Bayesian Analysis (2017) **12**, Number 4, pp. 969–989

Variable Selection in Seemingly Unrelated Regressions with Random Predictors

David Puelz*, P. Richard Hahn†, and Carlos M. Carvalho‡

Abstract. This paper considers linear model selection when the response is vector-valued and the predictors, either all or some, are randomly observed. We propose a new approach that decouples statistical inference from the selection step in a "post-inference model summarization" strategy. We study the impact of predictor uncertainty on the model selection procedure. The method is demonstrated through an application to asset pricing.

Keywords: decoupling shrinkage and selection, seemingly unrelated regressions, penalized utility selection.

1 Introduction and overview

This paper develops a method for parsimoniously summarizing the shared dependence of many individual response variables upon a common set of predictor variables drawn at random. The focus is on multivariate Gaussian linear models where an analyst wants to find, among p available predictors X, a subset which work well for predicting $q > 1$ response variables Y. The multivariate normal linear model assumes that a set of responses $\{Y_j\}_{j=1}^q$ are linearly related to a shared set of covariates $\{X_i\}_{i=1}^p$ via

$$Y_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p + \epsilon_j, \qquad \boldsymbol{\epsilon} \sim \mathrm{N}(0, \Psi), \tag{1}$$

where Ψ is a non-diagonal covariance matrix.

Bayesian variable selection in (single-response) linear models is the subject of a vast literature, from prior specification on parameters (Bayarri et al., 2012) and models (Scott and Berger, 2006) to efficient search strategies over the model space (George and McCulloch, 1993; Hans et al., 2007). For a more complete set of references, we refer the reader to the reviews of Clyde and George (2004) and Hahn and Carvalho (2015). By comparison, variable selection has not been widely studied in concurrent regression models, perhaps because it is natural simply to apply existing variable selection methods to each univariate regression individually. Indeed, such joint regression models go by the name "seemingly unrelated regressions" (SUR) in the Bayesian econometrics literature, reflecting the fact that the regression coefficients from each of the separate regressions can be obtained in isolation from one another (i.e., conducting estimation

*University of Texas McCombs School of Business, 2110 Speedway, Austin, Texas 78705, david.puelz@utexas.edu, url: http://faculty.mccombs.utexas.edu/David.Puelz/

†University of Chicago Booth School of Business, Harper Center 368, Chicago, IL, 80637, richard.hahn@chicagobooth.edu

‡University of Texas McCombs School of Business, 2110 Speedway, Austin, Texas 78705, carlos.carvalho@mccombs.utexas.edu

 DOI: 10.1214/17-BA1053

as if Ψ were diagonal). However, allowing non-diagonal Ψ can lead to more efficient estimation (Zellner, 1962) and can similarly impact variable selection (Brown et al., 1998; Wang, 2010).

This paper differs from Brown et al. (1998) and Wang (2010) in that we focus on the case where all or some of the predictor variables (the regressors, or covariates) are treated as random as opposed to fixed. Our goal will be to summarize codependence among multiple responses in *subsequent* periods, making the uncertainty in future realizations highly central to our selection objective. This approach is natural in many contexts (e.g., macroeconomic models) where the purpose of selection is inherently forward-looking. Measurement errors in the predictors may also contribute to randomness, so this approach is naturally applicable in an "errors-in-variables" context. To our knowledge, no existing variable selection methods are suitable in this context. The new approach is based on the sparse summary perspective outlined in Hahn and Carvalho (2015), which applies Bayesian decision theory to summarize complex posterior distributions. By using a utility function that explicitly rewards sparse summaries, a high dimensional posterior distribution is collapsed into a more interpretable sequence of sparse point summaries.

A related approach to variable selection in multivariate Gaussian models is the Gaussian graphical model framework (Jones et al., 2005; Dobra et al., 2004; Wang and West, 2009). In that approach, the full conditional distributions are represented in terms of a sparse $(p+q)$-by-$(p+q)$ precision matrix. By contrast, we partition the model into response and predictor variable blocks, leading to a distinct selection criterion that narrowly considers the p-by-q covariance between Y to X.

This paper is structured as follows. Section 2 describes the methodology. A methods overview is presented followed by three subsections discussing the details of the approach. Section 3 presents a simulation study utilizing the methodology with comparisons to alternative approaches. Section 4 demonstrates how the methodology works in practice by considering an application in asset pricing. Section 5 concludes.

2 Posterior summary variable selection

2.1 Methods overview

Posterior summary variable selection consists of three phases: *model specification and fitting*, *utility specification*, and *graphical summary*. Each of these steps is outlined below. Additional details of the implementation are described in Supplementary Appendices A and B (Puelz et al., 2017).

Step 1: Model specification and fitting

The statistical model may be described compositionally as $p(Y, X) = p(Y|X)p(X)$. For $(Y, X) \sim \mathrm{N}(\mu, \Sigma)$, the regression model (1) implies Σ has the following block structure:

$$\Sigma = \left[\begin{array}{c|c} \boldsymbol{\beta}^T \Sigma_x \boldsymbol{\beta} + \Psi & (\Sigma_x \boldsymbol{\beta})^T \\ \hline \Sigma_x \boldsymbol{\beta} & \Sigma_x \end{array}\right]. \tag{2}$$

We denote the unknown parameters for the full joint model as $\Theta = \{\mu_x, \mu_y, \Sigma_x, \boldsymbol{\beta}, \Psi\}$ where $\mu = (\mu_y^T, \mu_x^T)^T$ and $\Sigma_x = \text{cov}(X)$.

For a given prior choice $p(\Theta)$, posterior samples of all model parameters are computed by routine Monte Carlo methods, primarily Gibbs sampling. Details of the specific modeling choices and associated posterior sampling strategies are described in Supplementary Appendix A.

A notable feature of our approach is that *steps 2* (and *3*) will be unaffected by modeling choices made in *step 1* except insofar as they lead to different posterior distributions. In short, *step 1* is "obtain a posterior distribution"; posterior samples then become inputs to *step 2*.

Step 2: Utility specification

For our utility function we use the log-density of the regression $p(Y|X)$ above. It is convenient to work in terms of negative utility, or loss:

$$\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) = \frac{1}{2}(\tilde{Y} - \boldsymbol{\gamma}\tilde{X})^T \Omega (\tilde{Y} - \boldsymbol{\gamma}\tilde{X}),$$

where $\Omega = \Psi^{-1}$. Note that this log-density is being used in a descriptive capacity, not an inferential one; that is, all posterior inferences are based on the posterior distribution from *step 1*. The "action" $\boldsymbol{\gamma}$ is regarded as a point estimate of the regression parameters $\boldsymbol{\beta}$, which would be a good fit to *future* data $(\tilde{Y}, \tilde{X})$ drawn from the same model as the observed data given by $\mathbf{Y} \in \mathbb{R}^{N \times q}$ and $\mathbf{X} \in \mathbb{R}^{N \times p}$.

Taking expectations over the posterior distribution of all unknowns,

$$p(\tilde{Y}, \tilde{X}, \Theta | \mathbf{Y}, \mathbf{X}) = p(\tilde{Y}|\tilde{X}, \Theta) p(\tilde{X}|\Theta) p(\Theta|\mathbf{Y}, \mathbf{X})$$

yields expected loss

$$\mathcal{L}(\boldsymbol{\gamma}) \equiv \mathbb{E}[\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})] = \text{tr}[M\boldsymbol{\gamma} S \boldsymbol{\gamma}^T] - 2\text{tr}[A\boldsymbol{\gamma}^T] + \text{constant},$$

where $A = \mathbb{E}[\Omega \tilde{Y} \tilde{X}^T]$, $S = \mathbb{E}[\tilde{X}\tilde{X}^T] = \overline{\Sigma_x}$, and $M = \overline{\Omega}$, the overlines denote posterior means, and the final term is a constant with respect to $\boldsymbol{\gamma}$.

Finally, we add an explicit penalty, reflecting our preference for sparse summaries:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) \equiv \text{tr}[M\boldsymbol{\gamma} S \boldsymbol{\gamma}^T] - 2\text{tr}[A\boldsymbol{\gamma}^T] + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1, \tag{3}$$

where $\|\mathbf{vec}(\boldsymbol{\gamma})\|_1$ sums the absolute values of components in $\mathbf{vec}(\boldsymbol{\gamma})$. In practice, it is well known that the ℓ_1 penalty selects relevant components by shrinking irrelevant ones to zero.

Step 3: Graphical summary

Traditional applications of Bayesian decision theory derive *point-estimates* by minimizing expected loss for certain loss functions. The present goal is not an *estimator* per se,

but a parsimonious summary of information contained in a complicated, high dimensional posterior distribution. This distinction is worth emphasizing because we have not one, but rather a continuum of loss functions, indexed by the penalty parameter λ. This class of loss functions can be used to summarize the posterior distribution as follows.

Using available convex optimization techniques, expression (3) can be optimized efficiently for a range of λ values simultaneously. Posterior graphical summaries consist of two components. First, graphs depicting which response variables have non-zero $\boldsymbol{\gamma}^*_\lambda$ coefficients on which predictor variables can be produced for any given λ. Second, posterior distributions of the quantity

$$\Delta_\lambda = \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*_\lambda) - \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*)$$

can be used to gauge the impact λ has on the descriptive capacity of γ^*_λ. Here, $\boldsymbol{\gamma}^* = \gamma^*_{\lambda=0}$ is the unpenalized optimal solution to the minimization of loss (3). Note that these graphs defined by γ^*_λ provide appropriate variable selection for SUR models – different sets of predictors are connected to each of the responses.

The statistical model is given in (1) and (2); prior specification and model fitting details can be found in Supplementary Appendix A. To briefly summarize, we use a multivariate version of the priors presented in George and McCulloch (1993) and similar to Brown et al. (1998) and Wang (2010) for the exercises in this paper. We choose the g-prior parameter using an empirical Bayes procedure, and the marginal distribution of the predictors is modeled via a Gaussian linear latent factor model. In the following three subsections, we flesh out the details of *steps 2* and *3*, which represent the main contributions of this paper.

2.2 Deriving the sparsifying expected utility function

Define the optimal posterior summary as the $\boldsymbol{\gamma}^*$ minimizing some expected loss $\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \mathbb{E}[\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})]$. Here, the expectation is taken over the joint posterior predictive and posterior distribution: $p(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$.

As described in the previous section, our loss takes the form of a penalized log conditional distribution:

$$\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) \equiv \frac{1}{2}(\tilde{Y} - \boldsymbol{\gamma}\tilde{X})^T \Omega (\tilde{Y} - \boldsymbol{\gamma}\tilde{X}) + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1, \tag{4}$$

where $\Omega = \Psi^{-1}$, $\left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1 = \sum_j |\mathbf{vec}(\boldsymbol{\gamma})_j|$, and $\mathbf{vec}(\boldsymbol{\gamma})$ is the vectorization of the action matrix $\boldsymbol{\gamma}$. The first term of this loss measures the distance (weighted by the precision Ω) between the linear predictor $\boldsymbol{\gamma}\tilde{X}$ and a future response $\tilde{Y}$. The second term promotes a sparse optimal summary, $\boldsymbol{\gamma}$. The penalty parameter λ determines the relative importance of these two components. Expanding the quadratic form gives:

$$\begin{aligned}
\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) &= \frac{1}{2}\left(\tilde{Y}^T\Omega\tilde{Y} - 2\tilde{X}^T\boldsymbol{\gamma}^T\Omega\tilde{Y} + \tilde{X}^T\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}\tilde{X}\right) + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1 \\
&= \left(\tilde{X}^T\boldsymbol{\gamma}^T\Omega\boldsymbol{\gamma}\tilde{X} - 2\tilde{X}^T\boldsymbol{\gamma}^T\Omega\tilde{Y}\right) + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1 + \text{constant}.
\end{aligned}$$

Integrating over $(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$ (and dropping the constant) gives:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) &= \mathbb{E}[\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma})] \\ &= \mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T \Omega \boldsymbol{\gamma} \tilde{X} \tilde{X}^T]\right] - 2\mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T \Omega \tilde{Y} \tilde{X}^T]\right] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1 \\ &= \mathbb{E}\left[\mathrm{tr}[\boldsymbol{\gamma}^T \Omega \boldsymbol{\gamma} S]\right] - 2\mathrm{tr}[A\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1 \\ &= \mathrm{tr}[M\boldsymbol{\gamma} S \boldsymbol{\gamma}^T] - 2\mathrm{tr}[A\boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,\end{aligned} \tag{5}$$

where:

$$\begin{aligned} A &\equiv \mathbb{E}[\Omega \tilde{Y} \tilde{X}^T], \\ S &\equiv \mathbb{E}[\tilde{X} \tilde{X}^T] = \overline{\Sigma}_x, \\ M &\equiv \overline{\Omega}, \end{aligned} \tag{6}$$

and the overlines denote posterior means. Define the Cholesky decompositions $M = LL^T$ and $S = QQ^T$. Expression (5) can be formulated in the form of a standard penalized regression problem:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \left\| \left[Q^T \otimes L^T\right] \mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1} A Q^{-T}) \right\|_2^2 + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1, \tag{7}$$

with "covariates" $Q^T \otimes L^T$, "data" $L^{-1} A Q^{-T}$, and regression coefficients $\boldsymbol{\gamma}$ (see Supplementary Appendix B for details). Accordingly, (7) can be optimized using existing software such as the `lars` R package of Efron et al. (2004) and still yield sparse solutions.

The `lars` formulation of the utility function provides fast computation as well as flexibility. For example, suppose we wish to always include certain (potentially different) predictors in each SUR equation. This can be easily achieved by removing the ℓ_1 penalty on the relevant components of $\mathbf{vec}(\boldsymbol{\gamma})$ (since $\boldsymbol{\gamma}$ represents the dependence structure between the responses and predictors) by zeroing the appropriate penalty parameters λ.

What if only a subset of predictors are random?

Building on the previous derivation, we consider a scenario where some predictors are known, or fixed, and the remainder are random. This may occur when one would like to condition on a particular value of a predictor at some fixed future value. In this case, an expected utility function can be derived in a similar manner to the random predictors case.

Let the covariates X be divided into two pieces, those that are considered random: $X_r \in \mathbb{R}^{p_r}$, and those that are considered fixed: $X_f \in \mathbb{R}^{p_f}$, so that the column vector $X = [X_r^T \; X_f^T]^T \in \mathbb{R}^p$ and $p = p_r + p_f$. So, future values of the covariates are given by $\tilde{X} = [\tilde{X}_r^T \; X_f^T]^T$.

Conditioning on the fixed covariates, the distribution of unknowns is: $p(\tilde{Y}, \tilde{X}_r, \Theta | X_f)$ where Θ is a vector of parameters from a specified model. If we assume conditional independence, then we can write:

$$p(\tilde{Y}, \tilde{X}_r, \Theta | X_f) = p(\tilde{Y} | \tilde{X}_r, X_f, \Theta) p(\tilde{X}_r | X_f, \Theta) p(\Theta | X_f),$$

where, as before, $p(\Theta|X_f)$ is the posterior distribution of model parameters conditional on the fixed covariates. Following *step 1* of the methodology, any models may be chosen for the conditional $Y|X_r, X_f$ and the marginal $X_r|X_f$. For example, in the case of X following a multivariate normal distribution implied by a latent factor regression model, we automatically know the conditionals including $X_r|X_f$.

Define the following block structure for the action, $\boldsymbol{\gamma}$:

$$\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_r & \boldsymbol{\gamma}_f \end{bmatrix},$$

so that $\boldsymbol{\gamma}_r \in \mathbb{R}^{q \times p_r}$ and $\boldsymbol{\gamma}_f \in \mathbb{R}^{q \times p_f}$. We expand out (4) and drop terms that don't involve the action $\boldsymbol{\gamma}$:

$$\begin{aligned}\mathcal{L}_\lambda(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}) =& \frac{1}{2}\left(\tilde{X}_r^T \boldsymbol{\gamma}_r^T \Omega \boldsymbol{\gamma}_r \tilde{X}_r + X_f^T \boldsymbol{\gamma}_f^T \Omega \boldsymbol{\gamma}_f X_f - 2\tilde{X}_r^T \boldsymbol{\gamma}_r^T \Omega \tilde{Y} - 2X_f^T \boldsymbol{\gamma}_f^T \Omega \tilde{Y}\right) \\ &+ \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1 + \text{constant}.\end{aligned}$$

Taking expectations over $p(\tilde{Y}, \tilde{X}_r, \Theta|X_f)$ and dropping the one-half and constant, we obtain the integrated loss function:

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) =& \mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_r^T \Omega \boldsymbol{\gamma}_r \tilde{X}_r \tilde{X}_r^T]\right] - 2\mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_r^T \Omega \tilde{Y} \tilde{X}_r^T]\right] + \mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_f^T \Omega \boldsymbol{\gamma}_f X_f X_f^T]\right] \\ &- 2\mathbb{E}\left[\text{tr}[\boldsymbol{\gamma}_f^T \Omega \tilde{Y} X_f^T]\right] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1.\end{aligned}$$

We simplify the expectations in a similar way to our derivation of the original loss function presented at the beginning of Section 2.2.

$$\begin{aligned}\mathcal{L}_\lambda(\boldsymbol{\gamma}) =& \text{tr}[M \boldsymbol{\gamma}_r S_r \boldsymbol{\gamma}_r^T] - 2\text{tr}[A_r \boldsymbol{\gamma}_r^T] + \text{tr}[M \boldsymbol{\gamma}_f S_f \boldsymbol{\gamma}_f^T] - 2\text{tr}[A_f \boldsymbol{\gamma}_f^T] \\ &+ \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,\end{aligned}$$

where:

$$\begin{aligned} &A_r \equiv \mathbb{E}[\Omega \tilde{Y} \tilde{X}_r^T], \quad A_f \equiv \mathbb{E}[\Omega \tilde{Y} \tilde{X}_f^T], \\ &S_r \equiv \mathbb{E}[\tilde{X}_r \tilde{X}_r^T], \quad S_f = X_f X_f^T, \\ &M \equiv \overline{\Omega}. \end{aligned}$$

Combining the matrix traces, we simplify the loss function as follows:

$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \text{tr}[M \boldsymbol{\gamma} \dot{S} \boldsymbol{\gamma}^T] - 2\text{tr}[\dot{A} \boldsymbol{\gamma}^T] + \lambda \left\|\mathbf{vec}(\boldsymbol{\gamma})\right\|_1,$$

where:

$$\dot{S} \equiv \begin{bmatrix} S_r & 0 \\ 0 & S_f \end{bmatrix}, \quad \dot{A} \equiv \begin{bmatrix} A_r \\ A_f \end{bmatrix}.$$

This form is analogous to the loss function derived when all predictors are assumed random, now for a case when fixed *and* random predictors are present. This can similarly be formulated into a penalized regression problem. The full derivation of the lasso form of this problem is presented in Supplementary Appendix B.

The effect on *step 3* of assuming only a subset of predictors are random is intuitive. Less statistical uncertainty will propagate into the Δ_λ metric, and the algorithm will favor denser graphs. This will be shown in the results section where we study two extremes: all random predictors and all fixed predictors.

2.3 Sparsity-utility trade-off plots

Rather than attempting to determine an "optimal" value of λ, we advocate displaying plots that reflect the utility attenuation due to λ-induced sparsification. We define the "loss gap" between a λ-sparse solution, $\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}_\lambda^*)$, and the optimal unpenalized (non-sparse, $\lambda = 0$) summary, $\mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*)$ as

$$\Delta_\lambda = \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}_\lambda^*) - \mathcal{L}(\tilde{Y}, \tilde{X}, \Theta, \boldsymbol{\gamma}^*).$$

As a function of $(\tilde{Y}, \tilde{X}, \Theta)$, Δ_λ is itself a random variable which we can sample by obtaining posterior draws from $p(\tilde{Y}, \tilde{X}, \Theta \mid \mathbf{Y}, \mathbf{X})$. The posterior distribution(s) of Δ_λ (for various λ) therefore reflects the deterioration in utility attributable to "sparsification". Plotting these distributions as a function of λ allows one to visualize this trade-off. Specifically, $\pi_\lambda \equiv \Pr(\Delta_\lambda < 0 \mid \mathbf{Y}, \mathbf{X})$ is the (posterior) probability that the λ-sparse summary is no worse than the non-sparse summary. Using this framework, a useful heuristic for obtaining a single sparse summary is to report the sparsest model (associated with the highest λ) such that π_λ is higher than some pre-determined threshold, κ; we adopt this approach in our application section.

We propose summarizing the posterior distribution of Δ_λ via two types of plots. First, one can examine posterior means and credible intervals of Δ_λ for a sequence of models indexed by λ. Similarly, one can plot π_λ across the same sequence of models. Also, for a fixed value of λ, one can produce graphs where nodes represent predictor variables and response variables and an edge is drawn between nodes whenever the corresponding element of γ_λ^* is non-zero. All three types of plots are exhibited in Section 4.

2.4 Relation to previous methods

Loss function (7) is similar in form to the univariate *DSS* (decoupled shrinkage and selection) strategy developed by Hahn and Carvalho (2015). Our approach generalizes Hahn and Carvalho (2015) by optimizing over the matrix $\boldsymbol{\gamma} \in \mathbb{R}^{qxp}$ rather than a single vector of regression coefficients, extending the sparse summary utility approach to seemingly unrelated regression models (Brown et al., 1998; Wang, 2010). Additionally, the present method considers random predictors, $\tilde{X}$, whereas Hahn and Carvalho (2015) considered only a matrix of fixed design points. The impact of accounting for random predictors on the posterior summary variable selection procedure is examined in more detail in the application section.

To the best of our knowledge, the most comparable method for analyzing sparse linear covariance structures are the SUR models described in Wang (2010), which utilize independent point-mass priors for each element of $\boldsymbol{\beta}$. Our method differs from this

approach for the following reasons. A sparse SUR model provides posterior draws of the coefficient matrix $\boldsymbol{\beta}$, but as in the simpler linear regression case described in Hahn and Carvalho (2015), obtaining a sparse summary is non-trivial. Two common approaches for extracting a summary from posterior draws of a sparse SUR model are either to report a maximum a posteriori estimate (MAP) or to hard-threshold posterior inclusion probabilities of matrix components of $\boldsymbol{\beta}$ describing the model sparsity pattern. Neither approach is fully satisfactory; the MAP estimate is not well-motivated if the goal is future prediction and approaches based on thresholding the edge inclusion probabilities fail to take into account co-dependence between individual edges coming in and out of the model together. By contrast, our method begins with a principled loss function; by focusing on the expected log-density of future predictions, our approach synthesizes information from all model parameters simultaneously in gauging how important they are for prediction. A comprehensive simulation comparing inclusion probability thresholding and our approach is presented in Section 3.

3 Simulation study

In this section, we present a simulation study to compare our posterior model selection summary to that of the median (posterior) probability model (using the model of Wang (2010)) for a fixed data generating process (DGP). This comparison is not meant to demonstrate the superiority of one method over the other, but rather to highlight that the two methods can give quite distinct summaries. More specifically, we observe that the median probability model (MPM) can differ substantially from our penalized utility summary when the predictor variables are highly correlated.

The data generating process is the following two equation seemingly unrelated regression model:

$$\begin{aligned} X &\sim N(0, \Sigma_x^{\text{sim}}), \\ Y|X &\sim N(\boldsymbol{\beta}_{\text{sim}}^T X, \Psi^{\text{sim}}), \end{aligned} \tag{8}$$

where:

$$\Sigma_x^{\text{sim}} = \begin{bmatrix} 1 & 0 & 0.9 & 0 & 0 \\ 0 & 1 & 0 & 0.8 & -0.3 \\ 0.9 & 0 & 1 & 0 & 0 \\ 0 & 0.8 & 0 & 1 & 0 \\ 0 & -0.3 & 0 & 0 & 1 \end{bmatrix}, \quad \boldsymbol{\beta}_{\text{sim}} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \Psi^{\text{sim}} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}.$$

Figure 1 graphically depicts the association structure encoded by $\boldsymbol{\beta}^{\text{sim}}$. In words, Y_1 is related to $\{X_1, X_2\}$ and Y_2 is related to $\{X_1, X_3, X_4\}$. Estimation of $\boldsymbol{\beta}^{\text{sim}}$ is complicated by the large positive correlation between $\{X_1, X_3\}$ and $\{X_2, X_4\}$ and negative correlation between $\{X_2, X_5\}$, as well as the non-diagonality of the residual variance Ψ^{sim}. We simulate 500 data sets, each with 50 samples of (Y, X).

As outlined in Section 2, the first step of our analysis consists of fitting a Bayesian model. We fit model (1) using a multivariate version of the priors presented in George

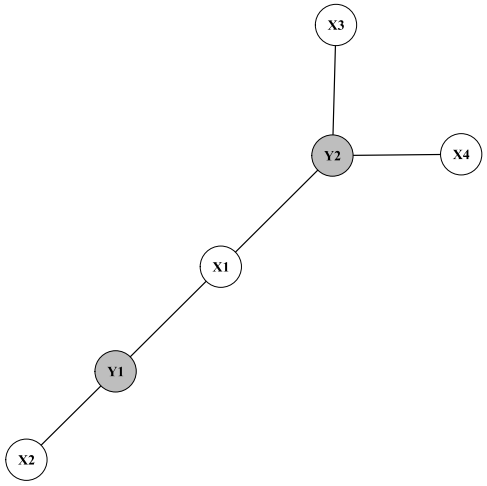
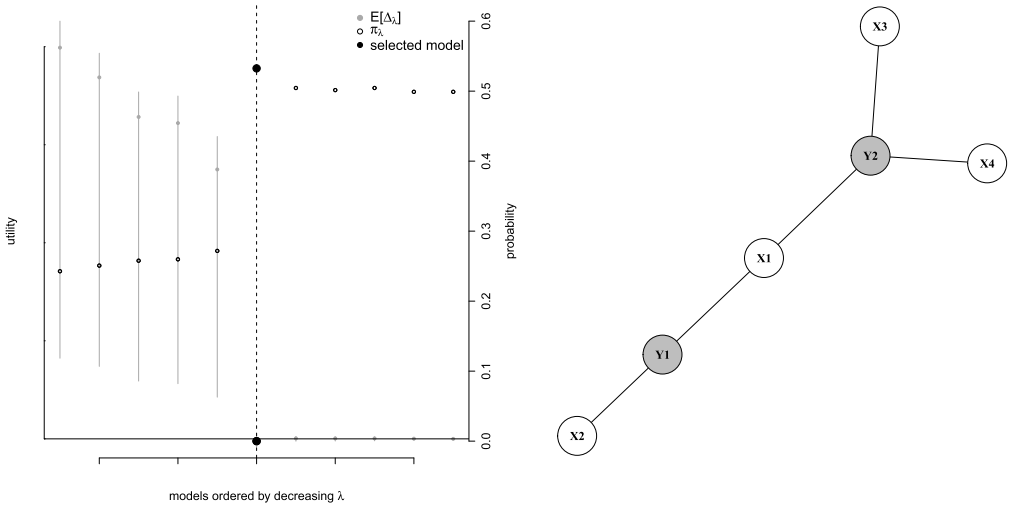


Figure 1: Graphical representation of the true graph defined by the coefficient matrix $\boldsymbol{\beta}^{\text{sim}}$.

and McCulloch (1993) and similar to Brown et al. (1998) and Wang (2010). Specifically, we use conjugate normal g-priors on the regression coefficients and choose the g parameter by an empirical Bayes procedure. The marginal distribution of the predictors is modeled via a Gaussian latent factor model. Details for all models and fitting algorithms are given in Supplementary Appendix A. Note that this is one of many such priors a researcher may choose.

Figures 2 display an example solution path for one of the data sets that selected the true model. The left axis is on the utility scale and shows the Δ_λ metric for decreasing penalty parameters λ. The right axis show the probability that the sparsified model is no worse than the saturated model π_λ. Posterior summary variable selection correctly identified the true model in 258 out of 500 simulated data sets at a 20% posterior uncertainty interval. Notice that the true model has a considerable jump in utility and π_λ on the left plot in Figures 2. In addition, the true model is contained in 400 out of the 500 posterior summary variable solution paths. This implies that our utility chooses the true model as the *best* 5-edge graph out of all possible graphs of equal size in 90% of the simulated data sets.

Next, we compare these results to a related method, based on the model of Wang (2010), which extends the stochastic search variable selection of George and McCulloch (1993) to seemingly unrelated regression models. Their model allows for sparsity in the $\boldsymbol{\beta}$ matrix and provides an inclusion probability for each entry in the matrix representing an edge in the graph. Details of the specific priors used may be found in Wang (2010). Although this model uses point-mass priors and produces posterior samples across various sparse regressions, the most common posterior summary, the (arithmetical) mean, generally produces *non-sparse* summaries. By contrast, among widely used summaries, the median probability model does provide a sparse point summary. In Barbieri and Berger (2004), the median probability model is shown to be optimal under squared



Figures 2: **(left)** Example of evaluation of Δ_λ and π_λ along the solution path for one of the simulated data sets where the true graph was correctly selected. Uncertainty bands are 20% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represent the graph selected and shown on the right. **(right)** The most selected graph for simulated data. This is the true graph given by $\boldsymbol{\beta}^{\text{sim}}$ and was selected for 258 out of the 500 simulated data sets and is present in 400 out of 500 posterior summary solution paths. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

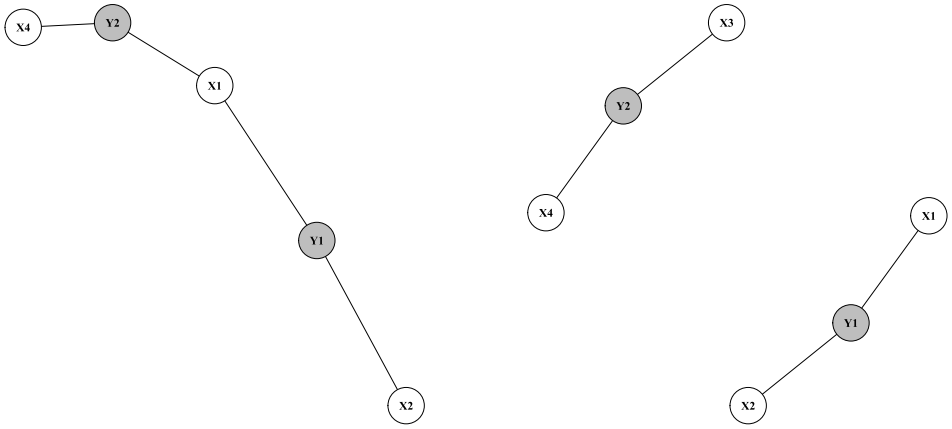
error prediction loss, as long as the predictor variables are mutually orthogonal. In practice, the median probability model is defined as the model containing all and only those variables with marginal inclusion probabilities greater than 1/2, whether or not the orthogonality condition is satisfied.

	covariate				
response	X_1	X_2	X_3	X_4	X_5
Y_1	0.8540	0.9588	0.1381	0.0335	0.0039
Y_2	0.5511	0.0467	0.5779	0.9492	0.0097

Table 1: Average edge inclusion probabilities for the $\boldsymbol{\beta}$ matrix across the 500 simulated data sets.

Table 1 shows the average edge inclusion probabilities for the $\boldsymbol{\beta}$ matrix across the 500 simulated data sets. Edge $\{Y_1, X_3\}$ was slightly sampled less than others, but the correlation in the predictors prohibits the inference from accurately ruling this out completely. The strong dependence between $\{X_1, X_3\}$ effects the inconclusive sampling of edges $\{Y_2, X_1\}$ and $\{Y_2, X_3\}$ as well – often when one edge is sampled, the other is excluded. Overall, the inclusion probabilities vary widely depending on the simulated data set.

Figures 3 depict the most common MPMs across the 500 data sets. The left graph was selected in 181 out of the 500 simulated data sets, and the right graph was selected in 149 out of the 500 data sets. The true graph shown in Figure 1 was only selected in 55 out of the 500 data sets.



Figures 3: The two most frequently appearing median probability models from the sparse SUR inference on each of the 500 simulated data sets. The left graph was selected in 181 out of the 500 simulated data sets, and the right graph was selected in 149 out of the 500 data sets.

To obtain a sense of how dissimilar the MPM summary can be compared to our approach, we tally how often the MPM appears in the posterior summary solution path displayed in, for example, the left side of Figures 2. A selected graph depends on the posterior uncertainty interval of Δ_λ; where a larger interval leads to sparser graphs. Therefore, if the MPM is contained in the solution path, it is a desirable model under our utility function modulo a degree of statistical uncertainty. We find that the MPM is contained in only 241 out of the 500 posterior summary solution paths. In other words, in 259 out of 500 solution paths, our utility function prefers a *different* model over the MPM of equal size (where size is measured by number of edges). Of the 241 occasions that they coincided, 55 of those recovered the true structure. We speculate that the difference between the two approach is largely due to strong correlation between predictors X_1 and X_3; our utility function explicitly considers this structure whereas the MPM formulation does not. In a similar simulation study with orthogonal predictors, the MPM recovers the true sparse structure in 488 out of 500 simulated data sets.

4 Applications

In this section, the sparse posterior summary method is applied to a data set from the finance (asset pricing) literature. A key component of our analysis will be a comparison between the posterior summaries obtained when the predictors are drawn at random versus when they are assumed fixed.

The response variables are returns on 25 tradable portfolios and our predictor variables are returns on 10 other portfolios thought to be of theoretical importance. In the asset pricing literature (Ross, 1976), the response portfolios represent assets to be priced (so-called *test assets*) and the predictor portfolios represent distinct sources of variation (so-called *risk factors*). More specifically, the test assets Y represent trading strategies based on company size (total value of stock shares) and book-to-market (the ratio of the company's accounting valuation to its size); see Fama and French (1992) and Fama and French (2015) for details. Roughly, these assets serve as a lower-dimensional proxy for the stock market. The risk factors are also portfolios, but ones which are thought to represent *distinct* sources of risk. What constitutes a distinct source of risk is widely debated, and many such factors have been proposed in the literature (Cochrane, 2011). Moreover, finding a small subset of factors (even from these 10) is useful for a finance researcher by providing *ease of interpretation*. If 3 factors are good enough predictively *and* easier for the finance researcher to grasp mentally, then this dimension reduction is useful; even in this moderately sized problem. We use monthly data from July 1963 through February 2015, obtained from Ken French's website:

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/`.

Our analysis investigates which subset of risk factors are most relevant (as defined by our utility function). As our initial candidates, we consider factors known in previous literature as: market, size, value, direct profitability, investment, momentum, short term reversal, long term reversal, betting against beta, and quality minus junk. Each factor is constructed by cross-sectionally sorting stocks by various characteristics of a company and forming linear combinations based on these sorts. For example, the value factor is constructed using the book-to-market ratio of a company. A high ratio indicates the company's stock is a "value stock" while a low ratio leads to a "growth stock" assessment. Essentially, the value factor is a portfolio built by going long stocks with high book-to-market ratio and shorting stocks with low book-to-market ratio. For detailed definitions of the first five factors, see Fama and French (2015). In the figures to follow, each is labeled as, for example, "Size2 BM3," to denote the portfolio buying stocks in the second quintile of size and the third quintile of book-to-market ratio.

Recent related work includes Ericsson and Karlsson (2004) and Harvey and Liu (2015). Ericsson and Karlsson (2004) follow a Bayesian model selection approach based off of inclusion probabilities, representing the preliminary inference step of our methodology. Harvey and Liu (2015) take a different approach that utilizes multiple hypothesis testing and bootstrapping.

4.1 Results

As outlined in Section 2.1, the first step of our analysis consists of fitting a Bayesian model. We fit model (1) using a multivariate version of the priors presented in George and McCulloch (1993) and similar to Brown et al. (1998) and Wang (2010). Specifically, we use conjugate normal g-priors on the regression coefficients and choose the g parameter by an empirical Bayes procedure. The marginal distribution of the predictors are modeled via a Gaussian latent factor model. Note that this is one of many such priors

a researcher may choose. The advantage of posterior summary variable selection is that any reasonable statistical model for the joint (Y, X) may be chosen.

Recalling the block structure for the covariance of the full joint distribution of (Y, X) from expression (2) we obtain posterior samples of Σ by sampling the conditional model parameters using a matrix-variate stochastic search algorithm (described below) and sampling the covariance of X from a latent factor model where it is marginally normally distributed. To reiterate our procedure is

- Σ_x is sampled from independent latent factor model,
- $\boldsymbol{\beta}$ and Ψ are sampled using a matrix-variate Markov Chain Monte Carlo algorithm.

The conditional model for $Y|X$ also includes the sampling of an indicator variable α that records if a given variable is non-zero (included in the model). In our simulation and application results, we fix α to the identity vector. This is done to emphasize that even when dense models are sampled in the inference step, our procedure has the ability to select a sparse set of predictors. Details of the model fitting algorithm may be found in Supplementary Appendix A.

In the subsections to follow, we will show a panel consisting of two figures. First, we plot the expectation of Δ_λ (and associated posterior credible interval) across a range of λ penalties. Recall, Δ_λ is the "loss gap" between a sparse summary and the best non-sparse (saturated) summary, meaning that smaller values are "better". Additionally, we plot the probability that a given model is no worse than the saturated model π_λ on this same figure, where "no worse" means $\Delta_\lambda < 0$. Note that even for very weak penalties (small λ), the distribution of Δ_λ will have non-zero variance and therefore even if it is centered about zero, some mass can be expected to fall above zero; practically, this means that $\pi_\lambda > 0.5$ is a very high score.

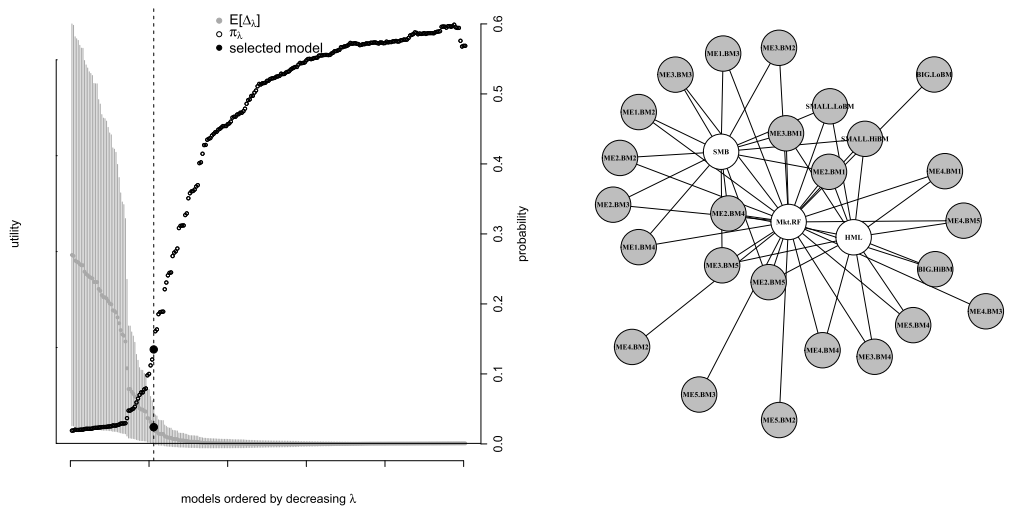
Second, we display a summary graph of the selected variables for the $\kappa = 12.5\%$ threshold. Recall that this is the highest penalty (sparsest graph) that is no worse than the saturated model with 12.5% posterior probability. For these graphs, the response and predictor variables are colored gray and white, respectively. A test asset label of, for example, "Size2 BM3," denotes the portfolio that buys stocks in the second quintile of size and the third quintile of book-to-market ratio. The predictors without connections to the responses under the optimal graph are not displayed.

These panels of two figures are shown in two scenarios:

1. Random predictors.
2. Fixed predictors.

Random predictors

This section introduces our baseline example where the risk factors (predictors) are random. We evaluate the set of potential models by analyzing plots such as the left plot



Figures 4: **(left)** Evaluation of Δ_λ and π_λ along the solution path for the 25 size/value portfolios modeled by the 10 factors. An analyst may use this plot to select a particular model. Uncertainty bands are 75% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represents the model selected and shown on the right. **(right)** The selected model for 25 size/value portfolios modeled by the 10 factors. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

in Figures 4. This shows Δ_λ and π_λ evaluated across a range of λ values. Additionally, we display the posterior uncertainty in the Δ_λ metric with gray vertical uncertainty bands; these are the centered $P\%$ posterior credible intervals where $\kappa = (1 - P)/2$. As the accuracy of the sparsified solution increases, the posterior of Δ_λ concentrates around zero by construction, and the probability of the model being no worse than the saturated model, π_λ, increases. We choose the sparsest model such that its corresponding $\pi_\lambda > \kappa = 12.5\%$. This model is displayed on the right in Figures 4 – also referred to as the "graphical summary".

The selected set of factors are the market (Mkt.RF), value (HML), and size (SMB). This three factor model is no worse than the saturated model with 12.5% posterior probability where all test assets are connected to all risk factors. Note also that in our selected model almost every test asset is distinctly tied to one of either value or size and the market factor. These are the three factors of Ken French and Eugene Fama's pricing model developed in Fama and French (1992). They are known throughout the finance community as being "fundamental dimensions" of the financial market, and our procedure is consistent with this widely held belief at a small κ level.

The characteristics of the test assets in the selected graph from Figure 4 are also important to highlight. The test portfolios that invest in small companies ("Size1" and "Size2") are primarily connected to the SMB factor which is designed as a proxy for the

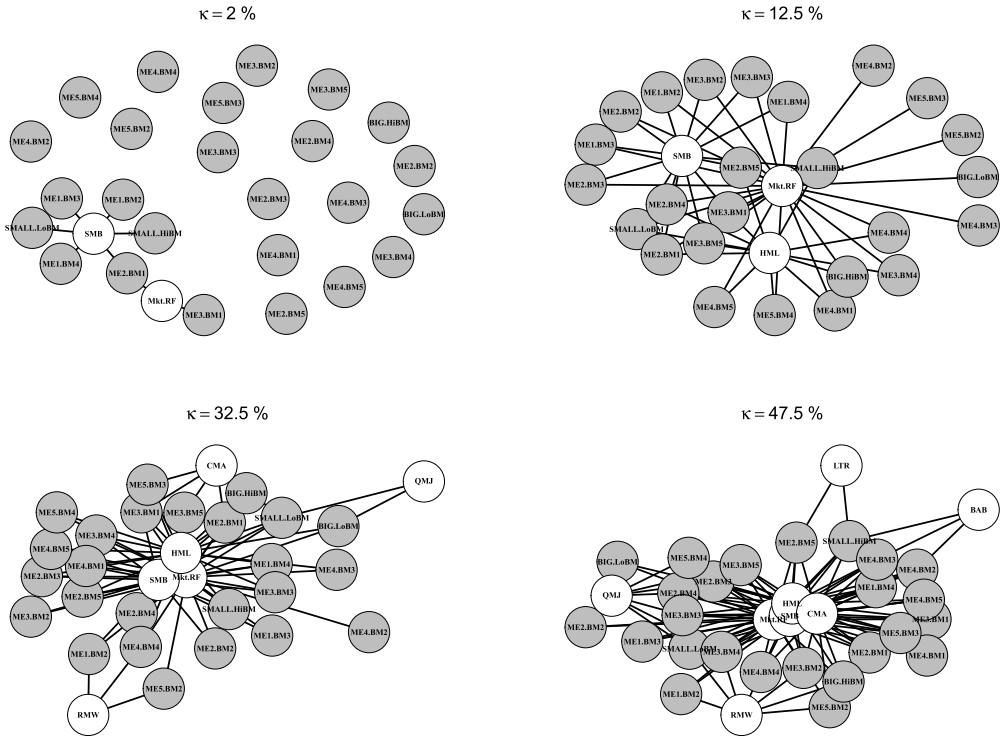


Figure 5: Sequence of selected models for varying threshold level κ under the assumption of **random predictors**.

risk of small companies. Similarly, the test portfolios that invest in high book-to-market companies ("BM4" and "BM5") have connections to the HML factor which is built on the idea that companies whose book value exceeds the market's perceived value should generate a distinct source of risk. As previously noted, all of the test portfolios are connected to the market factor suggesting that it is a relevant predictor even for the sparse $\kappa = 12.5\%$ selection criterion.

In Figure 5, we examine how different choices of the κ threshold change the selected set of risk factors. In this analysis, there is a trade-off between the posterior probability of being "close" to the saturated model and the utility's preference for sparsity. When the threshold is low ($\kappa = 2$ and 12.5%) the summarization procedure selects relatively sparse graphs with up to three factors (Mkt.RF, HML, and SMB). The market (Mkt.RF) and size (SMB) factors appear first, connected to a small number of the test assets ($\kappa = 2\%$). As the threshold is increased, the point summary becomes denser and correspondingly more predictively accurate (as measured by the utility function). The value factor (HML) enters at $\kappa = 12.5\%$ and quality minus junk (QMJ), investment (CMA), and profitability (RMW) factors enter at $\kappa = 32.5\%$. The graph for $\kappa = 32.5\%$ excluding QMJ is essentially the new five factor model proposed by Fama and French

(2015). The five Fama–French factors (plus OMJ, BAB and LTR with a small number of connections) persist up to the $\kappa = 47.5\%$ threshold. This indicates that, up to a high posterior probability, the five factor model of Fama and French (2015) does no worse than an asset pricing model with all ten factors connected to all test assets.

Notice also that our summarization procedure displays the specific relationship between the factors and test assets through the connections. Using this approach, the analyst is able to identify which predictors drive variation in which responses and at what thresholds they may be relevant. This feature is significant for summarization problems where individual characteristics of the test portfolios and their joint dependence on the risk factors may be a priori unclear.

As κ approaches the 50% threshold ($\kappa = 47.5\%$ in Figure 5), the model summary includes eight of ten factors. Requesting a summary with this level of certainty results in little sparsification. However, an additional contribution of a factor results in minor increases in out utility. Sparse posterior summarization applied in this context allows an analyst to study the impact of risk factors on pricing while taking uncertainty into account. Coming to a similar conclusion via common alternative techniques (e.g., component-wise ordinary least squares combined with thresholding by t-statistics) is comparatively ad hoc; our method is simply a perspicuous summary of a posterior distribution. Likewise, applying sparse regression techniques based on ℓ_1 penalized likelihood methods would not take into account the residual correlation Ψ, nor would that approach naturally accommodate random predictors.

Fixed predictors

In this section, we consider posterior summarization with the loss function derived under the assumption of *fixed predictors*. The analogous loss function when the predictor matrix is fixed at pre specified points $\mathbf{X}$ is:

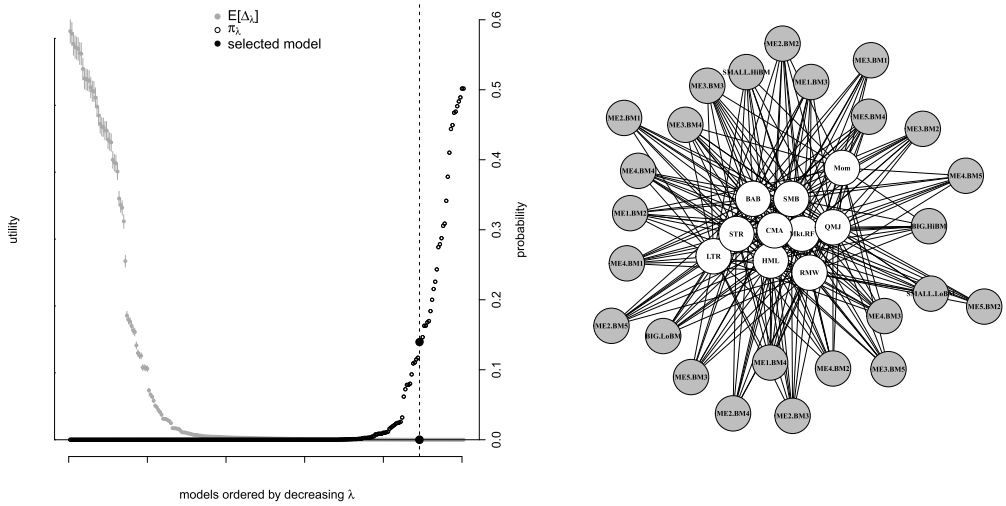
$$\mathcal{L}_\lambda(\boldsymbol{\gamma}) = \left\| \left[Q_f^T \otimes L^T \right] \mathbf{vec}(\boldsymbol{\gamma}) - \mathbf{vec}(L^{-1} A_f Q_f^{-T}) \right\|_2^2 + \lambda \left\| \mathbf{vec}(\boldsymbol{\gamma}) \right\|_1, \tag{9}$$

with $Q_f Q_f^T = \mathbf{X}^T\mathbf{X}$, $A_f = \mathbb{E}[\Omega \tilde{\mathbf{Y}}^T \mathbf{X}]$, and $M = \overline{\Omega} = LL^T$; compare to (6) and (7). The derivation of (9) is similar to the presentation in Section 2 and may be found in Supplementary Appendix C.

The corresponding version of the loss gap is

$$\Delta_\lambda = \mathcal{L}(\tilde{\mathbf{Y}}, \mathbf{X}, \Theta, \boldsymbol{\gamma}_\lambda^*) - \mathcal{L}(\tilde{\mathbf{Y}}, \mathbf{X}, \Theta, \boldsymbol{\gamma}^*),$$

which has distribution induced by the posterior over $(\tilde{\mathbf{Y}}, \Theta)$ rather than $(\tilde{Y}, \tilde{X}, \Theta)$ as before. By fixing $\mathbf{X}$, the posterior of Δ_λ has smaller dispersion which results in denser summaries for the same level of κ. For example, compare how dense the graph in Figures 4 is relative to the graph in Figures 6. The denser graph in Figures 6 contains all ten potential risk factors compared to just three in Figures 4, which correspond to the Fama–French factors described in Fama and French (1992). Recall that both graphs represent the sparsest model such that the probability of being no worse than the saturated model



Figures 6: **(left)** Evaluation of Δ_λ and π_λ along the solution path for the 25 size/value portfolios modeled by the 10 factors under the assumption of **fixed predictors**. An analyst may use this plot to select a particular model. Uncertainty bands are 75% posterior intervals on the Δ_λ metric. The large black dot and associated dashed line represents the model selected and shown on the right. **(right)** The selected model for 25 size/value portfolios modeled by the 10 factors. The responses and predictors are colored in gray and white, respectively. Edges represent nonzero components of the optimal action, $\boldsymbol{\gamma}$.

is greater than $\kappa = 12.5\%$ — the difference is that one of the graphs defines "worse-than" in terms of a fixed set of risk factor returns while the other acknowledges that those returns are themselves uncertain in future periods.

Figure 7 demonstrates this problem for several choices of the uncertainty level. Regardless of the uncertainty level chosen, the selected models contain most (if not all) of the ten factors and many edges. In fact, it is difficult to distinguish even the $\kappa = 2\%$ and $\kappa = 47.5\%$ models.

5 Conclusion

In this paper, we propose a model selection summary for multivariate linear models when future realizations of the predictors are unknown. Such models are widely used in many areas of science and economics, including genetics and asset pricing. Our utility-based sparse posterior summary procedure is a multivariate extension of the "decoupling shrinkage and selection" methodology of Hahn and Carvalho (2015). The approach we develop has three steps: (*i*) fit a Bayesian model, (*ii*) specify a utility function with a sparsity-inducing penalty term and optimize its expectation, and (*iii*) graphically summarize the posterior impact (in terms of utility) of the sparsity penalty. Our utility

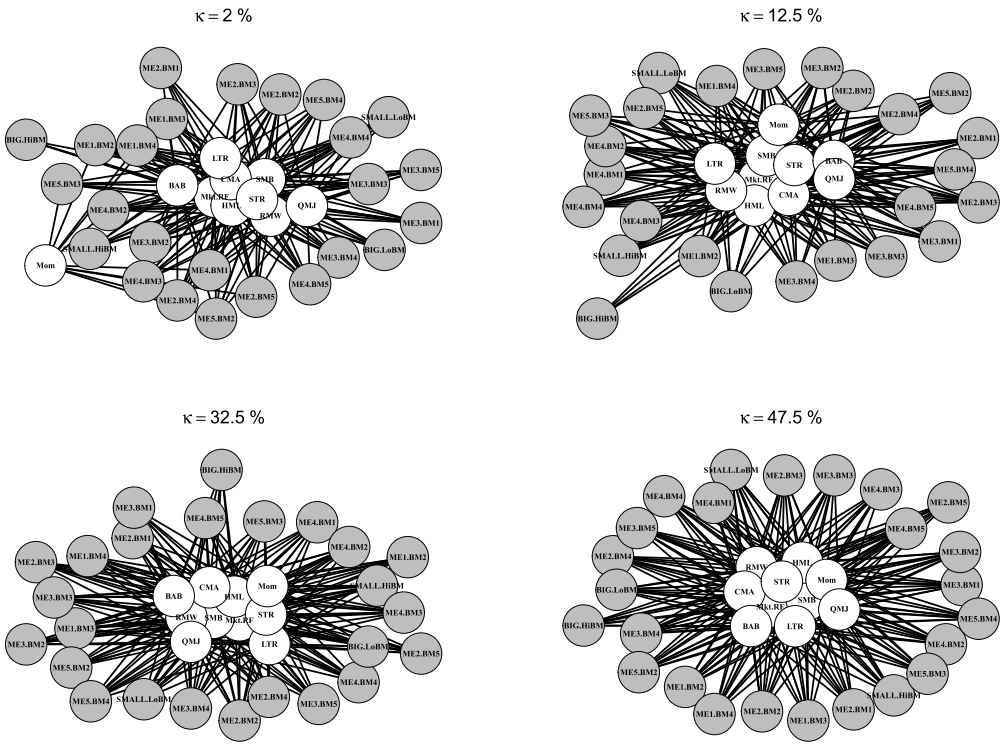


Figure 7: Sequence of selected models for varying threshold level κ under the assumption of **fixed predictors**.

function is based on the kernel of the conditional distribution responses given the predictors and can be formulated as a tractable convex program. We demonstrate how our procedure may be used in asset pricing under a variety of modeling choices.

The remainder of this discussion takes a step back from the specifics of the seemingly unrelated regressions model and considers a broader role for utility-based posterior summaries.

A paradox of applied Bayesian analysis is that posterior distributions based on relatively intuitive models like the SUR model are often just as complicated as the data itself. For Bayesian analysis to become a routine tool for practical inquiry, methods for summarizing posterior distributions must be developed apace with the models themselves. A natural starting point for developing such methods is decision theory, which suggests developing loss functions specifically geared towards practical posterior summary. As a matter of practical data analysis, articulating an apt loss function has been sorely neglected relative to the effort typically lavished on the model specification stage, specifically prior specification. Ironically (but not surprisingly) our application demonstrates that one's utility function has a dominant effect on the posterior summaries obtained relative to which prior distribution is used.

This paper makes two contributions to this area of "utility design". First, we propose that the likelihood function has a role to play in posterior summary apart from its role in inference. That is, one of the great practical virtues of likelihood-based statistics is that the likelihood serves to summarize the data by way of the corresponding point estimates. By using the log-density as our utility function applied to *future* data, we revive the fundamental summarizing role of the likelihood. Additionally, note that this approach allows three distinct roles for parameters. First, all parameters of the model appear in defining the posterior predictive distribution. Second, some parameters appear in *defining* the loss function; Ψ plays this role in our analysis. Third, some parameters define the action space. In this framework there are no "nuisance" parameters that vanish from the estimator as soon as a marginal posterior is obtained. Once the likelihood-based utility is specified, it is a natural next step to consider augmenting the utility to enforce particular features of the desired point summary. For example, our analysis above was based on a utility that explicitly rewards sparsity of the resulting summary. A traditional instance of this idea is the definition of high posterior density regions, which are defined as the *shortest, contiguous* interval that contains a prescribed fraction of the posterior mass.

Our second contribution is to consider not just one, but a range, of utility functions and to examine the posterior distributions of the corresponding posterior loss. Specifically, we compare the utility of a sparsified summary to the utility of the optimal non-sparse summary. Interestingly, these utilities are random variables themselves (defined by the posterior distribution) and examining their distributions provides a fundamentally Bayesian way to measure the extent to which the sparsity preference is driving one's conclusions. The idea of comparing a hypothetical continuum of decision-makers based on the posterior distribution of their respective utilities represents a principled Bayesian approach to exploratory data analysis. This is an area of ongoing research.

Supplementary Material

Supplement for Variable selection in seemingly unrelated regressions with random predictors (DOI: 10.1214/17-BA1053SUPP; .pdf).

References

Barbieri, M. M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics*, 870–897. MR2065192. doi: http://dx.doi.org/10.1214/009053604000000238. 977

Bayarri, M., Berger, J., Forte, A., and Garcia-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: http://dx.doi.org/10.1214/12-AOS1013. 969

Brown, P. J., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian variable selection and prediction." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3): 627–641. MR1626005. doi: http://dx.doi.org/10.1111/1467-9868.00144. 970, 972, 975, 977, 980

Clyde, M. and George, E. (2004). "Model uncertainty." *Statistical Science*, 19: 81–94. 969

Cochrane, J. H. (2011). "Presidential address: Discount rates." *The Journal of Finance*, 66(4): 1047–1108. 980

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). "Sparse graphical models for exploring gene expression data." *Journal of Multivariate Analysis*, 90(1): 196–212. MR2064941. doi: http://dx.doi.org/10.1016/j.jmva.2004.02.009. 970

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). "Least angle regression." *The Annals of Statistics*, 32(2): 407–499. 973

Ericsson, J. and Karlsson, S. (2004). "Choosing Factors in a Multifactor Asset Pricing Model: A Bayesian Approach." Technical report, Stockholm School of Economics. 980

Fama, E. F. and French, K. R. (1992). "The cross-section of expected stock returns." *The Journal of Finance*, 47(2): 427–465. 980, 982, 984

Fama, E. F. and French, K. R. (2015). "A five-factor asset pricing model." *Journal of Financial Economics*, 116(1): 1–22. 980, 983, 984

Garcia-Donato, G. and Martinez-Beneito, M. (2013). "On sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association*, 108(501): 340–352.

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 969, 972, 976, 977, 980

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective." *Journal of the American Statistical Association*, 110(509): 435–448. 969, 970, 975, 976, 985

Hans, C., Dobra, A., and West, M. (2007). "Shotgun stochastic search for "large p" regression." *Journal of the American Statistical Association*, 102(478): 507–516. 969

Harvey, C. R. and Liu, Y. (2015). "Lucky factors." *Available at SSRN 2528780*. 980

Jeffreys, H. (1961). "Theory of Probability (3rd edt.) Oxford University Press." MR0187257.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in stochastic computation for high-dimensional graphical models." *Statistical Science*, 388–400. 970

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008a). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008b). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481).

Murray, J. S., Dunson, D. B., Carin, L., and Lucas, J. E. (2013). "Bayesian Gaussian copula factor models for mixed data." *Journal of the American Statistical Association*, 108(502): 656–665.

Puelz, D., Hahn, P. R., and Carvalho, C. M. (2017). "Supplement for Variable selection in seemingly unrelated regressions with random predictors." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/17-BA1053SUPP. 970

Ross, S. A. (1976). "The arbitrage theory of capital asset pricing." *Journal of Economic Theory*, 13(3): 341–360. MR0429063. doi: http://dx.doi.org/10.1016/0022-0531(76)90046-6. 980

Scott, J. and Berger, J. (2006). "An exploration of aspects of Bayesian multiple testing." *Journal of Statistical Planning and Inference*, 136: 2144–2162. 969

Wang, H. (2010). "Sparse seemingly unrelated regression modelling: Applications in finance and econometrics." *Computational Statistics & Data Analysis*, 54(11): 2866–2877. MR2720481. doi: http://dx.doi.org/10.1016/j.csda.2010.03.028. 970, 972, 975, 976, 977, 980

Wang, H. and West, M. (2009). "Bayesian analysis of matrix normal graphical models." *Biometrika*, 96(4): 821–834. 970

Zellner, A. (1962). "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias." *Journal of the American Statistical Association*, 57(298): 348–368. 970

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6: 233–243.

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." *Trabajos de estadística y de investigación operativa*, 31(1): 585–603.

Zellner, A. and Siow, A. (1984). *Basic issues in econometrics*. University of Chicago Press Chicago.