrinkage (or model averaging) in the regression model enables the columns in $[\mathbf{F}_j\ \mathbf{Y}_j]$, be they factors or returns, that are least useful for predicting $\mathbf{y}_j$ to be down-weighted. That way, rather than having one parameter governing the trade-off, like α in Eq. (16), the $m-1$ Bayesian shrinkage regressions can choose the right balance of factors and returns for each asset. Prior knowledge that many assets will be independent when conditioned upon appropriate factors may reasonably translate into small π (controlling the level of sparsity) encoding a preference for a small proportion of nonzero components of $\boldsymbol{\beta}$ in the $\boldsymbol{\phi}$-space regressions.

4.3 Balancing portfolios and accounting for estimation risk

A portfolio is balanced by choosing m weights $\mathbf{w}$ describing the portion of the portfolio invested in each asset. A standard technique uses the mean and the covariance between

returns to obtain a *mean–variance efficient portfolio* (Markowitz 1959) by solving a quadratic program (QP). Common formulations include the following.[3]

A so-called *minimum variance* portfolio may be obtained by solving

$$\text{argmin}_{\mathbf{w}} \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}, \quad \text{subject to} \quad \mathbf{w}^\top \mathbf{1} = 1. \tag{18}$$

Typical extensions include capping the weights, e.g., $0 \leq w_j \leq 2/m$, for $j = 1, \ldots, m$.

The above formulation may be augmented to involve the estimated mean return. One way is to aim for a minimum expected return μ while minimizing the variance of the portfolio:

$$\text{argmin}_{\mathbf{w}} \mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}, \quad \text{subject to} \quad \mathbf{w}^\top \boldsymbol{\mu} \geq \mu, \quad \text{and} \quad \mathbf{w}^\top \mathbf{1} = 1. \tag{19}$$

Similar heuristic augmentations apply here as well. A common extension is to assume that there is a risk-free asset available, e.g., a Treasury bond, at rate of return R_f. Then the constraints may be relaxed to $\boldsymbol{\mu} \geq \mu + (1 - \mathbf{w}^\top \mathbf{1}) R_f \geq \mu$ and $\mathbf{w}^\top \mathbf{1} \leq 1$.

Given $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ the solutions to these QPs, which are strictly convex, are essentially trivial to obtain. Gramacy et al. (2008) showed that when shrinkage-based MLEs $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$—constructed from all available returns via the monotone factorized likelihood—are used to balance portfolios in this way, they outperform a wealth of alternatives based upon estimators that could only use the completely observed instances. In the Bayesian context there are several ready extensions to this approach. For example, the MAP parameterization can be used to balance the portfolio. Another sensible option is to use the posterior mean of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which we show empirically leads to improved estimators of a true (known) generating distribution [Section 4.4], and to improved portfolios [Section 5]. But the portfolio balancing problem is about choosing at time t, say, weights to maximize expected return and/or minimize variance under the posterior predictive distribution $p(\mathbf{y}^{(t+1)}|\mathbf{Y}^{(t)})$. Here $\mathbf{Y}^{(t)} \equiv \mathbf{Y}_{1:n,:}$ represents the returns available up to time t, and $\mathbf{y}^{(t+1)}$ is the vector of returns at time $t+1$.[4] *Parameter uncertainty* (a.k.a., estimation risk) is taken into account by integration (Zellner and Chetty 1965; Klein and Bawa 1976):

$$p(\mathbf{y}^{(t+1)}|\mathbf{Y}^{(t)}) = \int p(\mathbf{y}^{(t+1)}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{Y}^{(t)}) \, d\boldsymbol{\mu} d\boldsymbol{\Sigma}. \tag{20}$$

Calculating this integral (or its moments, for use in the QP) is not, in general, tractable. However, with i.i.d. MVN returns (or another elliptical distribution) the moments are easily obtained (Polson and Tew 2000) as $\boldsymbol{\mu}^{(t+1)} = E\{\mathbf{y}^{(t+1)}|\mathbf{Y}^{(t)}\} = E\{\boldsymbol{\mu}|\mathbf{Y}^{(t)}\}$ and $\boldsymbol{\Sigma}^{(t+1)} = E\{\boldsymbol{\Sigma}|\mathbf{Y}^{(t)}\} + \text{Var}\{\boldsymbol{\mu}|\mathbf{Y}^{(t)}\}$, i.e., via a conditional variance identity.

In the case of completely observed returns, the standard non-informative prior $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-\left(\frac{m+1}{2}\right)}$, and (importantly) more returns than assets ($n > m$), Polson and Tew (2000) show that there is "*no effect* of parameter uncertainty on the portfolio

[3] In all cases we have the tacit constraint that $0 \leq w_j \leq 1$, for all $j = 1, \ldots, m$.

[4] The i.i.d. assumptions erode the meaning of time. Notionally, t runs counter to $i = 1, \ldots, n$.

rule" since $\boldsymbol{\mu}^{(t+1)} = \hat{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}^{(t+1)} = c\hat{\boldsymbol{\Sigma}}$, where the constant c is available in closed form and is a function of n and m only. In the case of historical returns of varying length, and when $n \gg m$, Stambaugh (1997) shows that we again have that $\boldsymbol{\mu}^{(t+1)} = \hat{\boldsymbol{\mu}}$, and that $\boldsymbol{\Sigma}^{(t+1)}$ is available in closed form but is not a scalar multiple of $\hat{\boldsymbol{\Sigma}}$. It can be shown empirically that incorporating this parameter uncertainty (a.k.a., estimation risk) leads to improved investments.

We are motivated by the situations in which these analytical approaches do not apply, i.e., when $m \geq n$ or when $n_j \leq j$ for any $j = 1, \ldots, m$. Although Gramacy et al. (2008) extended the MLE approach to the $n_j \leq j$ setting by employing parsimonious regressions, accounting for estimation risk remained illusive. The problem is best exposed in the calculation of Stambaugh's $\tilde{V}$ in Eq. (69–71), pp. 302, where the resulting diagonal is negative when $m > n$. But under the fully Bayesian approach, where samples of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ may be taken from the posterior, the above conditional variance identity may be used to approximate $\boldsymbol{\Sigma}^{(t+1)}$ with arbitrary precision.

4.4 Empirical results and comparisons

As a first point of comparison we pit the MLE based point-estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ against the Bayesian alternative: posterior expectations. We simulated synthetic data from known $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, imposed a uniformly random monotone missingness pattern, and then calculated the expected (predictive) log likelihood (ELL) of the so-parameterized MVN distribution(s). The ELL of data sampled from a density p relative to a density q (usually estimated) is given by $E_p\{\log q\} = \int p(x) \log q(x)\, dx = H(p) - D_{\mathrm{KL}}(q \parallel p)$, where $H(p) = \int p \log p$ is the entropy of p, and $D_{\mathrm{KL}}(q \parallel p)$ is the Kullback–Leibler (KL) divergence between q and p. The entropy and KL divergence are known in closed form for MVN densities p and q. When q uses point-estimates $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, and p uses the truth $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the ELL is given by:

$$-\frac{1}{2}\log\{(2\pi e)^N |\boldsymbol{\Sigma}|\} - \frac{1}{2}\left(\log\frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}|} + \mathrm{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})\right). \tag{21}$$

Table 1 contains summary information about the relative rankings of the ELL calculations for nine Bayesian and nine MLE estimators in a series of 50 repeated experiments. The results in the top portion of the table are for random MVN parameterizations obtained using the `randmvn` function from the `monomvn` package, with argument `method="normwish"`, and $m = n = 100$. Uniform monotone missingness patterns are then obtained with `rmono`. The use of parsimonious $\boldsymbol{\phi}$-space regressions (lasso, NG, ridge) with model selection via RJ-MCMC is determined by the setting of $\delta \in [0, 1)$. A parsimonious regression is performed if $\delta n_j < j$, and OLS is used otherwise—OLS regression is never used when $\delta = 0$. Observe from this portion of the table that the Bayesian estimators always obtain a better rank than the MLE ones. In the Bayesian case, ridge regression nudges out the lasso/NG. This makes sense because the data generation method (`"normwish"`) does not allow for any marginal or conditional independencies, i.e., zeros in $\boldsymbol{\Sigma}$ or $\boldsymbol{\Sigma}^{-1}$. The situation is reversed for the MLE, where the

	Fully Bayesian									MLE/CV					
	NG			Lasso			Ridge			Lasso			Ridge		
δ	0.9	0.2	0	0.9	0.2	0	0.9	0.2	0	0.9	0.2	0	0.9	0.2	0
	`normwish`, $m = 100$, $n = 100$														
min	8	3	3	8	3	3	3	1	1	10	10	10	12	11	10
avg	9	4	5	9	4	5	7	1	12	13	11	12	14	13	12
max	9	7	7	9	6	7	7	2	2	15	15	15	15	14	15
	`parsimonious`, $m = 100$, $n = 100$														
min	7	2	1	7	3	1	9	5	5	10	7	6	13	13	12
avg	8	3	1	8	4	2	10	6	5	12	10	10	15	14	13
max	10	4	2	10	4	3	11	7	6	15	15	12	15	14	15
	`normwish`, $m = 100$, $n = 1000$														
min	7	1	3	7	1	3	5	1	1	10	10	10	12	11	10
avg	8	3	5	9	3	6	7	1	2	14	12	12	14	13	12
max	9	6	6	9	6	7	9	4	4	15	15	15	15	14	13
	`parsimonious`, $m = 100$, $n = 1000$														
min	7	4	1	7	3	1	7	6	3	11	7	4	12	11	10
avg	9	4	1	9	5	2	10	6	3	12	9	8	15	14	13
max	11	6	2	11	5	2	11	7	4	15	15	15	15	14	13

Table 1: Summary of the rankings (by ELL (21)) of fifteen parsimonious regression methods used in the monotone MVN algorithm on 50 randomly generated MVN parameterizations, via two data generation "`methods`".

lasso comes out on top. In all cases but the MLE/CV lasso implementation, a lower value of δ gives improved performance, indicating that parsimonious regressions yield improvements even when they are not strictly necessary. However, lower δ comes with a higher computational cost. Given that the improvement of $\delta = 0$ over $\delta = 0.2$ is slight, the higher setting may be preferred. Finally, observe that the Bayesian lasso edges out the NG.

The results of a similar experiment, except using `method="parsimonious"`, are summarized in the second portion of the table. This data generation mechanism allows for conditional and marginal independence in the data by building up $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ sequentially, via randomly generating $\boldsymbol{\beta}_j$ (possibly having zero-entries) and applying Φ^{-1} as in Eq. (1). Here the number of nonzero entries of (each) $\boldsymbol{\beta}_j$ follows a $\text{Bin}(j, 0.1)$ distribution. In this case the results in the table indicate that the lasso/NG is the winner. The distinction between Bayesian and MLE/CV methods is as before. This time, NG edges out the Bayesian lasso. Finally, the bottom two portions of the table report ranks for the same two experiments described above, but using $n = 1000$ so that parsimonious regressions are needed less often. These results are similar to the $n = 100$ case except that the distinction between the Bayesian and MLE/CV ranks is blurred somewhat.

5 Asset management by portfolio balancing

Here we return to the motivating asset management problem from Section 1. We examine the characteristics of minimum variance portfolios (19) constructed using estimates

of $\mathbf{\Sigma}$ based upon historical monthly returns through a Monte Carlo experiment of repeated investment exercises. The experimental setup closely mirrors the one used by Gramacy et al. (2008), modeled after Chan et al. (1999). The data consist of returns of common domestic stocks traded on the NYSE and the AMEX from April 1968 until 1998 that have a share price greater than \$5 and a market capitalization greater than 20% based on the size distribution of NYSE firms. All such "qualifying stocks" are used—not just ones that survived to 1998. Since the i.i.d. assumption is only valid locally (in time) due to the conditional heteroskedastic nature of financial returns, estimators of $\mathbf{\Sigma}$ are constructed based upon (at most) the most recently available 60 months of historical returns. Short selling is not allowed; all portfolio weights must be non-negative. Although practitioners often impose a heuristic cap on the weights of balanced portfolios, e.g., at 2%, in order to "tame occasional bold forecasts" (Chan et al. 1999) or to "curb the effects ... of poor estimators" (Jagannathan and Ma 2003), we specifically do not do so here in order to fully expose the relative qualities of the estimators in question.

Our analysis in Section 3.3 suggests that the benefit of modeling Student-t errors will lead to minor improvements in our estimators based upon just 60 or fewer historical returns. Jacquier et al. (2004) showed that Bayes factors give strong preference to (the simpler) normal model over the Student-t at return frequencies less than weekly—supporting the *aggregation normality* of monthly returns. Models with Student-t errors were included in initial versions of our exercise, but they performed no better than their normal counterparts. This, and the extra computational burden required by the extra $O(n)$ extra latent variables, led us to exclude the Student-t comparators from the experiment reported on below.

The Monte Carlo experiment consists of 50 random repeated paths through 26 years, starting in April 1972. In each year 250 "qualifying stocks" with at least 12 months of historical returns are chosen randomly without replacement. Using at most the last 60 returns of the 250 assets, estimates of the covariance matrix $\mathbf{\Sigma}$ of monthly excess returns (over the monthly Treasury Bill rate) are calculated under our various methods and used to construct minimum variance portfolios. The portfolios are then held (fixed) for the year. To assess their quality and characteristics we follow Chan et al. (1999) in using the following: (annualized) mean return and standard deviation; (annualized) Sharpe ratio (average return in excess of the Treasury bill rate divided by the standard deviation); (annualized) tracking error (standard deviation of the return in excess of the S&P500); correlation to the market (S&P500); average number of stocks with weights above 0.5%. Generally speaking, portfolios with high mean return and low standard deviation, i.e., with large Sharpe ratio, are preferred. Sharpe ratios being roughly equal, we prefer those with lower tracking error.

Table 2 summarizes the results. It is broken into five sections, vertically. The first section gives results for the equal- and value-weighted portfolios. The second section uses standard estimators of $\mathbf{\Sigma}$ based only upon the complete data. The "min" estimator uses only the last 12 months of historical returns, whereas "com" uses the maximal complete history available. Both use standard estimators (i.e., via the `cov` function in R). The "rm" estimator is similar but discards any assets without the full 60 months of historical returns. These three rows in the table highlight that the more historical

	mean	sd	sharpe	te	cm	wmin
eq	0.149	0.188	0.431	0.063	0.949	0
vw	0.134	0.162	0.404	0.032	0.980	45
com	0.151	0.182	0.457	0.107	0.812	26
min	0.150	0.183	0.448	0.106	0.816	29
rm	0.131	0.130	0.486	0.095	0.802	16
fmin	0.141	0.146	0.498	0.085	0.844	39
fcom	0.143	0.146	0.509	0.087	0.840	38
frm	0.136	0.130	0.519	0.117	0.685	21
ridge	0.158	0.165	0.540	0.122	0.717	18
bridge	0.140	0.129	0.554	0.089	0.829	27
lasso	0.149	0.149	0.543	0.054	0.940	69
blasso	0.144	0.136	0.561	0.078	0.871	39
bng	0.144	0.136	0.560	0.078	0.872	39
fridge	0.158	0.164	0.549	0.121	0.719	19
bfridge	0.142	0.129	0.571	0.085	0.846	34
flasso	0.150	0.148	0.552	0.056	0.935	69
bflasso	0.148	0.138	0.573	0.070	0.898	51
bfng	0.148	0.138	0.574	0.071	0.896	50

Table 2: Comparing statistics summarizing the returns of yearly buy-and-hold portfolios generated over 50 repeated random paths through the 26 years of monthly historical returns.

returns (within the five-year window) that can be used to estimate covariances the better. The third section, containing the same acronyms with a leading "f", incorporates the value-weighted factor on same subset of returns. The improved characteristics in the table show that good factors can be quite helpful.

The results are further improved when the estimators exploit the tractable factorization of the likelihood under the monotone missingness pattern using shrinkage regression ($\delta = 0.2$), as shown in the penultimate section of the table. Notice that the Sharpe ratios indicate that the fully Bayesian estimators (with a "b" prefix) outperform the classical alternative, and that the lasso/NG methods are better than the ridge. In addition to fully accounting for all posterior uncertainties, the Bayesian estimators have the advantage of being able to deal with "gaps" in the data, via MDA [Section 4.1], and can account for estimation risk [Section 4.3]. Observe that these estimators distribute the weight less evenly among the assets, having fewer assets with $\geq 0.5\%$ of the weight on average compared to their classical counterparts. The higher concentration of weight on the appropriate assets leads to lower variance portfolios which deviate further from the market, hence the somewhat higher tracking error. The results for the lasso are nearly identical to those under the extended NG formulation.

The final section of the table shows that incorporating the value-weighted factor (now with $\delta = 0$) leads to further improvements. A sensible (prior) belief that the

presence of a good factor causes many pairs of assets to be conditionally independent [Section 4.3] allows us to dial down the hierarchical prior on the proportion of nonzero regression coefficients: $\pi \sim \text{Beta}(1, 100)$. The results in the table suggest that the factor causes weight to be more evenly distributed in the case of the Bayesian estimators, but not the classical ones.

The first eight rows (first three sections) of the table, and those corresponding to "lasso", "ridge", "flasso" and "fridge", are nearly identical to ones in a similar table from Gramacy et al. (2008). Any variation is due to different random seeds. This calibration allows us to draw comparisons to the other classical `monomvn` estimators including ones based upon PCR, etc. In short, the fully Bayesian approach(es) reign supreme. The improvements may appear to be modest at a glance. But in light of the fact that financial markets are highly unpredictable they are actually quite substantial. As a matter of curiosity we also calculated statistics under the posterior mean parameterization, i.e., without accounting for estimation risk. The Sharpe ratios were: 0.549 (0.562 with the factor) under the ridge, 0.554 (0.562) under the lasso, and 0.553 (0.563) under the NG. These numbers point to an improvement over the classical approach using the posterior mean, but indicate that the incorporation of estimation risk is crucial to get the best portfolio weights.

Figure 4 summarizes the variability in the Monte Carlo experiment showing the distribution (via boxplots) of the Sharpe ratios and tracking error obtained for each of the 50 random paths through the 26 years, thereby complementing the averages presented in Table 2. We can immediately see that the classical ridge regression approach is highly variable, often yielding extremely low Sharpe ratios and high tracking error. The Bayesian approach offers a dramatic improvement here. In the case of the lasso/NG we can see that the variability of the Sharpe ratios for the Bayesian implementations are higher than their classical counterparts. However, it is crucial to observe that the boxplots extend in the direction of larger Sharpe ratios, offering improved estimators.

6 Discussion

We have shown how the classical (MLE/CV) shrinkage approach of Gramacy et al. (2008) to joint multivariate inference under monotone missingness may be treated in a fully Bayesian way to great effect. The Bayesian approach facilitates extensions to deal with "gaps" in the monotone pattern via MDA and heavy-tailed data, and can account for estimation risk. None of these features could be accommodated by the classical approach. In synthetic data experiments we demonstrated the descriptive, predictive, and inferential superiority of the Bayesian methods. Using real financial returns data we showed how the fully Bayesian approach leads to portfolios with lower variability and thus higher Sharpe ratio.

Our methods bear some similarity to other recent approaches to covariance estimation. Levina et al. (2008) and Carvalho and Scott (2009) offer priors on covariance matrices with shrinkage based on Cholesky decompositions. Our monotone factorization (originally: Andersen 1957) can be seen as a special case where the column order,

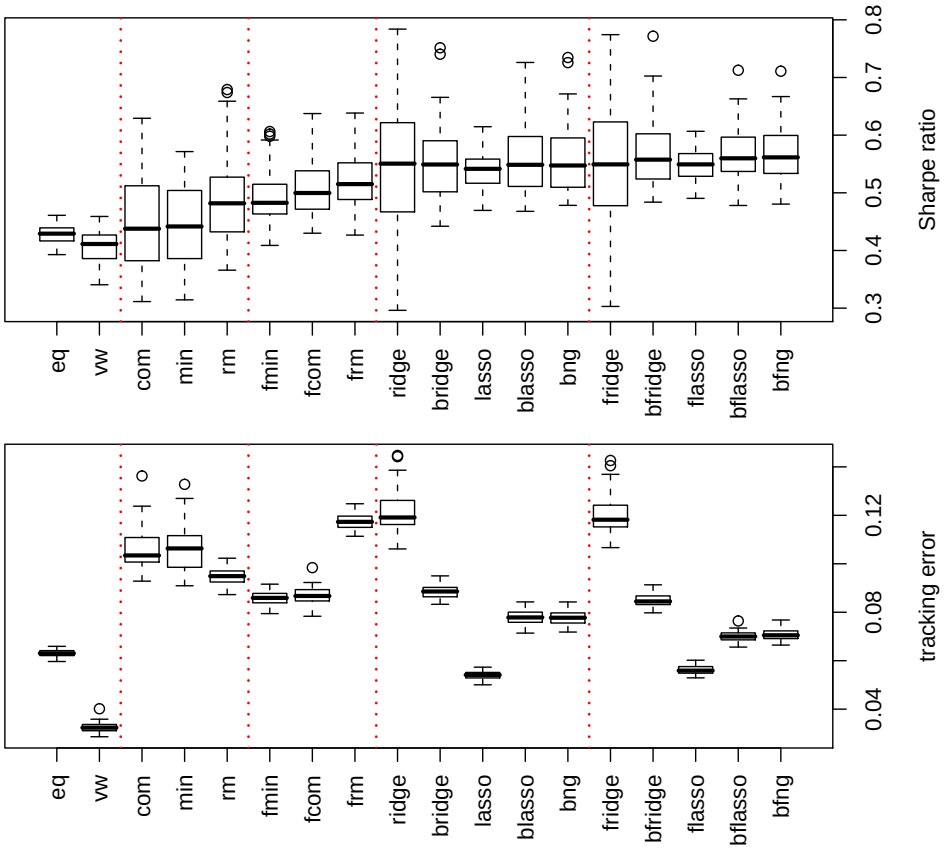

Figure 4: Boxplots of Sharpe ratios *(top)* and the tracking error *(bottom)* obtained over 50 random paths through 26 years. The vertical bars correspond to horizontal ones in Table 2.

and thus the underlying graphical model structure, is fixed by historical availability. This has certain advantages in our context (relative inferential simplicity, tractability, MDA and heavy tail extensions), but it may not be optimal in complete data cases. Liu (1996, 1995) provides a model for Bayesian robust multivariate joint inference for data exhibiting a monotone missingness pattern. The posteriors are derived using extensions of Bartlett's decomposition, and "gaps" may be handled via MDA. Although a wealth of theoretical and empirical results are provided, it is not clear how the methods can be adapted to "big p small n" setting. A possible way forward involves Bayesian dynamic factor models (West 2003) which are designed for "big p small n" and retain the ability to handle missing data in a tractable way.

It is becoming well understood that the Laplace prior has many drawbacks, even

when generalized by the NG. For example, it is known to produce biased estimates of the nonzero coefficients and to underestimate of the number of zeros. A newly developed shrinkage prior for regression called the horseshoe (Carvalho et al. 2008) shows promise as a tractable alternative (i.e., via GS) without the bias problems. Incorporating horseshoe regression in our framework is part of our future work. Another obvious extension is to relax the i.i.d. assumption to obtain more dynamic estimators. One possible approach might be to use weighted regressions for the $\mathcal{BR}$'s with weights decaying back in time.

References

Andersen, T. (1957). "Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations Are Missing." *J. of the American Statistical Association*, 52, 200–203. 238, 239, 257

Brown, S. (1979). "The Effect of Estimation Risk on Capital Market Equilibrium." *J. of Financial and Quantitative Analysis*, 14, 215–220. 241

Carlin, B. P. and Polson, N. G. (1991). "Inference for nonconjugate Bayesian Models using the Gibbs sampler." *The Canadian Journal of Statistics*, 19, 4, 399–405. 242

Carlin, B. P., Polson, N. G., and Stoffer, D. S. (1992). "A Monte Carlo Approach to Nonnormal and Nonlinear State–Space Modeling." *J. of the American Statistical Association*, 87, 418, 493–500. 245

Carvalho, C., Polson, N., and Scott, J. (2008). "The horseshoe estimator for sparse signals." Discussion Paper 2008-31, Duke University Department of Statistical Science. 259

Carvalho, C. M. and Scott, J. G. (2009). "Objective Bayesian model selection in Gaussian graphical models." *Biometrika*, 96, 3, 497–512. 257

Chan, L. K., Karceski, J., and Lakonishok, J. (1999). "On Portfolio Optimization: Forecasting Covariances and Choosing the Risk Model." *The Review of Financial Studies*, 12, 5, 937–974. 239, 250, 255

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). "Least Angle Regression (with discussion)." *Annals of Statistics*, 32, 2. 241

Fama, E. and French, K. (1993). "Common Risk Factors in the Returns on Stocks and Bonds." *J. of Financial Economics*, 33, 3–56. 250

George, E. and McCulloch, R. (1993). "Variable selection via Gibbs sampling." *J. of the American Statistical Association*, 88, 881–889. 244, 245

Geweke, J. (1992). "Priors for microeconomic times series and their application." Tech. Rep. Institute of Empirical Macroeconomics Discussion Paper No.64, Federal Reserve Bank of Minneapolis. 246

— (1993). "Bayesian Treatment of the Independent Student–t Linear Model." *J. of Applied Econometrics*, Vol. 8, Supplement: Special Issue on Econometric Inference Using Simulation Techniques, S19–S40. 245, 246

— (1996). "Variable selection and model comparison in regression." In *Bayesian Statistics 5*, eds. J. Bernardo, J. Berger, A. Dawid, and A. Smith, 609–620. Oxford Press. 244

Godsill, S. (2001). "On the relationship between Markov chain Monte Carlo methods for model uncertainty." *J. of Computational and Graphical Statistics*, 10, 2, 239–248. 247

Gramacy, R. B. (2009). *The* `monomvn` *package: Estimation for multivariate normal and Student-t data with monotone missingness*. R package version 1.8. 239

Gramacy, R. B., Lee, J. H., and Silva, R. (2008). "On estimating covariances between many assets with histories of highly variable length." Tech. Rep. 0710.5837, arXiv. Url: http://arxiv.org/abs/0710.5837. 237, 238, 241, 242, 251, 252, 253, 255, 257

Green, P. (1995). "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika*, 82, 711–732. 244

Griffin, J. E. and Brown, P. J. (2010). "Inference with Normal–Gamma prior distributions in regression problems." *Bayesian Analysis*, 5, 1, 171–188. 242, 243, 247

Hans, C. (2008). "Bayesian lasso regression." Tech. Rep. 810, Department of Statistics, The Ohio State University, Columbus, OH 43210. 243, 244, 245, 247

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. 241, 242

Hoerl, A. and Kennard, R. (1970). "Ridge Regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12, 55–67. 241

Jacquier, E., Polson, N., and Rossi, P. E. (2004). "Bayesian analysis of stochastic volatility models with fat-tails and correlated errors." *J. of Econometrics*, 122, 185–212. 247, 255

Jagannathan, R. and Ma, T. (2003). "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps." *J. of Finance*, 58, 4, 1641–1684. 239, 255

Klein, R. and Bawa, V. (1976). "The Effect of Estimation Risk on Optimal Portfolio Choice." *J. of Financial Econometrics*, 3, 215–231. 241, 252

Ledoit, O. and Wolf, M. (2002). "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection." *J. of Emperical Finance*, 10, 603–621. 251

Levina, E., Rothman, A., and Zhu, J. (2008). "Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty." *Annals of Applied Statistics*, 2, 1, 245–263. 257

Li, K. (1988). "Imputation using Markov chains." *J. of Statistical Computation*, 30, 57–79. 250

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley. 237, 239

Liu, C. (1993). "Bartlett's decomposition of the posterior distribution of the covariance for normal monotone ignorable missing data." *J. of Multivariate Analysis*, 46, 198–206. 241

— (1995). "Monotone Data Augmentation Using the Multivariate t Distribution." *J. of Multivariate Analysis*, 53, 139–158. 258

— (1996). "Bayesian Robust Multivariate Linear Regression With Incomplete Data." *J. of the American Statistical Association*, 91, 435, 1219–1227. 258

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York: John Wiley. 237, 252

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *J. of the American Statistical Association*, 103, 482, 681–686. 242, 243, 244

Polson, N. and Tew, B. (2000). "Bayesian Portfolio Selection: An Empirical Analysis of the S&P 500 Index 1970–1996." *J. of Business & Economic Statistics*, 18, 2, 164–173. 252

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC. 237, 239, 241, 250

R Development Core Team (2007). R*: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. 239

Stambaugh, R. F. (1997). "Analyzing Investments Whose Histories Differ in Lengh." *J. of Financial Economics*, 45, 285–331. 238, 239, 241, 253

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *J. of the Royal Statistical Society, Series B*, 58, 267–288. 241

Troughton, P. T. and Godsill, S. J. (1997). "A reversible jump sampler for autoregressive time series, employing full conditionals to achieve efficient model space moves." Tech. Rep. CUED/F-INFENG/TR.304, Cambridge University Engineering Department. 244

West, M. (2003). "Bayesian factor regression models in the "large p, small n" paradigm." *Bayesian Statistics 7*, 723–732. 258

Zellner, A. and Chetty, V. (1965). "Prediction and Decision Problems in Regression Models From the Bayesian Point of View." *J. of the American Statistical Association*, 605–616. 252

Acknowledgments

This work was partially supported by Engineering and Physical Sciences Research Council Grant EP/D065704/1. We would like to thank an anonymous referee, and the associate editor, whose many helpful comments improved the paper