Strategic Bayesian Asset Allocation

Vadim Sokolov* and Michael Polson†

First Draft: November 2018
This Draft: November 2019

Abstract

Strategic asset allocation requires an investor to select stocks from a given basket of assets. The perspective of our investor is to maximize risk-adjusted alpha returns relative to a benchmark index. Historical returns are used to provide inputs into an optimization algorithm. Our approach uses Bayesian regularization to not only provide stock selection but also optimal sequential portfolio weights. By incorporating investor preferences with a number of different regularization penalties we extend the approaches of Black and Litterman [1992] and Puelz et al. [2015]. We tailor standard sparse MCMC algorithms to calculate portfolio weights and perform selection. We illustrate our methodology on stock selection from the SP100 stock index and from the top fifty holdings of two hedge funds Renaissance Technologies and Viking Global. Finally, we conclude with directions for future research.

1 Introduction

Strategic asset allocation requires an investor to select stocks from a given basket of assets with a goal to outperform a benchmark index by a given margin of alpha. We propose to use Bayesian regularization, based on predicted expected returns and volatility from historical returns, to simultaneously perform asset selection and optimal portfolio allocations. Our approach builds on the asset allocation model of Black and Litterman [1992] which has become the gold standard for practitioners. This approach incorporates investors' preferences on expected returns. Puelz et al. [2015] provide a similar analysis using shrinkage and selection to select stocks using regularized regression.

Our methodology builds on traditional stock selection techniques in a number of ways. First, we recast portfolio selection as a regularization problem with a goal to optimize expected returns subject to a desired number of stocks in the portfolio. Second, we use sparsity inducing prior distributions such as spike-and-slab, horseshoe, and Lasso to perform selection. Thirdly, the output from our algorithm naturally orders assets to

---

*Vadim Sokolov is Assistant Professor in Operations Research at George Mason University. email:vsokolov@gmu.edu

†Michael Polson as an Analyst at Bates White. email:michael.alan.polson@gmail.com

include into the optimal portfolio in a one-by-one fashion so that performance can be dynamically tracked over time. Our contributions include

- Duality between hierarchical linear models and constrained optimization problem. Hierarchical linear model allows for more flexible set of prior distributions and allows to use existing MCMC algorithms for finding optimal portfolios

- Duality between uncertainty in input data and allocation weights. This provides a new interpretation of sparsity inducing the Laplace prior which corresponds to Lasso estimator.

- Extending the Black-Litterman framework to sparsity-inducing priors

- Our empirical results demonstrating effectiveness of several sparsity-inducing priors on portfolio selection and its performance.

The traditional Markowitz [1952] mean-variance approach is optimal under the assumption that historical means and variances provide good estimates of future expected returns and volatilities. A number of practical considerations, such as preferences to include as few stocks as possible into portfolio, needs further research. Markowitz portfolio selection is also known to be sensitive to errors in estimation and can lead to extreme weights which hinders its wide spread adoption. Black-Litterman use shrinkage (a.k.a. ridge regression) to improve performance and to allow for investor preferences. Polson and Tew [2000] use sparsity inducing priors that impose investors' beliefs and show how to estimate covariance matrices for large-scale portfolio problems.

Black and Litterman [1992] proposed using Gaussian regularization to perform the shrinkage and frame this approach in the context of combining quantitative and subjective beliefs into a predictive model of expected returns. We show how to incorporate sparsity and selection. While the Markowitz approach relies on convex optimization, BL relies on Bayesian posterior mode updates of portfolio weights, and as we will show later leads to identical solutions under certain assumptions. Stevens [1998] provided the connection between statistical and mean-variance based approaches. Bertsimas et al. [2012] show how the original BL approach can be recast as an optimization problem.

Our contribution is both methodological and empirical. On the methodological side, we develop a new optimal asset allocation algorithm using sparsity. On the empirical side, we compute asset allocation and selection for both static buy-and-hold and dynamic optimal re-balancing. We evaluate three sparsity inducing priors for portfolio allocation, namely Laplace, Horseshoe [Carvalho et al., 2010] and spike-and-slab. The empirical performance of sparsity approaches is demonstrated by comparing performance of selected portfolios with SP500 returns. The LARS algorithm [Efron et al., 2004] is particularly effective in finding the optimal portfolio posterior mode for the Laplace prior. MCMC works well for Horseshoe Hahn et al. [2019], and Single Best Replacement (SBR) Polson and Sun [2019] algorithm for finding posterior mode for spike-and-slab.

Our approach naturally addresses the problem of portfolio selection and the issue of over fitting is alleviated by applying regularization techniques, see Kandel et al. [1995] who address the problem of distribution shifts with Bayesain techniques. Carrasco and

Noumon [2011] apply regularization technique for the degeneracy of the estimated covariance matrix of a large number of assets.

1.1 Connections with Previous Work

Mean-variance portfolio selection has a long-standing place in financial econometrics with ground-breaking work of De Finetti [1940] and Markowitz [1952]. Scalability of optimization algorithms and incorporation of transactions costs was addressed by Perold [1984]. A number of authors have provided Bayesian solutions [Barberis, 2000], using a the predictive model of stock returns together with regularization.

Our work builds on other Bayesian stock selection strategies such as those based on factor modeling, see Black and Litterman [1992], Carvalho et al. [2011], Aguilar and West [2000], Puelz et al. [2015], Getmansky et al. [2015]. Our main assumption is predictability Barberis [2000], Kandel and Stambaugh [1996] of stock returns, which is the main justification that an investor should heavily invest into stocks due to the large equity premium. A related approach is to regularize with ℓ_1 penalties (lasso) which has been well studied in the portfolio optimization literature, see for example [Fastrich et al., 2015]. ℓ_1-based approaches have their shortcomings including over-shrinkage and the inability to recover sparse signals for highly dependent data.

To illustrate this issue, several authors have proposed non-convex approaches. Gasso et al. [2009] and Giuzio and Paterlini [2018] use ℓ_q penalties to address the issue of highly dependent data and allocate portfolios during a crisis. Other non-convex penalties include smoothly clipped absolute deviations (SCAD) Fan and Li [2001] and its linear approximation Zhang [2009]. Bridge [Polson et al., 2014] or ℓ_q penalty Frank and Friedman [1993] is a generalization of more widely used ℓ_1 (LASSO) and ℓ_2 (Ridge) penalties. [Bourgeron et al., 2018] show that the mean-variance optimization approach will play an arbitrage between assets with similar return and volatility and show that regularization techniques are required to find a robust asset allocations.

A Bayesian approach naturally allows an investor to incorporate uncertainty about mean-variance parameters. For example, Carvalho et al. [2011] address the problem of change in covariance estimates by dynamically updating it as new observations arrive [Jacquier and Polson, 2012]. Robust minimax optimization techniques were recently proposed to account for uncertainty in covariance matrix estimator and to solve for the worst-case scenario, see Ismail and Pham [2019]. Our work builds on Puelz et al. [2015] who use Bayesian techniques to design a mean-variance portfolio with a small number of assets and analyze the trade off between optimality and number of assets to be included. Kozak et al. [2019] design sparse factor models for analysis of large number of cross-sectional stock returns. Jacquier and Polson [2012] proposed decision-theoretic framework for asset allocations that relies on Bayesian analysis, see Avramov and Zhou [2010] for further discussion.

2   Strategic Bayesian Asset Allocation

Traditional mean-variance portfolio optimization assumes that returns at time t of asset i, $r_{it} = p_{it}/p_{i(t-1)}$ of each of n assets follow a distribution with mean $\mu \in R^n$ and covariance matrix $\Sigma \in R^{n \times n}$. Here p_{it} is the price of asset i at time t. Returns of each asset are assumed to be a weak-stationary stochastic process.

   The investor's objective is to minimize the variance (risk) of the portfolio while having a guaranteed return ρ. A portfolio is defined by a vector of weights (allocations), denoted by $w = (w_1, \ldots, w_p)$. Thus, the variance of the portfolio is given by $w^T \Sigma w$ and the expected return by $\mu^T w$.

   Then, the optimal portfolio is found by solving the following optimization problem

$$\underset{w}{\text{minimize}} \quad w^T \Sigma w \tag{1}$$
$$\text{subject to} \quad r^T w = \rho, \quad 1^T w = 1, \quad w_i \geq 0$$

where the weight w_i is the amount of asset i held throughout the period. Positivity constraint $w_i \geq 0$ are added to guarantee that only long positions are to be included.

2.1   Extending Black-Litterman

The first widely used regularization approach was proposed by Black and Litterman [1992]. The BL model uses quadratic regularization, which can be interpreted as a mechanism to integrate quantitative and traditional portfolio building strategies. The BL model assumes a normal prior over investor's beliefs over future returns. The objective function then combines loss minimization with the regularization term that encodes investors' beliefs. In other words, the BL model combines quantitative and traditional management approaches and allows to update currently held beliefs using observed data (returns) to form new opinions.

   We begin by providing a framework to determine an optimal portfolio allocation from the perspective of an informed investor. Let π denote an expected return vector for the n assets. Following Black-Litterman, let π^* denote beliefs about expected returns over and above those implied by π. Expected returns π, are typically determined from an equilibrium model, such as CAPM or FF3/5.

   Specifically, under CAPM model , we have

$$\pi = \beta(\mu_m - r_f) \quad \text{where} \quad \beta = \text{Cov}(r, r^T w_m)/\sigma_m^2.$$

Here $w_m^T r$ is the return on the (global) market, r_f is a risk-free rate, σ_m^2 is the market variance, and r is the vector of asset returns. We can then write $\pi = \sigma \Sigma w_m$, where $\Sigma = \text{Cov}(r, r^T)$. Here, $\sigma = (\mu_m - r_f)/\sigma_m$ is the Sharpe ratio of the market index. From this, we can calculate optimal weights given expected returns, namely $w_m = (\sigma \Sigma)^{-1} \pi$.

   This duality allows us to move between weights and expected returns. Informed investor beliefs are typically given by a constraint of the form $P\pi^* = q$, where P is a vector and P_i identifies if there is an opinion about asset i, and Q is the vector of opinions. As

a simple example, suppose that $P = (-1, 1, 0, \ldots, 0)$ and $q = (0.5, 0, \ldots, 0)$ express the belief that the first asset will outperform the second by 0.5% over the period in question. Uncertainty in this belief can be expressed as $P\pi^* \sim N(q, \tau\Omega)$, where τ measures the strength of such a belief. BL show how Bayesian updating lets the investor to calculate the distribution over the new beliefs $p(\pi^* \mid \pi)$

$$\pi \mid \pi^* \sim N(\pi^*, \Sigma), \quad P\pi^* \sim N(q, \tau\Omega)$$

combined with Bayes rule,

$$\pi^* \mid \pi \sim N\left(\mu, \left((\tau\Sigma)^{-1} + P^T\Omega^{-1}P\right)^{-1}\right)$$

$$\mu = \left((\tau\Sigma)^{-1} + P^T\Omega^{-1}P\right)^{-1}\left((\tau\Sigma)^{-1}\pi + P^T\Omega^{-1}Q\right)$$

Regularization is now a central tool to allow investors to perform stock selection. This requires a selection of a norm on the portfolio weights. Several authors have applied ℓ_1 regularization. DeMiguel et al. [2009] build on the work of Jagannathan and Ma [2003] and Ledoit and Wolf [2004] and propose a general mean-variance portfolio allocation framework in which norm of portfolio weights is constrained. They show duality of constraint-based approach and Bayesian approach in which investor assigns prior distribution for each of the weights. Lobo et al. [2007] show that inclusion of transaction costs make ℓ_1 regularized formulation to be non-convex and propose convex relaxations that can be efficiency solved. Brodie et al. [2009] showed that regularization techniques do improve predictive power of statistical models for stock portfolios. Fan et al. [2012] show that regularized mean-variance approach allows to achieve similar performance to the theoretically optimal portfolio while using covariance matrix estimated from a sample.

2.2 Portfolio Weight Regularization

In order to apply Bayesian inference algorithms, we re-cast the optimization problem as a hierarchical Bayesian linear model. To do this, let the return of the portfolio at time t is given by the return vector $r_t = (r_{1t}, \ldots, r_{pt})$. Then, the empirical estimate for the variance of the portfolio is given by

$$w^T\Sigma w = \text{Var}\left(w^T r\right) = \text{E}\left((r^T w - \mu^T w)(r^T w - \mu^T w)^T\right) = \text{E}\left((r^T w - \rho)(r^T w - \rho)^T\right)$$

The empirical risk is given by $w^T\hat{\Sigma}w = (1/T)\sum_{t=1}^{T}(\rho - r_t^T w)^2$.

We re-write risk minimization objective as a least-squares problem

$$\begin{aligned}
\underset{w}{\text{minimize}} \quad & (1/T)\|\boldsymbol{\rho} - Rw\|_2^2 \\
\text{subject to} \quad & \mu^T w = \rho \\
& 1^T w = 1,
\end{aligned}$$

were $\boldsymbol{\rho} = (\rho, \ldots, \rho) \in R^T$.

5

The return matrix R is typically ill-conditioned. This can lead to unstable numerical solutions of the above problem. It is a usual problem when assets are highly correlated, then the columns of matrix R become almost linearly dependent and the matrix becomes ill-conditioned. One approach to stabilize the solution and to find sparse portfolios is to add a regularization penalty to the objective function Brodie et al. [2009]. Brodie et al. [2009] analyzed the case where the penalty function is based on absolute value $\phi(w) = \sum_i |w_i|$ and showed that it leads to a stable solution. Our approach builds on the work of Puelz et al. [2015] who viewed the absolute value penalty as a way to incorporate investor's desire for a simple portfolio. They take a similar view as Black and Litterman and show that investor's preference to allocated her wealth among a small number of assets.

Sparse portfolios allow us to reduce transaction costs by eliminating certain stocks and to minimize the number of stocks that an investor need to follow and research. Another, interpretation of the penalized objective Puelz et al. [2015] is that it allows to incorporate investor's preference with regards to number of stocks to be included into portfolio

$$
\begin{aligned}
\underset{w}{\text{minimize}} \quad & (1/T)\|\boldsymbol{\rho} - Rw\|_2^2 + \tau\phi(w) \\
\text{subject to} \quad & \mu^T w = \rho \\
& 1^T w = 1
\end{aligned}
$$

Noting that we do not include the positivity constraint into our regularized formulation and allow for short positions in the portfolio. A regularization penalty added to the objective function allows to stabilize portfolio Brodie et al. [2009]. Thus the positive weight constraint can be excluded in a regularized formulation. In order to satisfy the $1^T w = 1$ constraint, we modify the problem by subtracting first column of the R matrix from other columns and then estimate $p - 1$ linear coefficients of the modified problem $w_2, \ldots, w_p$ and finally calculate the remaining weight $w_1 = 1 - \sum_{i=2}^p w_i$.

Denote the penalized empirical risk now by $g(w) = (1/T)\|\boldsymbol{\rho} - Rw\|_2^2 + \tau\phi(w)$. The corresponding Lagrangian dual function associated with the optimization problem with the $1^T w = 1$ constraint excluded is given by

$$
h(\lambda) = \inf_w \left[g(w) + \lambda(\rho - \mu^T w) \right].
$$

The dual function yields lower bounds on the optimal portfolio w^*. For any λ, we have

$$
h(\lambda) \leq g(w^*).
$$

For a specific value of $\lambda^* \in \arg\max_\lambda g(\lambda)$, we have $h(\lambda^*) = g(w^*)$. Since $\lambda\rho$ term does not depend on w, we re-write the problem as

$$
\underset{w}{\text{minimize}} \quad (1/T)\|\boldsymbol{\rho} - Rw\|_2^2 + \tau\phi(w) - \lambda\mu^T w. \tag{2}
$$

We cam select the value of the dual variables λ and τ using cross-validation or naturally as stocks get added to the portfolio.

The penalty terms $-\lambda\mu^T w$ and $\tau\phi(w)$ in (2) can also be viewed as judgement of an investor that need to be incorporated into portfolio allocation decision making. To interpret those terms as prior judgement, we re-write the optimization problem as a Bayesian inference problem.

2.3   Duality between weight and moment regularization

Bayesian techniques for portfolio selection can be divided into two categories, those that assume priors on the expected returns and variance-covariance matrix and those that place priors on the portfolio weights. Under certain conditions regularizing priors on the weights are equivalent to approaches that assume uncertainty in the observed data. For example, Jagannathan and Ma [2003] show that adding non-negative weight constraint is equivalent to shrinking elements of the covariance matrix, which leads to reduced risk portfolio and more stable allocations. This follows directly from the KKT (Karush-Kuhn-Tucker) optimality condition for the portfolio optimization problem with an additional constraints $0 \geq w_i \leq \bar{w}$, so that

$$\sum_j \Sigma_{ij} w_j - \gamma_i + \sigma_i = 0, \quad i = 1, \ldots, N$$

$$\gamma_i \geq 0, \quad \gamma_i = 0, \text{ if } w_i > 0$$
$$\delta_i \geq 0, \quad \delta_i = 0 \text{ if } w_i < \bar{w}$$

Here γ's and δ's are Lagrange multipliers. Thus, solving constrained problem is equivalent to solving the unconstrained problem with $\tilde{\Sigma} = \Sigma + \Delta - A$, where $\Delta_{ij} = \sigma_i + \sigma_j$, and $A_{ij} = \gamma_i + \gamma_j$. Thus, the Lagrange multipliers γ associated with the $w_i \geq 0$ constraint are effectively shrink the elements of the covariance matrix Σ. More specifically, Σ_{ij} is reduced by $\gamma_i + \gamma_j$, and Σ_{ii} is reduced by $2\gamma_i$.

More generally, Xu et al. [2009] and [Bertsimas et al., 2016] provide duality between robust optimization which assumes uncertainty over return data and linear models with weight regularization. Specifically, one can show that a robust min-max formulation of a maximum likelihood estimation for a linear model

$$\underset{w \in \mathbb{R}^P}{\text{minimize}} \quad \left(\underset{\Delta R \in U}{\text{maximize}} \quad \|\rho - (R + \Delta R)w\|_2^2 \right)$$

with feature-wise uncertainty set given by $U = \{\Delta R \mid \|\Delta R_i\|_2 \leq c, i = 1, \ldots, p\}$ is equivalent to an ℓ_1 penalized log-likelihood maximization

$$\underset{w}{\text{minimize}} \quad \|\rho - Rw\|_2^2 + \sum_{i=1}^{p} c_i |w_i|.$$

We now turn to stock selection.

2.4   Stock Selection via Bayesian Inference

The key insight is that the optimization problem defined by (2) is equivalent to finding a mode of a posterior distribution for a linear Gaussian model with exponential prior on the parameters and sparsity prior (regularization).

$$\rho = Rw + \epsilon, \text{ where } w \sim p(w)f(w), \ \epsilon \sim N(0, \sigma_e^2 I). \tag{3}$$

Here the prior is a product of factors with distributional assumptions $p(w) = \exp(-\lambda \mu^T w)$ and $f(w) \propto \prod_{i=1}^{p} \exp(\tau \phi(w_i))$.

Since $-\log p(\rho \mid R, w) \propto \|\rho - Rw\|_2^2$, the mode of the log-posterior distribution over the coefficients of the above linear model is given by

$$\log p(w \mid R, \rho) = c + \log p(\rho \mid R, w) + \log p(w) + \log f(w).$$

This is equal to the solution of the optimization problem given by (2). The exponential prior corresponds to the equality constraint $\mu^T w = \rho$.

The exponential prior $\exp(-\lambda \mu^T w)$ is conjugate and the posterior can be analytically calculated as follows

$$p(w \mid R, \rho, \lambda) \propto \exp\left(-\frac{1}{2}\|\rho - Rw\|^2 - \lambda \mu^T w\right)$$

$$\propto \exp\left(-\frac{1}{2}w^T R^T R w + w^T R^T \rho - \lambda \mu^T w\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(w - (R^T R)^{-1}(R^T \rho - \lambda \mu)\right)^T R^T R \left(w - (R^T R)^{-1}(R^T \rho - \lambda \mu)\right)\right).$$

By combining the likelihood and exponential prior p, we get the normal posterior with mean $(R^T R)^{-1}(R^T \rho - \lambda \mu)$ and covariance $(R^T R)^{-1}$. The resulting linear model is

$$\tilde{\rho} = Rw + \epsilon, \text{ where } w \sim f(w), \ \epsilon \sim N(0, \sigma_e^2 I). \tag{4}$$

where $\tilde{\rho} = R(R^T R)^{-1}(R^T \rho - \lambda \mu)$. The corresponding optimization problem is then

$$\underset{w}{\text{minimize}} \quad \|\tilde{\rho} - Rw\|_2^2 + \tau \phi(w) \tag{5}$$

Sparsity-inducing prior $f(w)$ and the corresponding penalty $\phi(w)$ leads to stable numerical solution robust to estimation errors in covariance and allow for sparse portfolios.

Since matrix R can be ill-conditioned in practice, we use a QR-decomposition to calculate $\tilde{\rho}$ which has better numerical properties compared to a LU decomposition Golub and Van Loan [1996]. Further, as we have discussed earlier by adding regularization terms for the weight parameters is equivalent to assuming uncertainty in the input data and thus is robust to perturbations in the R matrix.

3 Sparsity-Inducing Portfolio Priors

Bayesian formulation of portfolio selection problem allows to gain insight and to provide an alternative interpretation of the constraints and the corresponding penalty terms. Additionally, it allows to quantify uncertainty over the portfolio weights. Fully Bayesian inference that relies on MCMC algorithms allows to calculated credible intervals and thus to assess uncertainty. Efficient MCMC algorithms can be constricted by exploiting latent variable tricks. For example, the ℓ_1 penalty $\phi(w) = \|w\|_1 = \sum_{i=1}^n |w_i|$ corresponds to Laplace prior distribution $f(w_i) = (1/(2b))\exp(-|w_i|/b)$. Using a latent variable trick allows us to re-write this prior as a scale mixture of normals [Andrews and Mallows, 1974, West, 1987, Carlin and Polson, 1991].

Specifically, introduce latent variable γ_i with an exponential distribution,

$$w_i \mid \sigma^2, \gamma \sim N(0, \gamma\sigma^2)$$
$$\gamma \mid \lambda \sim Exp(\lambda^2/2)$$
$$\sigma^2 \sim \pi(\sigma^2).$$

There is an equivalence with the ℓ_1 penalty obtained by integrating out γ

$$p(w_i \mid \sigma^2, \lambda) = \int_0^\infty \frac{1}{\sqrt{2\pi\gamma\sigma^2}} \exp\left(-\frac{w_i^2}{2\sigma^2\gamma}\right) \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2\gamma}{2}\right) d\gamma = \frac{\lambda}{2\sigma} \exp(-\lambda|w_i|/\sigma).$$

In general, many widely used priors can be represented as variance-mean mixtures, using latent variable. The resulting model is linear with heteroscedastic errors [Polson and Scott, 2013]:

$$p(w_i \mid \lambda) = \int_0^\infty \phi(w_i \mid \mu + \kappa\gamma_i^{-1}, \lambda^2\gamma_i^{-1}) p(\gamma_i) d\gamma_i,$$

where $\phi(a \mid m, v)$ is the density function of normal variable with mean m and variance v.

3.1 Elastic Net Priors

The original BL Black and Litterman [1992] framework combines model-driven and traditional portfolio management into in a joint framework. Simply put, it allows us to combine predictions from quantitative models together with with investor beliefs.

We modify this approach and introduce a sparse Black-Litterman selection. Our Bayesian methodology updates opinions of an investor with model-based predictions to form a new set of predictions while maintaining sparse portfolio.

$$p(w_i \mid \lambda_1, \lambda_2, \sigma^2) = \frac{\lambda_1}{2\sigma} \exp(-\lambda_1|w_i|/\sigma) \frac{\sqrt{\lambda_2}}{\sqrt{2\pi\sigma}} \exp(-\frac{\lambda_2}{\sigma^2} w_i \Omega^{-1} w_i).$$

Here we have two regularization parameters λ_1 and λ_2 to choose. The parameter λ_2 is set proportional to the strength of the investor's beliefs in their side-information and not bu the cross-validation. A Gibbs sampler can be constructed by representing the Laplace prior as a scale mixture of normal Carlin and Polson [1991]. For an alternative empirical Bayes approach to select values of the hyperparameters for cases when design matrix is highly collinear see Liu et al. [2018].

The posterior MAP estimate can be computed as a fast convex optimization problem by combining the two normal terms and then using the LARS algorithm. This provide a scalable version of the Black-Litterman that allow us to perform selection. Similar approach was considered by Ho et al. [2015] who shoed that elastic net regularization allows to improve out-of-sample performance of the mean-variance portfolio. We can still use LARS and thus see how stocks are sequentially added (without finding an optimal λ_1 by cross-validation).

3.2 Horseshoe Priors

Horseshoe priors belongs to a class of global-local class of priors and is defined by global parameter τ that does not depend on index i and local parameter λ_i which is different for each parameter w_i. The prior is defined by

$$w_i \mid \lambda_i, \tau \sim N(0, \tau^2 \lambda_i^2).$$

The global hyper-parameter τ shrinks all parameters towards zero. The prior for the local parameter λ_i has a tail that decays slower than an exponential rate and thus allows w_i not to be shrunk. A horseshoe prior assumes half-Cauchy distribution over λ_i and τ

$$\lambda_i \sim C^+(0,1), \quad \tau \sim C^+(0,1).$$

Being constant at the origin, the half-Cauchy prior has nice risk properties near the origin [Polson and Scott, 2009]. Polson and Scott [2010] warn against using empirical Bayes or cross-validation approaches to estimate τ, due to the fact that MLE estimate of τ is always in danger of collapsing to the degenerate $\hat{\tau} = 0$ [Tiao and Tan, 1965].

A feature of the horseshoe prior is that it possesses both tail-robustness and sparse-robustness properties [Bhadra et al., 2017a]; meaning that an infinite spike at the origin and very heavy tail that still ensures integrability. The horseshoe prior can also be specified as

$$w_i \mid \lambda_i, \tau \sim N(0, \lambda_i^2), \quad \lambda_i \mid \tau \sim C^+(0, \tau), \quad \tau \sim C^+(0,1)$$

The log-prior of the horseshoe cannot be calculated analytically, but a tight lower bound [Carvalho et al., 2010] can be used instead

$$\phi_{HS}(w_i|\tau) = -\log p_{HS}(w_i|\tau) \geq -\log \log \left(1 + \frac{2\tau^2}{w_i^2} \right). \tag{6}$$

The motivation for the horseshoe penalty arises from the analysis of the prior mass and influence on the posterior in **both** the tail and behavior at the origin. The latter provides the key determinate of the sparsity properties of the estimator.

When Metropolis-Hasting MCMC is applied to horseshoe regression, it suffers from sampling issues. The funnel shape geometry of the horseshoe prior is makes it challenging for MCMC to efficiently explore the parameter space. Piironen and Vehtari [2017] proposed to replace Cauchy prior with half-t prior with small degrees of freedom and showed improved convergence behavior for NUTS sampler Hoffman and Gelman [2014]. Makalic and Schmidt [2016] proposed using a scale mixture representation of half-Cauchy which leads to conjugate hierarchy and allows a Gibbs sample to be used. Johndrow et al. [2017] proposed two MCMC algorithms to calculate posteriors for horseshoe priors. The first algorithm addresses computational cost problem in high dimensions by approximating matrix-matrix multiplication operations. For further details on computational issues and packages for horseshoe sampling, see Bhadra et al. [2017b]. An issue of high dimensionality was also addressed by Bhattacharya et al. [2016].

One approach is to replace the thick-tailed half-Cauchy prior over λ_j with half-t priors using small degrees of freedom. This leads to the sparsity-sampling efficiency trade-off

problem. Larger degrees of freedom for a half-t distribution will lead to more efficient sampling algorithms, but will be less sparsity inducing. For cases with large degrees of freedom, tails of half-t are slimmer and we are required to choose large τ to accommodate large signals. However, priors with a large τ are not able to shrink coefficients towards zero as much.

3.3 Spike-and-slab Prior

Spike-and-slab is another sparsity inducing prior widely used in Bayesian analysis. It assumes that the prior is a mixture of point-mass δ distribution and Gaussian distribution Polson and Sun [2019]

$$ w_i \mid a, \sigma^2 \sim (1-a)\delta_0 + aN\left(0, \sigma^2\right). $$

Here $a \in (0, 1)$ controls the overall sparsity in w and σ^2 allows for non-zero weights. By setting $w_i = \gamma_i \alpha_i$, we get a Bernoulli-Gaussian mixture model given by

$$ \begin{aligned} \gamma_i \mid a &\sim \text{Bernoulli}(a) ; \\ \alpha_i \mid \sigma^2 &\sim N\left(0, \sigma^2\right). \end{aligned} \tag{7} $$

Since γ and α are independent, we can write the joint density function as a product

$$ \begin{aligned} p\left(\gamma, \alpha \mid a, \sigma^2\right) &= \prod_{i=1}^{p} p\left(\alpha_i, \gamma_i \mid a, \sigma^2\right) \\ &= a^{\|\gamma\|_0}(1-a)^{p-\|\gamma\|_0}\left(2\pi\sigma^2\right)^{-\frac{p}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{p}\alpha_i^2\right\}. \end{aligned} $$

Here $\|\gamma\|_0 = \sum_i I(\gamma_i \neq 0)$ is the number of non-zero entries in the vector γ, and p is the length of vector w. It can be shown that finding the MAP estimator for the linear model given by Equation 4 with Spike-and-Slab prior is equivalent to solving the following optimization problem for γ and α Soussen et al. [2011], Polson and Sun [2019]

$$ \underset{\gamma, \alpha}{\text{minimize}} \quad \|\tilde{\rho} - R_\gamma \alpha_\gamma\|_2^2 + \frac{\sigma_e^2}{\sigma^2}\|\alpha\|_2^2 + 2\sigma_e^2\log\left(\frac{1-a}{a}\right)\|\gamma\|_0. $$

Here $R_\gamma := [R_i]_{i \in S}$ is the matrix with columns that have index inside set S, and S is the set of "active explanatory variables" with $S = \{i : \gamma_i = 1\}$ and $\alpha_\gamma := (\alpha_i)_{i \in S}$ be their corresponding coefficients.

3.4 Laplace Prior (Lasso)

A double exponential (Laplace) prior distribution [Carlin and Polson, 1991] for each weight w_i was previously shown to be an effective mechanism to regularize the portfolio [Brodie et al., 2009] and to incorporate investor's preferences for the number of assets in the optimal portfolio [Puelz et al., 2015].

$$ p(w_i \mid b) = 1/(2b)\exp(-|w_i|/b). $$

The log-posterior is then given by

$$\log p(w \mid R, \tilde{\rho}, b) \propto \|\tilde{\rho} - Rw\|_2^2 + \frac{\sigma_e^2}{b}\|w\|_1.$$

For $b > 0$, the posterior mode is equivalent to the ℓ_1-penalized estimate with $\lambda = \sigma_e^2/b$. Large variance b of the prior is equivalent to the small penalty weight λ in the ℓ_1-penalized objective function.

Park and Casella [2008] represent the Laplace prior is a scale Normal mixture to develop a Gibbs sampler that iteratively samples w and b to estimate joint distribution over $(\hat{w}, \hat{b})$. Thus, we do not need to apply cross-validation to find optimal value of b, the Bayesian algorithm calculates it automatically.

Given data $D = (R, \tilde{\rho})$, where R is the $n \times p$ matrix of standardized regressors and y is the n-vector of outputs. Implement a Gibbs sampler for this model when Laplace prior is used for model coefficients w_i. Use scale mixture normal representation.

$$w \mid \sigma^2, \tau_1, \dots, \tau_p \sim N(0, \sigma^2 D_\tau)$$
$$D_\tau = \text{diag}(\tau_1^2, \dots, \tau_p^2)$$
$$\tau_i^2 \mid \lambda \sim Exp(\lambda^2/2)$$
$$p(\sigma^2) \propto 1/\sigma^2.$$

Then the complete conditional required for Gibbs sampling are given by

$$w \mid D, D_\tau \sim N(A^{-1}R^T y, \sigma^2 A^{-1}), \quad A = R^T R + D_\tau^{-1}$$
$$\sigma^2 \mid w, D, D_\tau \sim \text{InverseGamma}\left((n-1)/2 + p/2, (\tilde{\rho} - Rw)^T(\tilde{\rho} - Rw)/2 + w^T D_\tau^{-1} w/2\right)$$
$$1/\tau_j^2 \mid w_j, \lambda \sim \text{InverseGaussian}\left(\sqrt{\frac{\lambda^2 \sigma^2}{w_j^2}}, \lambda^2\right)$$

The formulas above assume that R is standardized, e.g. observations for each feature are scaled to be of mean zero and standard deviation one, and $\tilde{\rho}$ is assumed to be centered.

We can use initialize the parameters as follows

$$w = (R^T R + I)^{-1} R^T \tilde{\rho}$$
$$q = \tilde{\rho} - Rw$$
$$\sigma^2 = q^T q/n$$
$$\tau^{-2} = 1/w^2$$
$$\lambda = p\sqrt{\sigma^2}/\|w\|_1.$$

where n is number of observations and p is number of columns (inputs) in matrix X.

There are several efficient optimization algorithms to compute mode of the posterior distribution for a Laplace prior. The most widely used approaches in applications are LARS [Efron et al., 2004] and coordinate descent [Friedman et al., 2010]. The advantage

of LARS compared to other optimization techniques is that it provides a way to compute the sequence of solutions for different values of the penalty weight $2\sigma^2/b$. Coordinate descent algorithm which updates one parameter at a time, holding the others fixed was shown to be more computationally efficient.

Finally, it is important to note the posterior contraction properties of all of these sparsity priors. Not all of the priors have similar minimax rates and properties. For example, lasso can underperform priors such as the horseshoe. For example, [Castillo et al., 2015] analyses Bayesian linear regression with sparsity-inducing prior of the following form

$$(S,\beta) \sim \pi_p(|S|) \binom{p}{|S|}^{-1} g_s(\beta_S)\delta_0(\beta_{\bar{S}})$$

here $S \subset \{1,\ldots,p\}$, is a random subset of cardinality $|S|$ and $\beta_S = \{\beta_i \ : \ i \in S\}$, and $\bar{S} = \{1,\ldots,p\} - S$. The prior π_p performs the model selection by inducing sparsity and g_s density model only non-zero coordinates. Song and Liang [2017] extend this analysis and to the cases when shrinkage priors have heavy and fat tail and also show how estimation algorithm for hierarchical models with spars-inducing priors adopt to unknown σ_e.

Spike and slab priors are a special case of this prior with π_p being the Binomial and g_s being the product of Laplace of Normal univariate distributions.

They show that under certain conditions on the design matrix, the posterior distribution is shown to contract at the optimal rate for recovery of the unknown sparse vector, and to give optimal prediction of the response vector, when g_S is the product of Laplace densities $(1/2)\lambda \exp(-\lambda|\beta|)$ and the inverse scale parameter is bounded by $\|X\|/p \leq \lambda \leq 2\bar{\lambda}$, where $\bar{\lambda} = 2\|X\|\sqrt{\log p}$, with $\|X\|^2 = \max_i \text{trace}(R^T R)_i$.

Van Der Pas et al. [2014] show that the τ parameter of the horseshoe can be interpreted as the proportion of nonzero weights up to a logarithmic factor. Further, they show that when true sparsity level p_n is know and τ is set to $(p_n/n)\sqrt{\log(n/p_n)}$ then the horseshoe estimator attains the minimax ℓ_2 risk and posterior contraction is bounded above by this number, when p_n is estimated from data (empirical Bayes) or a hierarchical prior is used [van der Pas et al., 2017a].

van der Pas et al. [2017b] show that the horseshoe estimator (the posterior mean) leads to credible balls and marginal credible intervals that have an optimal size if the sparsity level of the prior is set correctly. Their proofs are derived for the posterior under the sparse normal means problem.

4   Applications

Our sparse portfolio selection model is to construct optimal portfolios made by three models with sparsity-inducing priors; Laplace, horseshoe [Carvalho et al., 2010], and spike-and-slab. The LARS algorithm Efron et al. [2004] is used to find posterior mode for the Laplace model, MCMC to generate samples form the horseshoe Hahn et al. [2019] model, and Single Best Replacement (SBR) Polson and Sun [2019] algorithm to find the posterior mode of the spike-and-slab.

We demonstrate how Laplace regularized portfolios allocations and the corresponding LARS algorithms leads to an intuitive way to select an optimal portfolio and assign

a selection order of the stocks to be included into the portfolio. We use daily returns from three different portfolios. One portfolio corresponds to a widely used stock index (SP100) and two portfolios of stocks managed by two different hedge funds, namely Viking Global Investors and Renaissance Technologies. We use the top 50 holdings of each portfolio and apply our selection algorithm to design a sparse portfolio with minimal risk level while guaranteeing to out-perform the SP500 index. We use daily returns during the period from 2016-02-23 to 2018-02-15 (500 trading days) as our training data and returns for the period 2018-02-16 to 2019-02-22 for calculating out-of-sample performance of our portfolios. An optimal penalty parameter λ can be calculated using cross-validation, but we also show that output of the LARS algorithm allows us to infer λ given the desired number of assets in the portfolio.

4.1 Small-sized Stock Portfolio

First, we demonstrate how the LARS algorithm finds the posterior mode under ℓ_1 regularization and also provides a natural ranking of the importance of individual stocks. At every step of the LARS algorithm, a new variable enters the active set and thus it performs the same number of steps as the number of variables. The order in which LARS adds the variables corresponds to their importance, implying that the variables added in the beginning lead to a model that fits the training data well and of a low variance.

To empirically evaluate our selection procedure we take the top nine holdings from the SP100, Renaissance and Viking Global. Then the LARS algorithm adds one stock at a time to the portfolio and evaluates the out-of-sample performance at each time. We select an optimal portfolio that leads to the best out-of-sample returns. Figure 1(a) shows the weights assigned by the LARS algorithm at each iteration and the step at which the optimal portfolio was achieved.

Visualizing outputs of the LARS algorithms gives an inventor a way to interpret the importance of each of the stocks in the portfolio and suggest modifications. If an investor wants a smaller portfolio he/she can remove stocks added later by the algorithm.

The next question is whether the LARS selected portfolio does out-perform the naive equally weighted portfolio or traditional Markowitz portfolio selected by solving problem 1. We compare our optimal portfolio to performance of the SP500 index. Figure 1(b) shows the cumulative return (growth of $1 invested) of the LARS selected optimal portfolio and compares it with the naive and SP500 portfolios.

Table 1 shows the out-of-sample mean and standard deviation of the returns as well as Sharpe ratio for the selected portfolios.

14

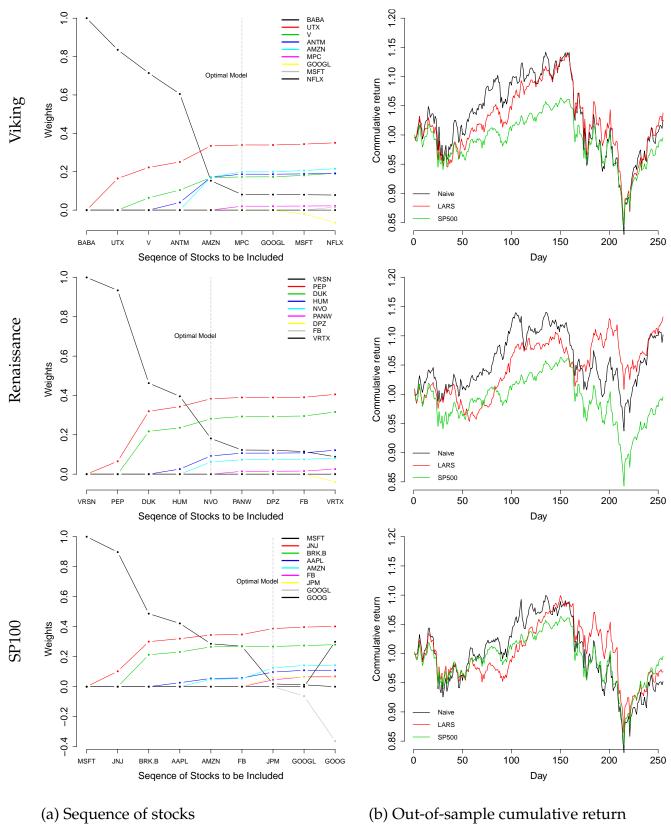(a) Sequence of stocks   (b) Out-of-sample cumulative return

Figure 1: (a) Sequence of stocks to be added to the portfolio as the function of number of stocks allowed (b) Out-of-sample cumulative return of an optimal LARS portfolio compared with naive portfolio and SP500

	Viking			Renaissance			SP100			
	Naive	LARS	QP	Naive	LARS	QP	Naive	LARS	QP	SP500
μ	0.066	0.12	0.12	0.1	0.099	0.065	-0.029	-0.011	-0.038	0.0086
σ	4.033	3.67	3.65	2.8	2.383	2.349	3.779	3.467	3.476	2.6004
μ/σ	4.124	7.94	8.37	9.3	10.466	7.026	-1.95	-0.764	-2.787	0.8376

Table 1: Out of sample performance of LARS selected portfolio compared with Naive portfolio and SP500

The LARS algorithms selected 5 stocks (VRSN, PEP, DUK, HUM and NVO) and the resulting portfolio is less risky and has the same average return when compared to the naive portfolio.

Plots shown in Figure 1 allow us to visualize the ranking of the stocks in the portfolio and to allow investor to decide how to increase or decrease the number of positions in a portfolio. Further, the optimal allocations calculated by LARS lead to a portfolio with lower risk (standard deviation of 0.008) and of higher return (mean of 0.0005) when compared to SP500 and naive equal weights allocations.

4.2 Large-sized Stock Portfolio

To demonstrate further, how our portfolio allocation algorithm can be used for selecting from a larger sets of stocks, we again compare models with Laplace (ℓ_1), Horseshoe and ℓ_0 regularization. We compare the shrinkage effect and empirical out-of-sample performance of these three selection approaches.

The question is whether the weight shrinkage introduced by the LARS algorithm does effect the portfolio performance and whether Horseshoe or ℓ_0 selectors lead to sparser portfolios. We select portfolios using the LARS for Laplace prior (ℓ_1), MCMC algorithm for Horseshoe prior, and Single Best Replacement (SBR) algorithm for Spike-and-Slab prior (ℓ_0), and non-regularized least-squares approach. We apply all four algorithms to select portfolio from top 35 holdings of Viking and Renaissance hedge funds as well as from SP100 stocks.

Table 2 shows the out-of-sample mean (μ), standard deviation (σ), and sharpe ratio (μ/σ) of the daily returns multiplied by 252 as well as the number of stocks selected ($\|w\|_0$). Figure 2 shows the cumulative return.

	Viking				Renaissance				SP100			
	LARS	HS	LM	L0	LARS	HS	LM	L0	LARS	HS	LM	L0
μ	0.13	0.14	0.16	0.16	0.14	0.17	0.11	0.18	0.051	0.055	0.077	0.066
σ	3.21	3.2	3.3	3.23	2.19	2.14	2.26	2.05	2.401	2.386	2.533	2.52
μ/σ	10.45	10.96	12.02	12.34	16.47	20.4	12.72	22.15	5.39	5.779	7.613	6.578
$\|w\|_0$	22	11	31	11	25	12	32	4	16	9	32	4

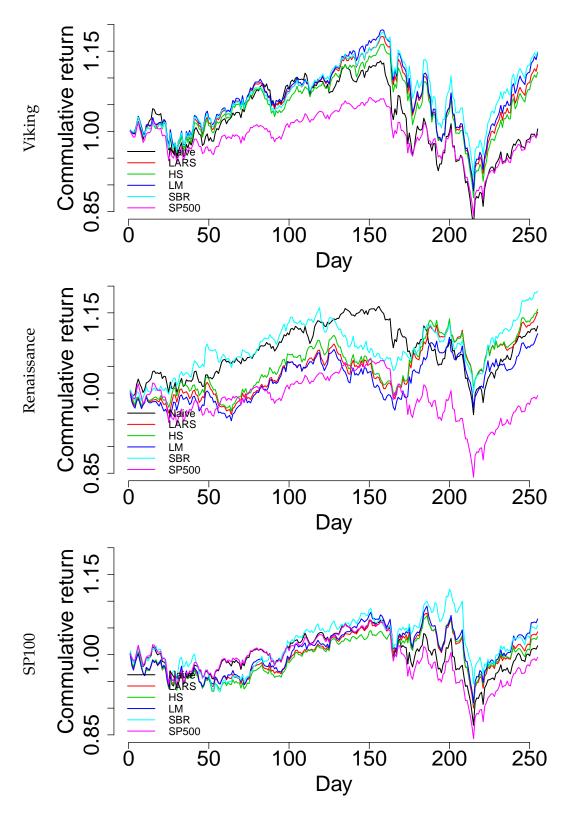Table 2: Out of sample performance of LARS selected portfolio compared with Naive portfolio and SP500

Figure 2: Out-of-sample cumulative return of an optimal LARS portfolio compared with naive portfolio and SP500

Horseshoe and ℓ_0 selectors do out-perform the LARS selector and lead to sparser portfolios. From practical standpoint, the ℓ_0 selector not only leads to the best performing and the most sparse portfolio, it is also the easier to use when compared to the horseshoe selector. ℓ_0 requires investor to specify one parameter instead of two as in the horseshoe and the penalty term in ℓ_0 is arguably more interpretable.

5   Discussion

The goal of our paper is to present an extension of the Black-Litterman portfolio framework. A quadratic programming optimization problem is recast as an inference problem for hierarchical Bayesian linear model. We have shown how linear constraints of the optimization problem can be formulated as the product of exponential priors and sparsity priors. The main advantage of our formulation is the ability to incorporate an investor's subjective opinion about which assets to be included into the portfolio. Specifically, we demonstrated how sparsity priors stabilize the portfolio selection and to select a small number of stocks to be included into the portfolio. The sparsity priors correspond to investors' preferences for portfolios with small number of stocks. We used our hierarchical formulation to demonstrate empirical performance of several sparsity-inducing priors versus the SP100 index. We have shown that Horseshoe and Spike-and-Slab (ℓ_0 penalty) priors not only lead to portfolios with smaller number of stocks but also have better out-of-sample performance when compared to Laplace prior (ℓ_1 penalty) and traditional Markowitz (un-regularized) portfolio selection procedure. Inclusion of short positions in regularized portfolio also leads to better performance and lower risk while maintaining stability of the portfolio (no extremely large weights). From a practical standpoint ℓ_0 penalty leads to the the best performing portfolio and requires an investor to specify only one parameter.

There are a number of direction for future research. For example, equivalence between non-convex penalties and hierarchical linear models, as well as further study of elastic net formulations of the Black-Litterman framework.

References

Fischer Black and Robert Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

David Puelz, P Richard Hahn, and Carlos M Carvalho. Sparse mean-variance portfolios: A penalized utility approach. *arXiv preprint arXiv:1512.02310*, 2015.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

Nicholas G Polson and Bernard V Tew. Bayesian portfolio selection: An empirical analysis of the S&P 500 index 1970–1996. *Journal of Business & Economic Statistics*, 18(2):164–173, 2000.

Guy VG Stevens. On the inverse of the covariance matrix in portfolio analysis. *The Journal of Finance*, 53(5):1821–1827, 1998.

Dimitris Bertsimas, Vishal Gupta, and Ioannis Ch Paschalidis. Inverse optimization: A new perspective on the black-litterman model. *Operations research*, 60(6):1389–1403, 2012.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The Horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

P Richard Hahn, Jingyu He, and Hedibert F Lopes. Efficient sampling for Gaussian linear regression with arbitrary priors. *Journal of Computational and Graphical Statistics*, 28(1): 142–154, 2019.

Nicholas G Polson and Lei Sun. Bayesian l 0-regularized least squares. *Applied Stochastic Models in Business and Industry*, 35(3):717–731, 2019.

Shmuel Kandel, Robert McCulloch, and Robert F Stambaugh. Bayesian inference and portfolio efficiency. *The Review of Financial Studies*, 8(1):1–53, 1995.

Marine Carrasco and Nérée Noumon. Optimal portfolio selection using regularization. Technical report, Working Paper, 2011.

Bruno De Finetti. *Il problema dei pieni*. Istituto italiano degli attuari, 1940.

Andre F Perold. Large-scale portfolio optimization. *Management Science*, 30(10):1143–1160, 1984.

Nicholas Barberis. Investing for the long run when returns are predictable. *Journal of Finance*, 55(1):225–264, 2000.

Carlos M Carvalho, Hedibert F Lopes, and Omar Aguilar. Dynamic stock selection strategies: A structured factor model framework. *Bayesian Statistics*, 9:1–21, 2011.

Omar Aguilar and Mike West. Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3):338–357, 2000.

Mila Getmansky, Peter A Lee, and Andrew W Lo. Hedge funds: A dynamic industry in transition. *Annual Review of Financial Economics*, 7:483–577, 2015.

Shmuel Kandel and Robert F Stambaugh. On the predictability of stock returns: an asset-allocation perspective. *The Journal of Finance*, 51(2):385–424, 1996.

Björn Fastrich, Sandra Paterlini, and Peter Winker. Constructing optimal sparse portfolios using regularization methods. *Computational Management Science*, 12(3):417–434, 2015.

Gilles Gasso, Alain Rakotomamonjy, and Stéphane Canu. Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *IEEE Transactions on Signal Processing*, 57(12):4686–4698, 2009.

Margherita Giuzio and Sandra Paterlini. Un-diversifying during crises: Is it a good idea? *Computational Management Science*, Dec 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Tong Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *The Annals of Statistics*, 37(5A):2109–2144, 2009.

Nicholas G Polson, James G Scott, and Jesse Windle. The Bayesian Bridge. *Journal of the Royal Statistical Society: Series B*, 76(4):713–733, 2014.

Ildiko E Frank and Jerome H Friedman. A Statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

Thibault Bourgeron, Edmond Lezmi, and Thierry Roncalli. Robust asset allocation for robo-advisors. *Available at SSRN 3261635*, 2018.

Eric Jacquier and Nicholas G Polson. Asset allocation in finance: A Bayesian perspective. *Hierarchinal models and MCMC: a Tribute to Adrian Smith*, pages 56–59, 2012.

Amine Ismail and Huyên Pham. Robust Markowitz mean-variance portfolio selection under ambiguous covariance matrix. *Mathematical Finance*, 29(1):174–207, 2019.

Serhiy Kozak, Stefan Nagel, and Shrihari Santosh. Shrinking the cross-section. *Journal of Financial Economics*, 2019.

Doron Avramov and Guofu Zhou. Bayesian portfolio analysis. *Annu. Rev. Financ. Econ.*, 2(1):25–47, 2010.

Victor DeMiguel, Lorenzo Garlappi, Francisco J Nogales, and Raman Uppal. A generalized approach to portfolio optimization: Improving performance by constraining portfolio norms. *Management Science*, 55(5):798–812, 2009.

Ravi Jagannathan and Tongshu Ma. Risk reduction in large portfolios: Why imposing the wrong constraints helps. *The Journal of Finance*, 58(4):1651–1683, 2003.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Miguel Sousa Lobo, Maryam Fazel, and Stephen Boyd. Portfolio optimization with linear and fixed transaction costs. *Annals of Operations Research*, 152(1):341–365, 2007.

Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences*, 106(30):12267–12272, 2009.

Jianqing Fan, Jingjin Zhang, and Ke Yu. Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498):592–606, 2012.

Huan Xu, Constantine Caramanis, and Shie Mannor. Robust regression and lasso. In *Advances in Neural Information Processing Systems*, pages 1801–1808, 2009.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

HG Golub and Charles F Van Loan. Matrix computations, johns hopkins uni. *Press, London*, 1996.

David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the RSS, Series B*, pages 99–102, 1974.

Mike West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

Bradley P Carlin and Nicholas G Polson. Inference for nonconjugate Bayesian models using the Gibbs sampler. *Canadian Journal of statistics*, 19(4):399–405, 1991.

Nicholas G Polson and James G Scott. Data augmentation for non-Gaussian regression models using variance-mean mixtures. *Biometrika*, 100(2):459–471, 2013.

Chang Liu, Yue Yang, Howard Bondell, and Ryan Martin. Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior. *arXiv preprint arXiv:1810.00739*, 2018.

Michael Ho, Zheng Sun, and Jack Xin. Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM Journal on Financial Mathematics*, 6(1):1220–1244, 2015.

Nicholas G Polson and James G Scott. Alternative global–local shrinkage rules using hypergeometric–beta mixtures. *Technical report 14*, 2009.

Nicholas G Polson and James G Scott. Shrink Globally, Act Locally: Sparse Bayesian regularization and prediction. *Bayesian statistics*, 9:501–538, 2010.

George C Tiao and WY Tan. Bayesian analysis of random-effect models in the analysis of variance. I. posterior distribution of variance-components. *Biometrika*, 52(1/2):37–53, 1965.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. The Horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017a.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the Horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Enes Makalic and Daniel F Schmidt. A simple sampler for the Horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2016.

James E Johndrow, Paulo Orenstein, and Anirban Bhattacharya. Bayes shrinkage at gwas scale: Convergence and approximation theory of a scalable mcmc algorithm for the horseshoe prior. *arXiv preprint arXiv:1705.00841*, 2017.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon T Willard. Lasso meets Horseshoe. *Statistical Science (to appear)*, 2017b.

Anirban Bhattacharya, Antik Chakraborty, and Bani K Mallick. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991, 2016.

Charles Soussen, Jérôme Idier, David Brie, and Junbo Duan. From Bernoulli–Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59 (10):4572–4584, 2011.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

Ismaël Castillo, Johannes Schmidt-Hieber, and Aad Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Qifan Song and Faming Liang. Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*, 2017.

Stéphanie L Van Der Pas, Bas JK Kleijn, and Aad W Van Der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618, 2014.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Adaptive posterior contraction rates for the Horseshoe. *Electronic Journal of Statistics*, 11(2):3196–3225, 2017a.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274, 2017b.