Journal of Machine Learning Research 20 (2019) 1–39 Submitted 5/18; Revised 11/18; Published 3/19

Prediction Risk for the Horseshoe Regression

Anindya Bhadra BHADRA@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Jyotishka Datta JD033@UARK.EDU
Department of Mathematical Sciences
University of Arkansas
Fayetteville, AR 72701, USA

Yunfan Li LI896@PURDUE.EDU
Department of Statistics
Purdue University
West Lafayette, IN 47907, USA

Nicholas G. Polson NGP@CHICAGOBOOTH.EDU
Booth School of Business
University of Chicago
Chicago, IL 60637, USA

Brandon Willard BRANDONWILLARD@GMAIL.COM
Booth School of Business
University of Chicago
Chicago, IL 60637, USA

Editor: Robert McCulloch

Abstract

We show that prediction performance for global-local shrinkage regression can overcome two major difficulties of global shrinkage regression: (i) the amount of relative shrinkage is monotone in the singular values of the design matrix and (ii) the shrinkage is determined by a single tuning parameter. Specifically, we show that the horseshoe regression, with heavy-tailed component-specific local shrinkage parameters, in conjunction with a global parameter providing shrinkage towards zero, alleviates both these difficulties and consequently, results in an improved risk for prediction. Numerical demonstrations of improved prediction over competing approaches in simulations and in a pharmacogenomics data set confirm our theoretical findings.

Keywords: Global-local Priors, Principal Components, Shrinkage Regression, Stein's Unbiased Risk Rstimate

1. Introduction

We develop theoretical results on prediction risk in the high-dimensional linear regression model

$$y = X\beta + \epsilon, \tag{1}$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}, \beta \in \mathbb{R}^p$, $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$ with $p > n$, and the design matrix X is assumed fixed. Let $\hat{\beta}$ denote the estimate of β based on the observed data y and design X. Let y^* denote a future observation generated from the same model, independent of y. Define the quadratic predictive risk

$$R = \mathbb{E}_{y^*,y|X,\beta}(y^* - X\hat{\beta})^2, \tag{2}$$

where the subscript denotes that the expectation is with respect to the data generating distribution, with X and β held fixed. We focus on comparing estimators $\hat{\beta}$ according to the criterion of Equation (2) in a *non-asymptotic* fixed n, fixed $p > n$ setting. Our approach follows the paradigm of Stein (1956), with risk results that are valid for all p and n, rather than asymptotic oracle properties. Our specific contribution, established by Theorem 5.1 and Corollary 5.1, is to identify the shortcomings of some commonly used global shrinkage estimators in prediction, with shrinkage driven by a single tuning parameter, and to demonstrate that under certain conditions, suitably-chosen component-specific local shrinkage parameters can result in theoretically lower predictive risk.

1.1. Connections with existing global shrinkage regression approaches

We define shrinkage estimators with a single tuning parameter as "global." Examples include ridge regression (Hoerl and Kennard, 1970) and principal components regression or PCR (Jolliffe, 1982), and they remain popular in prediction under the high-dimensional model of Equation (1). Shrinkage methods enjoy a number of advantages over simultaneous shrinkage and selection-based methods such as the lasso (Tibshirani, 1996) and comfortably outperform them in predictive performance in certain situations. Prominent among these is when the predictors are correlated and the resulting lasso estimate is unstable, whereas ridge or PCR estimates are not (see, e.g, the discussion in Chapter 3 of Hastie et al., 2009). On the theoretical side, Polson and Scott (2012a) used a representation devised by Frank and Friedman (1993) to show that many commonly used high-dimensional shrinkage regression estimates, such as the estimates of ridge regression, regression with g-prior (Zellner, 1986) and PCR, can be viewed as posterior means under a unified framework of "global" shrinkage prior on the regression coefficients that are suitably orthogonalized. Polson and Scott (2012a) also demonstrated that purely global shrinkage regression methods suffer from two major difficulties: (i) the amount of relative shrinkage is monotone in the singular values of the design matrix and (ii) the shrinkage is determined by a single tuning parameter. Both of these factors can translate to poor out of sample prediction performance, which they demonstrated numerically.

Polson and Scott (2012a) further provided numerical evidence that the difficulties mentioned above can be resolved by allowing "local," component-specific shrinkage terms, in conjunction with a global shrinkage parameter as used in ridge or PCR, giving rise to the

so-called "global-local" shrinkage regression models. Specifically, Polson and Scott (2012a) demonstrated by simulations that using the horseshoe prior of Carvalho et al. (2010) on the regression coefficients performed well over a variety of competitors in terms of predictive performance, including the lasso, ridge, PCR and sparse partial least squares (Chun and Keles, 2010). While these empirical results are encouraging, a theoretical investigation of the conditions required for the horseshoe regression to outperform a global shrinkage regression method such as ridge or PCR in terms of prediction has been lacking. Our work bridges this theoretical gap by developing formal tools for comparing the finite sample predictive risk for shrinkage methods.

1.2. Regression with non-convex penalties

The ℓ_1 or ℓ_2 penalties that correspond to lasso or ridge regression are convex. While this simplifies the computation, it also results in a number of drawbacks such as bias in estimating large signals (Fan and Li, 2001). This problem can be remedied using non-convex ℓ_q penalties for $0 < q < 1$, but this introduces other problems such as non-uniqueness of solutions and greater computational burden. Prominent examples of non-convex penalties include the smoothly clipped absolute deviation or SCAD (Fan and Li, 2001) and the minimax concave penalty of MCP (Zhang, 2010). In particular, the MCP estimate enjoys a number of asymptotic optimality properties and conditions required for an iterative computational algorithm to reach the global optimum are available (Mazumder et al., 2011). The optimality results, however, are valid only in an asymptotic regime with various assumptions on the design X and do not characterize finite sample risk properties. Nevertheless, the univariate MCP estimator is identical to the firm shrinkage estimator of Gao and Bruce (1997), who provide explicit finite sample expressions for predictive risk, a fact utilized later in Section 6.

1.3. Finite sample estimates of predictive risk

The quadratic risk in Equation (2) involves the future observation y^* and must be estimated. Developing a formal estimate based on the training data (X, y) to compare predictive performance of competing regression methods is important in both frequentist and Bayesian settings. This is because the frequentist tuning parameter or the Bayesian hyper-parameters can then be chosen to minimize the estimated predictive risk, if prediction of future observations is the main modeling goal. A *finite sample unbiased estimate* of R in Equation (2) is given by Stein's unbiased risk estimate or SURE (Stein, 1981).

We will focus on SURE as our estimate of R for the remainder of this article, which is an example of a model-based covariance penalty. Other examples of covariance penalties include Mallows' C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1974) and risk inflation criterion (Foster and George, 1994). Nonparametric penalties include the generalized cross validation of Craven and Wahba (1978), which has the advantage of being model free but usually produces a prediction error estimate with high variance (Efron, 1983). The relationship between the covariance penalties and nonparametric approaches were further explored by Efron (2004), who showed the covariance penalties to be a Rao-Blackwellized version of the nonparametric penalties. Thus, Efron (2004) concluded that model-based penalties such as SURE or Mallows' C_p (the two coincide for models where the fit is linear

in the response variable) offer substantially lower variance in estimating the prediction risk, assuming of course the model is true. From a computational perspective, calculating SURE, when it is explicitly available, is substantially less burdensome than performing cross validation, which usually requires several Monte Carlo replications. Furthermore, SURE, an estimate of quadratic risk in prediction, also has connections with the Kullback–Leiber risk for the predictive density (George et al., 2006).

1.4. Outline of main contributions

Our main contribution is to analyze the finite sample predictive risk of global shrinkage regression methods, examine where these methods fall short, and demonstrate a remedy using local shrinkage parameters. The main results are summarized as follows.

1.4.1. Theoretical findings

The key technique to our innovation is an orthogonalized representation first employed by Frank and Friedman (1993) that allows shrinkage regression estimates to be viewed as posterior means under some suitable priors. This is formulated in Section 2. Using this representation in Sections 3 and 4, we devise general, explicit and numerically stable techniques for computing SURE for regression models that can be employed to compare the performances of global as well as horseshoe regressions. We characterize the finite sample risk properties of all competing methods by computing expectations of SURE. Consequently, all results provided in our article are valid under minimal assumptions on the design matrix, similar to the risk results by Stein (1956), where the only requirement is $n > 2$. This is at a contrast with most existing results in linear regression focusing on asymptotic minimax risk that require various assumptions on the singular values of X (e.g., Raskutti et al., 2011; Castillo et al., 2015; Dobriban and Wager, 2017).

Using the developed tools for SURE, we provide explicit finite sample risk comparisons between the global ridge and global-local horseshoe regressions in Section 5, where for all methods the tuning parameter is chosen to optimize SURE. Specifically, we demonstrate that the horseshoe regression can outperform the optimal ridge regression in prediction when most true signals are zero, but a few are large. We also compare risk of the horseshoe regression with non-convex penalized likelihood approaches such as MCP in Section 6 and show that when most of the true signals are away from zero, the risk of MCP can be quite large, unlike that of the horseshoe regression.

1.4.2. Empirical findings

Extensive numerical results are provided in Section 7 and Supplementary Section S.1. Our simulation results treat three distinct regimes: (i) sparse-robust: where most true signals are zero and a few are large, (ii) null: where all signals are zero and (iii) dense: where all signals are large. Our major finding is that the horseshoe regression outperforms the other methods in (i). Moreover, it is not much worse than ridge in (iii) and adaptive lasso in (ii), which are usually the best performers in these settings. Being a shrinkage estimate, the results for the horseshoe are numerically stable, unlike that of the selection-based estimators in the dense case. We conclude with a demonstration on real data in Section 8 and by outlining some possible extensions of the current work in Section 9.

2. Shrinkage regression estimates as posterior means

Let $X = UDW^T$ be the singular value decomposition of the design matrix. Let $D = \text{diag}(d_1, \ldots, d_n)$ with $d_1 \geq \ldots \geq d_n > 0$ and $\text{Rank}(D) = \min(n, p) = n$. Define $Z = UD$ and $\alpha = W^T\beta$. Then the regression model of Equation (1) can be reformulated as:

$$y = Z\alpha + \epsilon. \tag{3}$$

The ordinary least squared (OLS) estimate of α is $\hat{\alpha} = (Z^TZ)^{-1}Z^Ty = D^{-1}U^Ty$. Following the original results by Frank and Friedman (1993), several authors have used the well-known orthogonalization technique (Polson and Scott, 2012a; Clyde et al., 1996; Denison and George, 2012) to demonstrate that the estimates of many shrinkage regression methods can be expressed in terms of the posterior mean of the "orthogonalized" regression coefficients α under the following hierarchical model:

$$(\hat{\alpha}_i \mid \alpha_i, \sigma^2) \overset{ind}{\sim} \mathcal{N}(\alpha_i, \sigma^2 d_i^{-2}), \tag{4}$$

$$(\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2), \tag{5}$$

with $\sigma^2, \tau^2 > 0$. The global term τ controls the amount of shrinkage and the fixed λ_i^2 terms depend on the method at hand. Given λ_i and τ, the estimate for β under the global shrinkage prior, denoted by $\tilde{\beta}$, can be expressed in terms of the posterior mean estimate for α as follows:

$$\tilde{\alpha}_i = \frac{\tau^2\lambda_i^2 d_i^2}{1 + \tau^2\lambda_i^2 d_i^2}\hat{\alpha}_i, \quad \text{and} \quad \tilde{\beta} = \sum_{i=1}^{n} \tilde{\alpha}_i w_i, \tag{6}$$

where $\tilde{\alpha}_i = E(\alpha_i \mid \tau, \lambda_i^2, X, y)$; w_i is a $p \times 1$ vector and is the ith column of the $p \times n$ matrix W and the term $\tau^2\lambda_i^2 d_i^2/(1+\tau^2\lambda_i^2 d_i^2) \in (0, 1)$ is the shrinkage factor. The expression from Equation (6) makes it clear that it is the orthogonalized OLS estimates $\hat{\alpha}_i$s that are shrunk. We shall show that this orthogonalized representation is also particularly suitable for calculating the prediction risk estimate. The reason is tied to the independence assumption that is now feasible in Equations (4) and (5). To give a few concrete examples, we note below that several popular shrinkage regression models fall under the framework of Equations (4–5):

1. For ridge regression, $\lambda_i^2 = 1, \forall i$, and we have $\tilde{\alpha}_i = \tau^2 d_i^2\hat{\alpha}_i/(1 + \tau^2 d_i^2)$.

2. For K component PCR, λ_i^2 is infinite for the first K components and then 0. Thus, $\tilde{\alpha}_i = \hat{\alpha}_i$ for $i = 1, \ldots, K$ and $\tilde{\alpha}_i = 0$ for $i = K + 1, \ldots, n$.

3. For regression with g-prior, $\lambda_i^2 = d_i^{-2}$ and we have $\tilde{\alpha}_i = \tau^2\hat{\alpha}_i/(1+\tau^2)$ for $i = 1, \ldots, n$.

This shows the amount of relative shrinkage $\tilde{\alpha}_i/\hat{\alpha}_i$ is constant in d_i for PCR and g-prior and is monotone in d_i for ridge regression. In none of these cases it depends on the OLS estimate $\hat{\alpha}_i$ (consequently, on y). In the next section we quantify the effect of this behavior on the prediction risk estimate.

3. Stein's unbiased risk estimate for global shrinkage regression

Define the fit $\tilde{y} = X\tilde{\beta} = Z\tilde{\alpha}$, where $\tilde{\alpha}$ is the posterior mean of α. As noted by Stein (1981), the fitted risk is an underestimation of the prediction risk, and SURE for prediction is defined as:

$$SURE = ||y - \tilde{y}||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \tilde{y}_i}{\partial y_i},$$

where the $\sum_{i=1}^{n}(\partial\tilde{y}_i/\partial y_i)$ term is also known as the "degrees of freedom" (Efron, 2004). By Tweedie's formula (Masreliez, 1975; Pericchi and Smith, 1992) that relates the posterior mean with the marginals; we have for a Gaussian model of Equations (4–5) that: $\tilde{\alpha} = \hat{\alpha} + \sigma^2 D^{-2}\nabla_{\hat{\alpha}} \log m(\hat{\alpha})$, where $m(\hat{\alpha})$ is the marginal for $\hat{\alpha}$. Noting $y = Z\hat{\alpha}$ yields $\tilde{y} = y + \sigma^2 U D^{-1}\nabla_{\hat{\alpha}} \log m(\hat{\alpha})$. Using the independence of α_is, the formula for SURE becomes

$$SURE = \sigma^4 \sum_{i=1}^{n} d_i^{-2} \left\{ \frac{\partial}{\partial \hat{\alpha}_i} \log m(\hat{\alpha}_i) \right\}^2 + 2\sigma^2 \sum_{i=1}^{n} \left\{ 1 + \sigma^2 d_i^{-2} \frac{\partial^2}{\partial \hat{\alpha}_i^2} \log m(\hat{\alpha}_i) \right\}. \quad (7)$$

Thus, the prediction risk estimate for shrinkage regression can be quantified in terms of the first two derivatives of the log marginal for $\hat{\alpha}$. Integrating out α_i from Equations (4–5) yields in all these cases,

$$(\hat{\alpha}_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \mathcal{N}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

The marginal of $\hat{\alpha}$ is given by

$$m(\hat{\alpha}) \propto \prod_{i=1}^{n} \exp\left\{ -\frac{\hat{\alpha}_i^2/2}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)} \right\},$$

which yields

$$\frac{\partial \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i} = \frac{-\hat{\alpha}_i}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)}; \qquad \frac{\partial^2 \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i^2} = \frac{-1}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)}. \quad (8)$$

Therefore, Equation (7) reduces to the following expression for SURE for global shrinkage regressions: $SURE = \sum_{i=1}^{n} SURE_i$, where,

$$SURE_i = \frac{\hat{\alpha}_i^2 d_i^2}{(1 + \tau^2\lambda_i^2 d_i^2)^2} + 2\sigma^2 \frac{\tau^2\lambda_i^2 d_i^2}{(1 + \tau^2\lambda_i^2 d_i^2)}. \quad (9)$$

From a computational perspective, the expression in Equation (9) is attractive, as it avoids costly matrix inversions. For a given σ one can choose τ to minimize the prediction risk, which amounts to a one-dimensional optimization. Note that in our notation, $d_1 \geq d_2 \ldots \geq d_n > 0$. Clearly, this is the SURE when λ_is are fixed and finite (e.g., ridge regression). For K component PCR, only the first K terms appear in the sum. The d_i terms are features of the design matrix X and one may try to control the prediction risk by varying τ. When $\tau \to \infty$, $SURE \to 2n\sigma^2$, the risk of prediction with ordinary least squares (unbiased). When $\tau \to 0$, we get the mean-only model (zero variance): $SURE \to \sum_{i=1}^{n} \hat{\alpha}_i^2 d_i^2$. Regression models with $\tau \in (0, \infty)$ represent a bias-variance tradeoff. Following are the two major difficulties of global shrinkage regression:

1. The first term of Equation (9) shows that $SURE$ is increased by those components for which $\hat{\alpha}_i^2 d_i^2$ is large. Choosing a large τ alleviates this problem, but at the expense of an $SURE_i$ of $2\sigma^2$ even for components for which $\hat{\alpha}_i^2 d_i^2$ is small (due to the second term in Equation (9)). Thus, it might be beneficial to *differentially minimize* the effect of the components for which $\hat{\alpha}_i^2 d_i^2$ is large, while ensuring those for which $\hat{\alpha}_i^2 d_i^2$ is small make a contribution less than $2\sigma^2$ to $SURE$. Yet, regression models with λ_i fixed, such as ridge, PCR, regression with g-priors, provide no mechanism for achieving this, since the relative shrinkage, defined as the ratio $\tilde{\alpha}_i/\hat{\alpha}_i$, equals $\tau^2\lambda_i^2 d_i^2/(1+\tau^2\lambda_i^2 d_i^2)$, and is solely driven by a single quantity τ.

2. Equation (6) shows that the relative shrinkage for $\hat{\alpha}_i$ is monotone in d_i; that is, those $\hat{\alpha}_i$ corresponding to a smaller d_i are necessarily shrunk more (in a relative amount). This is only sensible in the case where one has reasons to believe the low variance eigen-directions (i.e., principal components) of the design matrix are not important predictors of the response variables, an assumption that can be violated in real data (Polson and Scott, 2012a).

In the light of these two problems, we proceed to demonstrate that putting a heavy-tailed prior on λ_is, in combination with a suitably small value of τ to enable global-local shrinkage can resolve both these issues. The intuition behind this is that a small value of a *global* parameter τ enables shrinkage towards zero for all the components while the heavy tails of the *local* or component-specific λ_i terms ensure the components with large values of $\hat{\alpha}_i d_i$ are not shrunk too much, and allow the λ_i terms to be learned from the data. Simultaneously ensuring both of these factors helps in controlling the prediction risk for both the noise as well as the signal terms.

4. Stein's unbiased risk estimate for the horseshoe regression

The global-local horseshoe shrinkage regression of Polson and Scott (2012a) extends the global shrinkage regression models of the previous section by putting a local (component-specific), heavy-tailed half-Cauchy prior on the λ_i terms that allow these terms to be learned from the data, in addition to a global τ. The model equations become:

$$(\hat{\alpha}_i \mid \alpha_i, \sigma^2) \overset{ind}{\sim} \mathcal{N}(\alpha_i, \sigma^2 d_i^{-2}), \tag{10}$$

$$(\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2), \tag{11}$$

$$\lambda_i \overset{ind}{\sim} C^+(0,1), \tag{12}$$

with $\sigma^2, \tau^2 > 0$ and $C^+(0,1)$ denotes a standard half-Cauchy random variable with density $p(\lambda_i) = (2/\pi)(1+\lambda_i^2)^{-1}$. The posterior mean $\tilde{\alpha}$ and the regression estimate $\tilde{\beta}$ are then obtained analogously to Equation (6), with the only difference being one uses the posterior mean $\mathbb{E}(\lambda_i \mid \hat{\alpha}_i, \tau)$ instead of a fixed λ_i. The marginal prior on α_is that is obtained as a normal scale mixture by integrating out λ_is from Equations (11) and (12) is called the horseshoe prior (Carvalho et al., 2010). Improved mean square error over competing approaches in regression has been empirically observed by Polson and Scott (2012a) with horseshoe prior on α_is. The intuitive explanation for this improved performance is that a heavy tailed prior of λ_i leaves the large α_i terms of Equation (11) un-shrunk in the posterior, whereas

the global τ term provides shrinkage towards zero for all components (see, for example, the discussion by Polson and Scott, 2012b; Bhadra et al., 2017; Carvalho et al., 2010, and the references therein). However, no explicit formulation of the prediction risk under horseshoe shrinkage is available so far and we demonstrate below the heavy-tailed priors on λ_i terms, in addition to a global τ, can be beneficial in controlling the overall prediction risk.

Under the model of Equations (10–12), after integrating out α_i from the first two equations, we have,

$$(\hat{\alpha}_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \mathcal{N}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

We have, $p(\lambda_i) \propto 1/(1+\lambda_i^2)$. Thus, the marginal of $\hat{\alpha}$, denoted by $m(\hat{\alpha})$, is given up to a constant of proportionality by

$$\begin{aligned} m(\hat{\alpha}) =& \prod_{i=1}^{n} \int_0^\infty \mathcal{N}(\hat{\alpha}_i \mid 0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)) p(\lambda_i) d\lambda_i \\ \propto& (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \int_0^\infty \exp\left\{ -\frac{\hat{\alpha}_i^2 d_i^2/2}{\sigma^2(1+\tau^2 d_i^2\lambda_i^2)} \right\} \frac{d_i}{(1+\tau^2 d_i^2\lambda_i^2)^{1/2}} \frac{1}{1+\lambda_i^2} d\lambda_i. \end{aligned} \tag{13}$$

This integral involves the normalizing constant of a compound confluent hypergeometric distribution that can be computed using a result of Gordy (1998).

Proposition 4.1 *(Gordy, 1998). The compound confluent hypergeometric (CCH) density is given by*

$$\mathrm{CCH}(x; p, q, r, s, \nu, \theta) \;=\; \frac{x^{p-1}(1-\nu x)^{q-1}\{\theta + (1-\theta)\nu x\}^{-r} \exp(-sx)}{B(p,q) H(p,q,r,s,\nu,\theta)},$$

for $0 < x < 1/\nu$, where the parameters satisfy $p > 0$, $q > 0$, $r \in \mathbb{R}$, $s \in \mathbb{R}$, $0 \le \nu \le 1$ and $\theta > 0$. Here $B(p,q)$ is the beta function and the function $H(\cdot)$ is given by

$$H(p,q,r,s,\nu,\theta) = \nu^{-p} \exp(-s/\nu) \Phi_1(q, r, p+q, s/\nu, 1-\theta),$$

where Φ_1 is the confluent hypergeometric function of two variables, given by

$$\Phi_1(\alpha, \beta, \gamma, x_1, x_2) \;=\; \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} \frac{(\alpha)_m (\beta)_n}{(\gamma)_{m+n} m! n!} x_1^m x_2^n, \tag{14}$$

where $(a)_k$ denotes the rising factorial with $(a)_0 = 1, (a)_1 = a$ and $(a)_k = (a+k-1)(a)_{k-1}$.

We present our first result in the following theorem and show that the marginal $m(\hat{\alpha})$ and all its derivatives lend themselves to a series representation in terms of the first and second moments of a random variable that follows a CCH distribution. Consequently, we quantify SURE for the horseshoe regression.

Theorem 4.1 *Denote $m'(\hat{\alpha}_i) = (\partial/\partial\hat{\alpha}_i) m(\hat{\alpha}_i)$ and $m''(\hat{\alpha}_i) = (\partial^2/\partial\hat{\alpha}_i^2) m(\hat{\alpha}_i)$. Then, the following holds.*

A. SURE for the horseshoe shrinkage regression model defined by Equations (10–12) is given by $SURE = \sum_{i=1}^{n} SURE_i$, *where the component-wise contribution* $SURE_i$ *is given by*

$$SURE_i \;=\; 2\sigma^2 - \sigma^4 d_i^{-2}\left\{\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}\right\}^2 + 2\sigma^4 d_i^{-2}\frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}. \tag{15}$$

B. Under independent standard half-Cauchy prior on λ_i*s, for the second and third terms in Equation (15) we have:*

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathbb{E}(Z_i), \quad \textit{and} \quad \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathbb{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathbb{E}(Z_i^2),$$

where $(Z_i \mid \hat{\alpha}_i, \sigma, \tau)$ *follows a* $\mathrm{CCH}(p = 1, q = 1/2, r = 1, s = \hat{\alpha}_i^2 d_i^2/2\sigma^2, v = 1, \theta = 1/\tau^2 d_i^2)$ *distribution.*

A proof is given in Appendix A.1. Theorem 4.1 provides a computationally tractable mechanism for calculating SURE for the horseshoe shrinkage regression in terms of the moments of CCH random variables. Gordy (1998) provides a simple formula for all integer moments of CCH random variables. Specifically, he shows if $X \sim \mathrm{CCH}(x; p, q, r, s, \nu, \theta)$ then

$$\mathbb{E}(X^k) = \frac{(p)_k}{(p+q)_k}\frac{H(p+k, q, r, s, \nu, \theta)}{H(p, q, r, s, \nu, \theta)}, \tag{16}$$

for integers $k \geq 1$. Moreover, as demonstrated by Gordy (1998), these moments can be numerically evaluated quite easily over a range of parameter values and calculations remain very stable. A consequence of this explicit formula for SURE is that the global shrinkage parameter τ can now be chosen to minimize SURE by performing a one-dimensional numerical optimization. Another consequence is that an application of Theorem 3 of Carvalho et al. (2010) shows

$$\lim_{|\hat{\alpha}_i|\to\infty}\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = \lim_{|\hat{\alpha}_i|\to\infty}\frac{\partial \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i} = 0,$$

with high probability, where $m(\hat{\alpha}_i)$ is the marginal under the horseshoe prior. Recall that the posterior mean $\tilde{\alpha}_i$ and the OLS estimate $\hat{\alpha}_i$ are related by Tweedie's formula as $\tilde{\alpha}_i = \hat{\alpha}_i + \sigma^2 d_i^{-2}\partial \log m(\hat{\alpha}_i)/\partial\hat{\alpha}_i$. Thus, $\tilde{\alpha}_i \approx \hat{\alpha}_i$, with high probability, as $|\hat{\alpha}_i| \to \infty$, for any fixed d_i and σ for the horseshoe regression. Since $\hat{\alpha}_i$ is unbiased for α_i, the resultant horseshoe posterior mean is also seen to be unbiased when $|\hat{\alpha}_i|$ is large. Compare with the resultant $\tilde{\alpha}_i$ for global shrinkage regression of Equation (6), which is monotone decreasing in d_i, and therefore can be highly biased if a true large $|\alpha_i|$ corresponds to a small d_i. Perhaps more importantly, we can use the expression from Theorem 4.1 to estimate the prediction risk of the horseshoe regression for the signal and the noise terms. First we treat the case when $|\hat{\alpha}_i|$ is large. We have the following result.

Theorem 4.2 *Define* $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$, $\theta_i = (\tau^2 d_i^2)^{-1}$. *For any* $s_i \geq 1, \theta_i \geq 1$, *we have for the horseshoe regression of Equations (10–12) that*

$$\left\{1 - \theta_i(\tilde{C}_1 + \tilde{C}_2)\frac{(1+s_i)}{s_i^2} - \theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2\frac{(1+s_i)^2}{s_i^3}\right\} \leq \frac{SURE_i}{2\sigma^2} \leq \left\{1 + 2\theta_i(1+s_i)\left(\frac{C_1}{s_i^2} + \frac{C_2}{s_i^{3/2}}\right)\right\},$$

where $C_1 = \{1 - 5/(2e)\}^{-1/2} \approx 3.53, C_2 = 16/15, \tilde{C}_1 = (1 - 2/e)^{-1/2} \approx 1.95, \tilde{C}_2 = 4/3$, *are constants.*

A proof is given in Appendix A.2. Our result is non-asymptotic, i.e., it is valid for any $s_i \geq 1$. However, an easy consequence is that $SURE_i \to 2\sigma^2$, almost surely, as $s_i \to \infty$, provided $\tau^2 \leq d_i^{-2}$. An intuitive explanation of this result is that component-specific shrinkage is feasible in the horseshoe regression model due to the heavy-tailed λ_i terms, which prevents the signal terms from getting shrunk too much and consequently, making a large contribution to SURE due to a large bias. With just a global parameter τ, this component-specific shrinkage is not possible. A comparison of $SURE_i$ resulting from Theorem 4.2 with that from Equation (9) demonstrates using global-local horseshoe shrinkage, we can rectify a major shortcoming of global shrinkage regression, in that the terms with large s_i do not make a large contribution to the prediction risk. Moreover, the main consequence of Theorem 4.2, that is $SURE_i \to 2\sigma^2$, almost surely, as $s_i \to \infty$, holds for a larger class of "global-local" priors, of which the horseshoe is a special case.

Theorem 4.3 *Consider the hierarchy of Equations (10–11) and suppose the prior on* λ_i *in Equation (12) satisfies* $p(\lambda_i^2) \sim (\lambda_i^2)^{a-1} L(\lambda_i^2)$ *as* $\lambda_i^2 \to \infty$, *where* $f(x) \sim g(x)$ *means* $\lim_{x\to\infty} f(x)/g(x) = 1$. *Assume* $a \leq 0$ *and* $L(\cdot)$ *is a slowly-varying function, defined as* $\lim_{|x|\to\infty} L(tx)/L(x) = 1$ *for all* $t \in (0, \infty)$. *Then we have* $SURE_i \to 2\sigma^2$, *almost surely, as* $s_i \to \infty$.

A proof is given in Appendix A.3. Densities that satisfy $p(\lambda_i^2) \sim (\lambda_i^2)^{a-1} L(\lambda_i^2)$ as $\lambda_i^2 \to \infty$ are sometimes called regularly varying or heavy-tailed. Clearly, the horseshoe prior is a special case, since for the standard half-Cauchy we have $p(\lambda_i) \propto 1/(1 + \lambda_i^2)$, which yields by a change of variables $p(\lambda_i^2) = (\lambda_i^2)^{-3/2}\{\lambda_i^2/(1 + \lambda_i^2)\}$, which is of the form $(\lambda_i^2)^{a-1} L(\lambda_i^2)$ with $a = -1/2$ since $L(\lambda_i^2) = \lambda_i^2/(1 + \lambda_i^2)$ is seen to be slowly-varying. Other priors that fall in this framework are the horseshoe+ prior of Bhadra et al. (2017), for which $p(\lambda_i) \propto \log(\lambda_i)/(\lambda_i^2 - 1) = \lambda_i^{-2} L(\lambda_i^2)$ with $L(\lambda_i^2) = \log(\lambda_i)\lambda_i^2/(\lambda_i^2 - 1)$. Ghosh et al. (2016) show that the generalized double Pareto prior (Armagan et al., 2013) and the three parameter beta prior (Armagan et al., 2011) also fall in this framework. Thus, Theorem 4.3 generalizes the main consequence of Theorem 4.2 to a broader class of priors in the asymptotic sense as $s_i \to \infty$.

Next, for the case when $|\hat{\alpha}_i|$ is small, we have the following result for estimating the prediction risk of the horseshoe regression.

Theorem 4.4 *Define* $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$ *and* $\theta_i = (\tau^2 d_i^2)^{-1}$. *Then the following statements are true for the horseshoe regression.*

A. $SURE_i$ *is an increasing function of* s_i *in the interval* $s_i \in [0, 1]$ *for any fixed* τ.

B. *When* $s_i = 0$, *we have that* $SURE_i$ *is a monotone increasing function of* τ, *and is bounded in the interval* $(0, 2\sigma^2/3]$ *when* $\tau^2 d_i^2 \in (0, 1]$.

C. *When* $s_i = 1$, *we have that* $SURE_i$ *is bounded in the interval* $(0, 1.93\sigma^2]$ *when* $\tau^2 d_i^2 \in (0, 1]$.

A proof is given in Appendix A.4. This theorem establishes that: (i) the terms with smaller s_i in the interval $[0, 1]$ contribute less to SURE, with the minimum achieved at $s_i = 0$ (these terms can be thought of as the noise terms) and (ii) if τ is chosen to be sufficiently small, the terms for which $s_i = 0$, has an upper bound on SURE at $2\sigma^2/3$. Note that the OLS estimator has risk $2\sigma^2$ for these terms. At $s_i = 0$, the PCR risk is either 0 or $2\sigma^2$, depending on whether the term is or is not included. A commonly used technique for shrinkage regressions is to choose the global τ to minimize a data-dependent estimate of the risk, such as C_L or SURE (Mallows, 1973). The ridge regression SURE at $s_i = 0$ is an increasing function of τ and thus, it might make sense to choose a small τ if all s_i terms were small. However, in the presence of some s_i terms that are large, ridge regression cannot choose a very small τ, since the large s_i terms will then be heavily shrunk and contribute too much to SURE. This is not the case with global-local shrinkage regression methods such as the horseshoe, which can still choose a small τ to mitigate the contribution from the noise terms and rely on the heavy-tailed λ_i terms to ensure large signals are not shrunk too much. Consequently, the ridge regression risk estimate is usually larger than the global-local regression risk estimate even for very small s_i terms, when some terms with large s_i are present along with mostly noise terms. At this point, the results concern the risk estimate (i.e., SURE) rather than risk itself, the discussion of which is deferred until Section 5.

To summarize the theoretical findings, Theorem 4.2 together with Theorem 4.4 establishes that the horseshoe regression is effective in handling both very large and very small values of $\hat{\alpha}_i^2 d_i^2$. Specifically, Theorem 4.4 asserts that a small enough τ shrinks the noise terms towards zero, minimizing their contribution to SURE. Whereas, according to Theorem 4.2, the heavy tails of the Cauchy priors for the λ_i terms ensure the large signals are not shrunk too much and ensures a SURE of $2\sigma^2$ for these terms, which is an improvement over purely global methods of shrinkage.

5. Prediction risk for the global and horseshoe regressions

In this section we compare the theoretical prediction risks of global and global-local horseshoe shrinkage regressions. While SURE is a data-dependent estimate of the theoretical risk, these two quantities are equal in expectation for all n. We use a concentration argument to derive conditions under which the horseshoe regression will outperform global shrinkage regression, e.g., ridge regression, in terms of predictive risk. While the analysis seems difficult for an arbitrary design matrix X, we are able to treat the case of ridge regression for orthogonal design, i.e., $X^T X = I$. Clearly, if the SVD of X is written as $X = UDV^T$, then we have $D = I$ and for ridge regression $\lambda_i = 1$ for all i. Thus, for orthogonal design, Equations (4) and (5) become

$$\begin{aligned} (\hat{\alpha}_i \mid \alpha_i, \sigma^2) &\overset{ind}{\sim} \mathcal{N}(\alpha_i, \sigma^2), \\ (\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) &\overset{ind}{\sim} \mathcal{N}(0, \sigma^2\tau^2), \end{aligned}$$

where τ is the global shrinkage parameter. Since the fit in this model is linear in $\hat{\alpha}_i$, SURE is equivalent to Mallows' C_L. Equation (14) of Mallows (1973) shows that if τ is chosen to

minimize C_L, then the optimal ridge estimate is given in closed form by

$$\alpha_i^\star = \left(1 - \frac{n\sigma^2}{\sum_{i=1}^n \hat{\alpha}_i^2}\right)\hat{\alpha}_i.$$

Alternatively, the solution can be directly obtained from Equation (9) by taking $d_i = \lambda_i = 1$ for all i and by setting $\tau^\star = \text{argmin}_\tau \sum_{i=1}^n SURE_i$. It is perhaps interesting to note that this "optimal" ridge estimate, where the tuning parameter is allowed to depend on the data, is no longer linear in $\hat{\alpha}$. In fact, the optimal solution $\alpha^\star$ can be seen to be closely related to the James–Stein estimate of α and its risk can therefore be quantified using the risk bounds on the James–Stein estimate. As expected due to the global nature of ridge regression, the relative shrinkage $\alpha_i^\star/\hat{\alpha}_i$ of the optimal solution only depends on $|\hat{\alpha}|^2 = \sum_{i=1}^n \hat{\alpha}_i^2$ but not on the individual components of $\hat{\alpha}$. Theorem 1 of Casella and Hwang (1982) shows that

$$1 - \frac{n-2}{n+|\alpha|^2} \leq \frac{R(\alpha,\alpha^\star)}{R(\alpha,\hat{\alpha})} \leq 1 - \frac{(n-2)^2}{n}\left(\frac{1}{n-2+|\alpha|^2}\right).$$

Consequently, if $|\alpha|^2/n \to c$ as $n \to \infty$ then the James–Stein estimate satisfies

$$\lim_{n\to\infty} \frac{R(\alpha,\alpha^\star)}{R(\alpha,\hat{\alpha})} = \frac{c}{c+1}.$$

Thus, $\alpha^\star$ offers large benefits over the least squares estimate $\hat{\alpha}$ for small c but it is practically equivalent to the least squares estimate for large c. The prediction risk of the least squares estimate for $p > n$ is simply $2n\sigma^2$, or an average component-specific risk of $2\sigma^2$. We first show that when true $\alpha_i = 0$, the component-specific risk bound of the horseshoe shrinkage regression with a fixed $\tau = 1$ (i.e., the case of purely local shrinkage) is less than $2\sigma^2$. We have the following result.

Theorem 5.1 *(Prediction risk for the purely local horseshoe regression). Let $D = I$ and let the global shrinkage parameter in the horseshoe regression be $\tau^2 = 1$. When true $\alpha_i = 0$, an upper bound of the component-wise risk of the purely local horseshoe regression is $1.75\sigma^2 < 2\sigma^2$.*

A proof can be found in Appendix A.5. The proof uses the fact that the actual risk can be obtained by computing the expectation of SURE. We split the domains of integration into three distinct regions and use the bounds on SURE from Theorems 4.2 and 4.4, as appropriate.

When true α_i is large enough, a consequence of Theorem 4.2 is that the component-specific risk for global-local shrinkage regression is $2\sigma^2$. This is because SURE in this case is almost surely equal to $2\sigma^2$ and $\hat{\alpha}_i$ is concentrated around true α_i. Therefore, it is established that if only a few components of true α are large and the rest are zero in such a way that $|\alpha|^2/n$ is large, then the horseshoe regression with fixed $\tau = 1$ outperforms ridge regression in terms of predictive risk. The benefit arises from a lower risk for the $\alpha_i = 0$ terms. On the other hand, if all components of true α are zero or all are large, the horseshoe regression need not outperform ridge regression.

Although Theorem 5.1 shows the horseshoe regression with a fixed $\tau = 1$ outperforms the optimal ridge regression in predictive risk when $\alpha = 0$, a useful corollary is that the optimal horseshoe regression still outperforms the optimal ridge regression, where the optimal global tuning parameters for both methods are chosen by minimizing their respective SURE.

Corollary 5.1 *(Prediction risk for the optimal horseshoe regression). Let* $SURE_{HS}(\tau = 1)$ *and* $SURE_{HS}(\tau = \tau^*_{HS})$ *denote the SURE for the horseshoe regression with fixed* $\tau = 1$ *and* $\tau = \tau^*_{HS} = \text{argmin}_\tau\,(SURE_{HS}(\tau))$. *Then, for any* α, $R(\alpha, \hat{\alpha}^{HS}(\tau = \tau^*_{HS})) \leq R(\alpha, \hat{\alpha}^{HS}(\tau = 1))$.

Proof

$$R(\alpha, \hat{\alpha}^{HS}(\tau = \tau^*_{HS})) = \mathbb{E}_{\hat{\alpha}|\alpha}(SURE_{HS}(\tau = \tau^*_{HS})) \leq \mathbb{E}_{\hat{\alpha}|\alpha}(SURE_{HS}(\tau = 1)) = R(\alpha, \hat{\alpha}^{HS}(\tau = 1)).$$

■

Clearly, τ^*_{HS} is a function of the data and this complicates exact prediction risk calculations for the optimal horseshoe regression as an expectation of SURE as in Theorem 5.1. It is not clear if an explicit minimizer of SURE analogous to Equation (14) of Mallows (1973) for ridge regression can be obtained for the horseshoe regression. Nevertheless, Corollary 5.1 shows the risk for the horseshoe regression can only decrease further if one sets $\tau = \tau^*_{HS}$, similar to the risk result of Stein (1956). This holds because the expectations of SURE are computed with respect to the distribution of $\hat{\alpha}$, which is independent of τ given true α.

6. Risk comparisons with other non-convex regressions

In this section we compare the risk of the proposed horseshoe regression with other approaches that are not shrinkage methods. Specifically, we consider the minimax concave penalty (MCP) of Zhang (2010). Again, for simplicity assume that the design matrix X is orthogonal. As pointed out by Zhang (2010), in this case the solution to the MCP estimator is available in closed form and reduces to the firm shrinkage estimator of Gao and Bruce (1997), which is given by

$$\delta_{\lambda,\gamma}(\hat{\alpha}_i) = \begin{cases} 0, & \text{if } |\hat{\alpha}_i| \leq \lambda, \\ \text{sign}(\hat{\alpha}_i)\frac{\gamma(|\hat{\alpha}_i| - \lambda)}{\gamma - 1}, & \text{if } \lambda \leq |\hat{\alpha}_i| \leq \gamma\lambda, \\ \hat{\alpha}_i, & \text{if } |\hat{\alpha}_i| > \gamma\lambda, \end{cases}$$

for $\lambda > 0$ and $\gamma > 1$. For a fixed λ, soft and hard thresholding estimators are obtained as $\gamma \to \infty$ and $\gamma \to 1+$ respectively. An explicit expression for the risk of this estimator is given in Theorem 1 of Gao and Bruce (1997), from which it can be seen easily that $R(\delta_{\lambda,\gamma}) > \lambda^2\{1/2 - \Phi(-2\lambda)\}$ when $\alpha_i = \lambda$ for any fixed $\lambda > 0$ and $\gamma > 1$, where $\Phi(\cdot)$ is the standard normal distribution function. Thus, for MCP to work well, a small value for λ is essential. However, λ is the threshold below which the estimates are shrunk to zero and a large λ is favored in a "dense" situation, where there are many true parameters several standard deviations away from zero. While this behavior is not necessarily a problem for the MCP, since it is designed with a sparse situation in mind, it is perhaps desirable to avoid a large risk at a given λ. The horseshoe regression achieves exactly that, since its component-specific risk in a dense case is $2\sigma^2$ by Theorem 4.2. In Supplementary Section S.1 we verify that the MCP performs worse than both global and global-local shrinkage methods in a dense situation under a variety of designs X.

Table 1: The true orthogonalized regression coefficients α_{0i}, their ordinary least squared estimates $\hat{\alpha}_i$, and singular values d_i of the design matrix, for $n = 100$ and $p = 500$.

i	α_{0i}	$\hat{\alpha}_i$	d_i	$\hat{\alpha}_i d_i$
1	0.10	0.10	635.10	62.13
2	-0.44	-0.32	3.16	-1.00
...	...	...	...	...
5	-0.13	0.30	3.05	0.91
6	10.07	10.22	3.02	30.88
...	...	...	...	...
29	0.46	0.60	2.53	1.53
30	10.47	11.07	2.51	27.76
...	...	...	...	...
56	0.35	0.57	2.07	1.18
57	10.23	10.66	2.07	22.05
...	...	...	...	...
66	-0.00	-0.35	1.90	-0.66
67	11.14	11.52	1.88	21.70
...	...	...	...	...
95	-0.82	-0.56	1.42	-0.79
96	9.60	10.21	1.40	14.26
...	...	...	...	...
100	0.61	0.91	1.27	1.15

7. Numerical examples

We simulate data where $n = 100$, and consider the cases $p = 100, 200, 300, 400, 500$. Let B be a $p \times k$ factor loading matrix, with all entries equal to 1. Let F_i be $k \times 1$ matrix of factor values, with all entries drawn independently from $\mathcal{N}(0, 1)$. The ith row of the $n \times p$ design matrix X is generated by a factor model, with number of factors $k = 8$, as follows:

$$X_i = BF_i + \xi_i, \quad \xi_i \sim \mathcal{N}(0, 0.1), \quad \text{for} \quad i = 1, \ldots, n.$$

Thus, the columns of X are correlated. Let $X = UDW^T$ denote the singular value decomposition of X. The observations y are generated from Equation (3) with $\sigma^2 = 1$, where for the true orthogonalized regression coefficients α_0, the 6, 30, 57, 67, and 96th components are randomly selected as signals, and the remaining 95 components are noise terms. Coefficients of the signals are generated by a $\mathcal{N}(10, 0.5)$ distribution, and coefficients of the noise terms are generated by a $\mathcal{N}(0, 0.5)$ distribution. For the case $n = 100$ and $p = 500$, some of the true orthogonalized regression coefficients α_0, their ordinary least squared estimates $\hat{\alpha}$, and the corresponding singular values d of the design matrix, are shown in Table 1.

Table 2 lists the SURE for prediction and actual out of sample sum of squared prediction error (SSE) for the ridge, lasso, PCR and horseshoe regressions. Out of sample prediction error of the adaptive lasso is also included in the comparisons, although we are unaware of a formula for computing the SURE for the adaptive lasso. SURE for ridge and PCR can be

Table 2: SURE and average out of sample prediction SSE (standard deviation of SSE) on one training set and 200 testing sets for the competing methods for $n = 100$, for ridge regression (RR), the lasso regression (LASSO), the adaptive lasso (A_LASSO), principal components regression (PCR) and the horseshoe regression (HS). The lowest SURE in each row is in italics and the lowest average prediction SSE is in bold. A formula for SURE is unavailable for the adaptive lasso.

	RR		LASSO		A_LASSO	PCR		HS	
p	SURE	SSE	SURE	SSE	SSE	SURE	SSE	SURE	SSE
100	159.02	168.24	125.37	128.98	127.22	162.23	179.81	*120.59*	**126.33**
		(23.87)		(18.80)	(18.10)		(25.51)		(18.77)
200	187.38	174.92	140.99	132.46	151.89	213.90	191.33	*139.32*	**126.99**
		(21.13)		(18.38)	(20.47)		(22.62)		(17.29)
300	192.78	191.91	*147.83*	145.04	153.64	260.65	253.00	151.24	**136.67**
		(22.95)		(19.89)	(21.19)		(26.58)		(18.73)
400	195.02	182.55	148.56	165.63	178.98	346.19	292.02	*147.69*	**143.91**
		(22.70)		(21.55)	(20.12)		(28.98)		(18.41)
500	196.11	188.78	159.95	**159.56**	186.23	386.50	366.88	*144.97*	160.11
		(22.33)		(19.94)	(23.50)		(39.38)		(20.29)

computed by an application of Equation (9) and SURE for the horseshoe regression is given by Theorem 4.1. SURE for the lasso is calculated using the result given by Tibshirani and Taylor (2012). In each case, the model is trained on 100 samples. We report the SSE on 100 testing samples, averaged over 200 testing data sets, and their standard deviations. For ridge, lasso, PCR and horseshoe regression, the global shrinkage parameters were chosen to minimize SURE for prediction. In adaptive lasso, the shrinkage parameters were chosen by cross validation due to SURE being unavailable. It can be seen that SURE in most cases are within one standard deviation of the actual out of sample prediction SSE, suggesting SURE is an accurate method for evaluating actual out of sample prediction performance. When $p = 100, 200, 300, 400$, horseshoe regression has the lowest prediction SSE. When $p = 500$, SSE of the lasso and horseshoe regression are close, and the lasso performs marginally better. The horseshoe regression also has the lowest SURE in all but one cases. Generally, SURE increases with p for all methods. The SURE for ridge regression approaches the OLS risk, which is $2n\sigma^2 = 200$ in these situations. SURE for PCR is larger than the OLS risk and PCR happens to be the poorest performer in most settings. Performance of the adaptive lasso also degrades compared to the lasso and the horseshoe, which remain the two best performers. Finally, the horseshoe regression outperforms the lasso in four out of the five settings we considered.

Figure 1 shows contribution to SURE by each component for $n = 100$ and $p = 500$, for ridge, PCR, lasso and horseshoe regressions. The components are ordered left to right on the x-axis by decreasing magnitude of d_i, and SURE for prediction on each component are shown on the y-axis. Note from Table 1 that the $6, 30, 57, 67$ and 96th components are the signals, meaning these terms correspond to a large α_0. The PCR risk on the 96th component is 203.22, which is out of range for the y-axis in the plot. For this data set, PCR selects

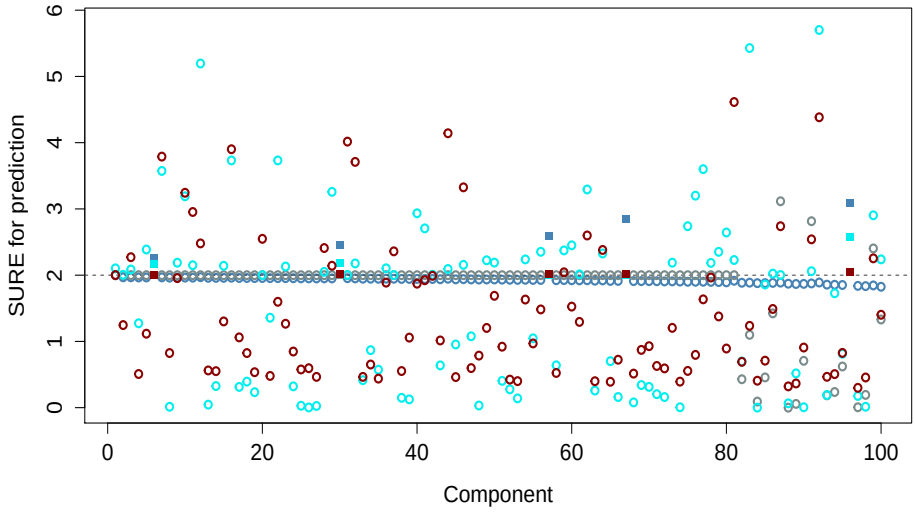

Figure 1: Component-wise SURE for ridge (blue), PCR (gray), lasso (cyan), and horseshoe regression (red), for $n = 100$ and $p = 500$. Signal components are shown in solid squares and noise components shown in blank circles. Dashed horizontal line is at $2\sigma^2 = 2$.

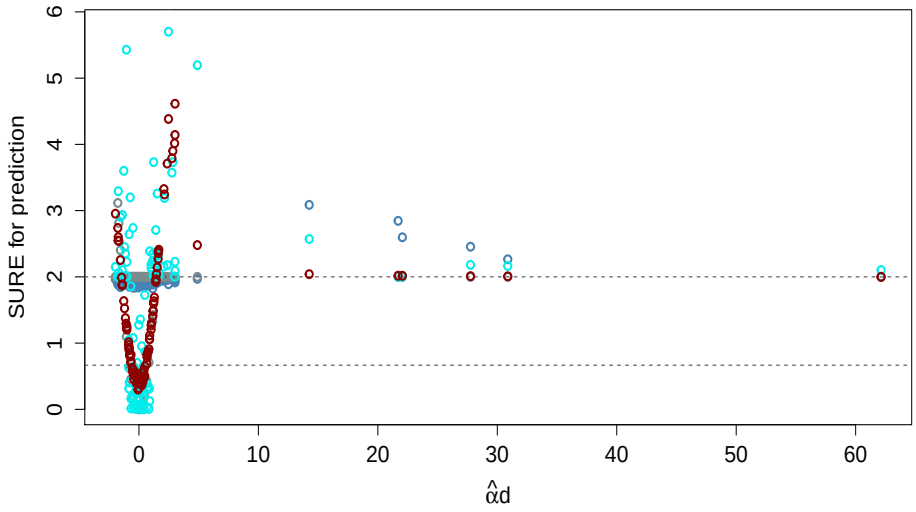

Figure 2: SURE for ridge (blue), PCR (gray), lasso (cyan) and horseshoe regression (red), versus $\hat{\alpha}d$, where $\hat{\alpha}$ is the OLS estimate of the orthogonalized regression coefficient, and d is the singular value, for $n = 100$ and $p = 500$. Dashed horizontal lines are at $2\sigma^2 = 2$ and $2\sigma^2/3 = 0.67$.

81 components, and therefore SURE for the first 81 components equal to $2\sigma^2 = 2$ and the SURE is equal to $\hat{\alpha}_i^2 d_i^2$ for $i = 82, \ldots, 100$. Component-wise SURE for ridge regression are large on the signal components, and is decreasing as the singular values d decrease on the other components. But due to the large global shrinkage parameter τ ridge must select in presence of both large signals and noise terms, the magnitude of improvement over the OLS risk $2\sigma^2$ is small for the noise terms. On the other hand, the horseshoe estimator does not shrink the components with large $\hat{\alpha}_i d_i$ heavily and therefore the horseshoe SURE on the signal components are almost equal to $2\sigma^2$ (according to Theorem 4.2). SURE for the horseshoe is also much smaller than $2\sigma^2$ on many of the noise components. Lasso also appears to be quite effective for the noise terms, but its performance for the signal components is generally not as effective as the horseshoe.

Figure 2 takes a fresh look at the same results and shows component-wise SURE plotted against $\hat{\alpha}_i d_i$. The signal components as well as the first component in Table 1 have $\hat{\alpha}_i d_i > 10$. Horseshoe SURE converges to $2\sigma^2$ for large $\hat{\alpha}_i d_i$, as expected from Theorem 4.2. For these components, the SURE for both ridge and lasso are larger than $2\sigma^2$, due to the bias introduced in estimating large signals by these methods (see also Theorem 1 of Carvalho et al., 2010). When ${\hat{\alpha}_i}^2 d_i^2 \approx 0$, risks for lasso and horseshoe are comparable, with lasso being slightly better. This is because an estimate can be exactly zero for the lasso, but not for the horseshoe, which is a shrinkage method (as opposed to a selection method). Nevertheless, the upper bound on SURE for the horseshoe regression at $2\sigma^2/3$ when ${\hat{\alpha}_i}^2 d_i^2 \approx 0$ and provided τ is chosen to be small enough so that $\tau^2 \leq d_i^{-2}$, as established by Theorem 4.4, can be verified from Figure 2.

Additional simulation results are presented in Supplementary Section S.1, where we (i) treat a higher dimensional case ($p = 1000$), (i) perform comparisons with non-convex MCP (Zhang, 2010) and SCAD (Fan and Li, 2001) regressions, (iii) explore different choices of X and (iv) explore the effect of the choice of α. The main finding is that the horseshoe regression is often the best performer when α has a sparse-robust structure as in Table 1, that is most elements are very small while a few are large so that $|\alpha|^2$ is large. This is consistent with the theoretical results of Sections 5 and 6.

8. Assessing out of sample prediction in a pharmacogenomics data set

We compare the out of sample prediction error of the horseshoe regression with ridge regression, PCR, the lasso, the adaptive lasso, MCP and SCAD on a pharmacogenomics data set. The data were originally described by Szakács et al. (2004), in which the authors studied 60 cancer cell lines in the publicly available NCI-60 database (https://dtp.cancer.gov/discovery_development/nci-60/). The goal here is to predict the expression of the human ABC transporter genes (responses) using some compounds or drugs (predictors) at which 50% inhibition of cellular growth for the cell lines are induced. The NCI-60 database includes the concentration level of 1429 such compounds, out of which we use 853, which did not have any missing values, as predictors. We investigate the expression levels of transporter genes A1 to A12 (except for A11, which we omit due to missing values), and B1. Thus, in our study X is a $n \times p$ matrix of predictors with $n = 60, p = 853$ and Y is a n-dimensional response vector for each of the 12 candidate transporter genes under consideration.

To test the performance of the methods, we split each data set into training and testing sets, with 75% (45 out of 60) of the observations in the training sets. We standardize each response by subtracting the mean and dividing by the standard deviation. We fit the model on the training data, and then calculate mean squared prediction error (prediction MSE) on the testing data. This is repeated for 20 random splits of the data into training and testing sets. The tuning parameters in ridge regression, the lasso, the adaptive lasso, SCAD and MCP are chosen by five-fold cross validation on the training data. Similarly, the number of components in PCR and the global shrinkage parameter τ for horseshoe regression are chosen by cross validation as well. It is possible to use SURE to select the tuning parameters or the number of components, but one needs an estimate of the standard deviation of the errors in high-dimensional regressions. This is a problem of recent interest, as the OLS estimate of σ^2 is not well-defined in the $p > n$ case. Unfortunately, some of the existing methods we tried, such as the method of moments estimator of Dicker (2014), often resulted in unreasonable estimates for σ^2, such as negative numbers. Thus, we stick to cross validation here, as it is not necessary to estimate the residual standard deviation in that case.

The average prediction MSE over 20 random training-testing splits for the competing methods is reported in Table 3. Average prediction MSE for responses A1, A8 and A10 are around or larger than 1 for all of the methods. Since the responses are standardized before analysis, we might conclude that none of the methods performed well for these cases. Among the remaining nine cases, the horseshoe regression substantially outperforms the other methods for A3, A4, A9, A12 and B1. It is comparable to PCR for A5 and A7, and is comparable to the adaptive lasso for A6, which are the best performers in the respective cases. Overall, the horseshoe regression performed the best in 5 among the total 12 cases we considered.

9. Concluding remarks

We outlined some situations where the horseshoe regression is expected to perform better compared to some other commonly used "global" shrinkage or selection alternatives for high-dimensional regression. Specifically, we demonstrated that the global term helps in mitigating the prediction risk arising from the noise terms, and an appropriate choice for the tails of the local terms is crucial for controlling the risk due to the signal terms. For this article we have used the horseshoe prior as our choice for the global-local prior. However, in recent years, several other priors have been developed that fall in this class. This includes the horseshoe+ (Bhadra et al., 2017, 2016), the three-parameter beta (Armagan et al., 2011), the normal-exponential-gamma (Griffin and Brown, 2010), the generalized double Pareto (Armagan et al., 2013), the generalized shrinkage prior (Denison and George, 2012) and the Dirichlet–Laplace prior (Bhattacharya et al., 2015). Empirical Bayes approaches have also appeared (Martin and Walker, 2014) and the spike and slab priors have made a resurgence due to recently developed efficient computational approaches (Ročková and George, 2014; Ročková and George, 2016). Especially in the light of Theorem 4.3, we expect the results developed in this article for the horseshoe to foreshadow similar results when many of these alternatives are deployed. A particular advantage of using the horseshoe prior seems to be the tractable expression for SURE, as developed in Theorem 4.1.

Table 3: Average out of sample mean squared prediction error computed on 20 random training-testing splits (number of splits out of 20 with lowest prediction MSE), for each of the 12 human ABC transporter genes (A1–A10, A12, B1) in the pharmacogenomics example. Methods under consideration are ridge regression (RR), principal components regression (PCR) , the lasso, the adaptive lasso (A_LASSO), the minimax concave penalty (MCP), the smoothly clipped absolute deviation (SCAD) penalty, and the horseshoe regression (HS). Lowest prediction MSE and largest number of splits with the lowest prediction MSE for each response in bold.

Response	RR	PCR	LASSO	A_LASSO	MCP	SCAD	HS
A1	1.12	1.10	**1.00**	**1.00**	1.01	1.06	1.30
	(2)	(5)	**(7)**	(2)	(1)	(1)	(2)
A2	1.00	1.04	0.95	0.93	**0.92**	0.99	1.15
	(3)	(1)	**(7)**	(5)	(1)	(0)	(3)
A3	0.77	0.91	1.11	0.90	0.92	1.06	**0.65**
	(1)	(0)	(0)	(0)	(1)	(0)	**(18)**
A4	0.92	0.95	0.97	0.96	0.93	0.99	**0.79**
	(2)	(0)	(2)	(2)	(2)	(0)	**(12)**
A5	0.82	**0.77**	1.06	0.81	0.83	0.94	0.79
	(1)	**(6)**	(4)	(1)	(2)	(0)	**(6)**
A6	0.93	0.92	0.98	**0.86**	0.87	0.90	0.95
	(4)	(0)	(3)	(5)	(0)	(2)	**(6)**
A7	0.92	**0.83**	0.92	0.93	0.99	0.93	0.85
	(0)	**(8)**	(1)	(4)	(0)	(0)	(7)
A8	1.08	1.05	1.14	**1.01**	**1.01**	1.15	1.34
	(6)	(4)	**(6)**	(4)	(0)	(0)	(0)
A9	0.57	0.64	0.81	0.67	0.77	0.68	**0.55**
	(4)	(0)	(0)	(6)	(0)	(1)	**(9)**
A10	1.18	1.04	**1.00**	1.01	**1.00**	1.06	1.33
	(0)	**(7)**	(4)	(3)	(2)	(0)	(4)
A12	1.01	1.12	1.09	1.01	1.02	1.05	**0.80**
	(0)	(0)	(2)	(2)	(1)	(0)	**(15)**
B1	0.53	0.59	0.70	0.63	0.91	0.70	**0.46**
	(1)	(0)	(3)	(2)	(1)	(3)	**(10)**

Whether this advantage translates to some of the other global-local priors mentioned above is an open question. Following the approach of Stein (1981), our risk results are developed in a non-asymptotic setting (finite n, finite $p > n$). In the normal means model, finite sample risk properties in estimation under heavy-tailed priors have been considered by Polson and Scott (2012b). However, their work does not consider (a) predictive risk or (b) a linear regression model. Global-local priors such as the horseshoe and horseshoe+ are known to be minimax in estimation in the Gaussian sequence model (van der Pas et al., 2014, 2016). For linear regression, frequentist minimax risk results are discussed by Raskutti et al. (2011); and Castillo et al. (2015) have shown that spike and slab priors achieve minimax prediction

risk in regression. Whether the prediction risk for the horseshoe regression is optimal in an asymptotic sense is an important question to investigate and recent asymptotic prediction risk results for ridge regression (Dobriban and Wager, 2017) should prove helpful for comparing with global shrinkage methods. Another possible direction for future investigation might be to explore the implications of our findings on the predictive density in terms of an appropriate metric, say the Kullback-Leibler loss, following the results of George et al. (2006).

Acknowledgements

The authors are grateful for constructive suggestions by the reviewers and the action editor. Bhadra and Polson are supported by Grant No. DMS-1613063 by the US National Science Foundation.

Appendix A. Proofs

A.1. Proof of Theorem 4.1

Part A follows from Equation (7) with standard algebraic manipulations. To prove part B, define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$. Then, from Equation (13)

$$\begin{aligned} m(\hat{\alpha}) &= (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n}\int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2) d_i z_i^{1/2}\left(\frac{z_i\tau^2 d_i^2}{1-z_i+z_i\tau^2 d_i^2}\right)\frac{1}{\tau d_i}(1-z_i)^{-1/2} z_i^{-3/2} dz_i \\ &= (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n}\int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2)(1-z_i)^{-1/2}\left\{\frac{1}{\tau^2 d_i^2}+\left(1-\frac{1}{\tau^2 d_i^2}\right)z_i\right\}^{-1} dz_i. \end{aligned}$$

From the definition of the compound confluent hypergeometric (CCH) density in Gordy (1998), the result of the integral is proportional to the normalizing constant of the CCH density and we have from Proposition 4.1 that,

$$m(\hat{\alpha}) \propto (2\pi\sigma^2)^{-n/2}\prod_{i=1}^{n} H\left(1,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right).$$

In addition, the random variable $(Z_i \mid \hat{\alpha}_i,\sigma,\tau)$ follows a $\text{CCH}(1,1/2,1,\hat{\alpha}_i^2 d_i^2/2\sigma^2,1,1/\tau^2 d_i^2)$ distribution. Lemma 3 of Gordy (1998) gives,

$$\frac{d}{ds}H(p,q,r,s,\nu,\theta) = -\frac{p}{p+q}H(p+1,q,r,s,\nu,\theta).$$

This yields after some algebra that,

$$\begin{aligned} \frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} &= -\frac{2}{3}\frac{H\left(2,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)}{H\left(1,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)}\frac{\hat{\alpha}_i d_i^2}{\sigma^2}, \\ \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} &= \frac{-\frac{2}{3}H\left(2,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)\frac{d_i^2}{\sigma^2}+\frac{8}{15}H\left(3,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)\frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}}{H\left(1,\frac{1}{2},1,\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2},1,\frac{1}{\tau^2 d_i^2}\right)}. \end{aligned}$$

The correctness of the assertion

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathbb{E}(Z_i), \quad \text{and} \quad \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathbb{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathbb{E}(Z_i^2),$$

can then be verified using Equation (16), completing the proof.

A.2. Proof of Theorem 4.2

Define $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$ and $\theta_i = (\tau^2 d_i^2)^{-1}$, with$\theta_i \geq 1, s_i \geq 1$. From Theorem 4.1, the component-wise SURE is

$$\begin{aligned} SURE_i =& 2\sigma^2 - 2\sigma^2\mathbb{E}(Z_i) - \hat{\alpha}_i^2 d_i^2\{\mathbb{E}(Z_i)\}^2 + 2\hat{\alpha_i}^2 d_i^2\mathbb{E}(Z_i^2) \\ =& 2\sigma^2[1 - \mathbb{E}(Z_i) + 2s_i\mathbb{E}(Z_i^2) - s_i\{\mathbb{E}(Z_i)\}^2], \end{aligned} \tag{A.1}$$

Thus,

$$2\sigma^2[1 - \mathbb{E}(Z_i) - s_i\{\mathbb{E}(Z_i)\}^2] \leq SURE_i \leq 2\sigma^2[1 + 2s_i\mathbb{E}(Z_i^2)].$$

To find bounds on SURE, we need upper bounds on $\mathbb{E}(Z_i^2)$ and $\mathbb{E}(Z_i)$. Clearly, $\theta_i^{-1} \leq \{\theta_i + (1-\theta_i)z_i\}^{-1} \leq 1$, when $\theta_i \geq 1$. Let $a_i = \log(s_i^{5/2})/s_i$. Then $a_i \in [0, 5/(2e))$ when $s_i \geq 1$. Now,

$$\mathbb{E}(Z_i^2) = \frac{\int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i},$$

An upper bound to the numerator of $\mathbb{E}(Z_i^2)$ can be found as follows.

$$\begin{aligned} &\int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i \\ &\leq \int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\ &= \int_0^{a_i} z_i^2(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i + \int_{a_i}^1 z_i^2(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\ &\leq (1-a_i)^{-\frac{1}{2}}\int_0^{a_i} z_i^2\exp(-s_i z_i)dz_i + \exp(-a_i s_i)\int_{a_i}^1 z_i^2(1-z_i)^{-\frac{1}{2}}dz_i \\ &= (1-a_i)^{-\frac{1}{2}}\frac{2}{s_i^3}\left\{1 - \left(1 + a_i s_i + \frac{a_i^2 s_i^2}{2}\right)\exp(-a_i s_i)\right\} + \exp(-a_i s_i)\int_{a_i}^1 z_i^2(1-z_i)^{-\frac{1}{2}}dz_i \\ &\leq \{1 - 5/(2e)\}^{-\frac{1}{2}}\frac{2}{s_i^3} + \frac{1}{s_i^{5/2}}\int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}dz_i \\ &= \frac{C_1}{s_i^3} + \frac{C_2}{s_i^{5/2}}, \end{aligned}$$

where $C_1 = \{1 - 5/(2e)\}^{-\frac{1}{2}} \approx 3.53$ and $C_2 = \int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}dz_i = \Gamma(1/2)\Gamma(3)/\Gamma(3.5) = 16/15$. Similarly, a lower bound on the denominator of $\mathbb{E}(Z_i^2)$ is

$$\begin{aligned}
&\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i \\
&\geq \theta_i^{-1}\int_0^1 \exp(-s_i z_i)dz_i \\
&= \theta_i^{-1}\left\{\frac{1-\exp(-s_i)}{s_i}\right\} \geq \frac{1}{\theta_i(1+s_i)},
\end{aligned}$$

Thus, combining the upper bound on the numerator and the lower bound on the denominator

$$\mathbb{E}(Z_i^2) \leq \theta_i(1+s_i)\left(\frac{C_1}{s_i^3} + \frac{C_2}{s_i^{5/2}}\right).$$

Thus,

$$\begin{aligned}
SURE_i &\leq 2\sigma^2[1 + 2s_i\mathbb{E}(Z_i^2)] \\
&\leq 2\sigma^2\left\{1 + 2\theta_i(1+s_i)\left(\frac{C_1}{s_i^2} + \frac{C_2}{s_i^{3/2}}\right)\right\}. \qquad \text{(A.2)}
\end{aligned}$$

An upper bound to the numerator of $\mathbb{E}(Z_i)$ can be found as follows. Let $\tilde{a}_i = \log(s_i^2)/s_i$. Then, $\tilde{a}_i \in [0, 2/e)$ for $s_i \geq 1$.

$$\begin{aligned}
&\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i \\
&\leq \int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\
&= \int_0^{\tilde{a}_i} z_i(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i + \int_{\tilde{a}_i}^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\
&\leq (1-\tilde{a}_i)^{-\frac{1}{2}}\int_0^{\tilde{a}_i} z_i\exp(-s_i z_i)dz_i + \exp(-\tilde{a}_i s_i)\int_{a_i}^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i \\
&= (1-\tilde{a}_i)^{-\frac{1}{2}}\frac{1}{s_i^2}\left\{1 - (1+\tilde{a}_i s_i)\exp(-\tilde{a}_i s_i)\right\} + \exp(-\tilde{a}_i s_i)\int_{\tilde{a}_i}^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i \\
&\leq (1-2/e)^{-\frac{1}{2}}\frac{1}{s_i^2} + \frac{1}{s_i^2}\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i \\
&= \frac{\tilde{C}_1}{s_i^2} + \frac{\tilde{C}_2}{s_i^2},
\end{aligned}$$

where $\tilde{C}_1 = (1-2/e)^{-1/2} \approx 1.95$ and $\tilde{C}_2 = \int_0^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i = \Gamma(1/2)\Gamma(2)/\Gamma(2.5) = 4/3$. The lower bound on the denominator is the same as before. Thus,

$$\mathbb{E}(Z_i) \leq \frac{\theta_i(1+s_i)}{s_i^2}\left(\tilde{C}_1 + \tilde{C}_2\right).$$

Thus,

$$\begin{aligned} SURE_i &\geq 2\sigma^2[1 - \mathbb{E}(Z_i) - s_i\{\mathbb{E}(Z_i)\}^2] \\ &\geq 2\sigma^2 \left\{1 - \theta_i(\tilde{C}_1 + \tilde{C}_2)\frac{(1+s_i)}{s_i^2} - \theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2\frac{(1+s_i)^2}{s_i^3}\right\}. \end{aligned} \tag{A.3}$$

Thus, combining Equations (A.2) and (A.3) we get

$$\left\{1 - \theta_i(\tilde{C}_1 + \tilde{C}_2)\frac{(1+s_i)}{s_i^2} - \theta_i^2(\tilde{C}_1 + \tilde{C}_2)^2\frac{(1+s_i)^2}{s_i^3}\right\} \leq \frac{SURE_i}{2\sigma^2} \leq \left\{1 + 2\theta_i(1+s_i)\left(\frac{C_1}{s_i^2} + \frac{C_2}{s_i^{3/2}}\right)\right\},$$

for $s_i \geq 1, \theta_i \geq 1$.

A.3. Proof of Theorem 4.3

Our proof is similar to the proof of Theorem 1 of Polson and Scott (2011). Note from Equations (10–11) that integrating out α_i we have

$$\hat{\alpha}_i \mid \lambda_i^2, \sigma^2, \tau^2 \overset{ind}{\sim} \mathcal{N}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

Let $p(\lambda_i^2) \sim (\lambda_i^2)^{a-1}L(\lambda_i^2)$, as $\lambda_i^2 \to \infty$ where $a \leq 0$. Define $u_i = \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)$. Then, as in Theorem 1 of Polson and Scott (2011), we have

$$p(u_i) \sim u_i^{a-1}L(u_i), \text{ as } u_i \to \infty.$$

The marginal of $\hat{\alpha}_i$ is then given by

$$m(\hat{\alpha}_i) = \int \frac{1}{\sqrt{2\pi u_i}} \exp\{-\hat{\alpha}_i^2/(2u_i)\}p(u_i)du_i.$$

An application of Theorem 6.1 of Barndorff-Nielsen et al. (1982) shows that

$$m(\hat{\alpha}_i) \sim |\hat{\alpha}_i|^{2a-1}L(|\hat{\alpha}_i|) \text{ as } |\hat{\alpha}_i| \to \infty.$$

Thus, for large $|\hat{\alpha}_i|$

$$\frac{\partial \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i} = \frac{(2a-1)}{|\hat{\alpha}_i|} + \frac{\partial \log L(|\hat{\alpha}_i|)}{\partial \hat{\alpha}_i}. \tag{A.4}$$

Clearly, the first term in Equation (A.4) goes to zero as $|\hat{\alpha}_i| \to \infty$. For the second term, we need to invoke the celebrated representation theorem by Karamata. A proof can be found in Bingham et al. (1989).

Result A.1 *(Karamata's representation theorem). A function L is slowly varying if and only if there exists $B > 0$ such that for all $x \geq B$ the function can be written in the form*

$$L(x) = \exp\left(\eta(x) + \int_B^x \frac{\varepsilon(t)}{t}\, dt\right),$$

where $\eta(x)$ is a bounded measurable function of a real variable converging to a finite number as x goes to infinity $\varepsilon(x)$ is a bounded measurable function of a real variable converging to zero as x goes to infinity.

Thus, using the properties of $\eta(x)$ and $\varepsilon(x)$ from the result above

$$\frac{d\log(L(x))}{dx} = \eta'(x) + \frac{\varepsilon(x)}{x} \to 0 \quad \text{as} \quad x \to \infty.$$

Using this in Equation (A.4) shows $\partial \log m(\hat{\alpha}_i)/\partial\hat{\alpha}_i \to 0$ as $|\hat{\alpha}_i| \to \infty$. By similar calculations, $\partial^2 \log m(\hat{\alpha}_i)/\partial^2\hat{\alpha}_i \to 0$ as $|\hat{\alpha}_i| \to \infty$. From Equation (7)

$$SURE_i = \sigma^4 d_i^{-2}\left\{\frac{\partial}{\partial\hat{\alpha}_i}\log m(\hat{\alpha}_i)\right\}^2 + 2\sigma^2\left\{1 + \sigma^2 d_i^{-2}\frac{\partial^2}{\partial\hat{\alpha}_i^2}\log m(\hat{\alpha}_i)\right\}.$$

Thus, $SURE_i \to 2\sigma^2$, almost surely, as $|\hat{\alpha}_i| \to \infty$.

A.4. Proof of Theorem 4.4

The proof of Theorem 4.4 makes use of technical lemmas in Appendix A.6.

Recall from Appendix A.1 that if we define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$ then the density of Z_i is given by

$$(Z_i \mid \hat{\alpha}_i, d_i, \tau, \sigma^2) \sim \text{CCH}\left(Z_i \mid 1, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right). \tag{A.5}$$

Then SURE is given by $SURE = \sum_{i=1}^n SURE_i$ with

$$\begin{aligned} SURE_i &= 2\sigma^2[1 - \mathbb{E}(Z_i) + 2s_i\mathbb{E}(Z_i^2) - s_i\{\mathbb{E}(Z_i)\}^2] \\ &= 2\sigma^2[1 - \mathbb{E}(Z_i) + s_i\mathbb{E}(Z_i^2) + s_i\text{Var}(Z_i)], \end{aligned} \tag{A.6}$$

where $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$. Thus,

$$\begin{aligned} \frac{\partial\{SURE_i\}}{\partial s_i} &= -2\sigma^2\frac{\partial\mathbb{E}(Z_i)}{\partial s_i} + 2\sigma^2\frac{\partial}{\partial s_i}\{s_i\mathbb{E}(Z_i^2)\} + 2\sigma^2\frac{\partial}{\partial s_i}\{s_i\text{Var}(Z_i)\} \\ &:= \text{I} + \text{II} + \text{III}. \end{aligned} \tag{A.7}$$

Now, as a corollary to Lemma A.1, $(\partial/\partial s_i)\mathbb{E}(Z_i) = \{\mathbb{E}(Z_i)\}^2 - \mathbb{E}(Z_i^2) = -\text{Var}(Z_i) < 0$, giving $\text{I} > 0$. The strict inequality follows from the fact that Z_i is not almost surely a constant for any $s_i \in \mathbb{R}$ and $(\partial/\partial s_i)\mathbb{E}(Z_i)$ is continuous at $s_i = 0$. Next, consider II. Define $\theta_i = (\tau^2 d_i^2)^{-1}$ and let $0 \le s_i \le 1$. Then,

$$\begin{aligned} \frac{\partial}{\partial s_i}\{s_i\mathbb{E}(Z_i^2)\} &= \mathbb{E}(Z_i^2) + s_i\frac{\partial}{\partial s_i}\mathbb{E}(Z_i^2) \\ &= \mathbb{E}(Z_i^2) + s_i\{\mathbb{E}(Z_i)\mathbb{E}(Z_i^2) - \mathbb{E}(Z_i^3)\} \qquad \text{(by Lemma A.1)} \\ &= s_i\mathbb{E}(Z_i)\mathbb{E}(Z_i^2) + \{\mathbb{E}(Z_i^2) - s_i\mathbb{E}(Z_i^3)\}. \end{aligned}$$

Now, clearly, the first term, $s_i\mathbb{E}(Z_i)\mathbb{E}(Z_i^2) \ge 0$. We also have $Z_i^2 - s_i Z_i^3 = Z_i^2(1 - s_i Z_i) \ge 0$ a.s. when $0 \le Z_i \le 1$ a.s. and $0 \le s_i \le 1$. Thus, the second term $\mathbb{E}(Z_i^2) - s_i\mathbb{E}(Z_i^3) \ge 0$. Putting the terms together gives $\text{II} \ge 0$. Finally, consider III. Denote $\mathbb{E}(Z_i) = \mu_i$. Then,

$$\begin{aligned} \frac{\partial}{\partial s_i}\{s_i\text{Var}(Z_i)\} &= \text{Var}(Z_i) + s_i\frac{\partial}{\partial s_i}\{\text{Var}(Z_i)\} \\ &= \text{Var}(Z_i) - s_i\frac{\partial^2\mathbb{E}(Z_i)}{\partial s_i^2} \\ &= \mathbb{E}\{(Z_i - \mu_i)^2\} - s_i\mathbb{E}\{(Z_i - \mu_i)^3\} \qquad \text{(by Lemma A.2)} \\ &= \mathbb{E}[(Z_i - \mu_i)^2\{1 - s_i(Z_i - \mu_i)\}]. \end{aligned}$$

Now, $(Z_i - \mu_i)^2\{1 - s_i(Z_i - \mu_i)\} \geq 0$ a.s. when $0 \leq Z_i \leq 1$ a.s. and $0 \leq s_i \leq 1$ and thus, III ≥ 0. Using I, II and III in Equation (A.7) yields $SURE_i$ is an increasing function of s_i when $0 \leq s_i \leq 1$, completing the proof of Part A.

To prove Part B, we need to derive an upper bound on SURE when $s_i = 0$. First, consider $s_i = 0$ and $0 < \theta_i \leq 1$. we have from Equation (A.6) that $SURE_i = 2\sigma^2(1 - \mathbb{E}Z_i)$. By Lemma A.3, $(\partial/\partial\theta_i)\mathbb{E}(Z_i) > 0$ and $SURE_i$ is a monotone decreasing function of θ_i, where $\theta_i = (\tau^2 d_i^2)^{-1}$. Next consider the case where $s_i = 0$ and $\theta_i \in (1, \infty)$. Define $\tilde{Z}_i = 1 - Z_i \in (0, 1)$ when $Z_i \in (0, 1)$. Then, by Equation (A.11) and a formula on Page 9 of Gordy (1998), we have that $\tilde{Z}_i$ also follows a CCH distribution. Specifically,

$$(\tilde{Z}_i \mid \hat{\alpha}_i, d_i, \tau, \sigma^2) \sim \text{CCH}\left(\tilde{Z}_i \mid \frac{1}{2}, 1, 1, -\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \tau^2 d_i^2\right),$$

and we have $SURE_i = 2\sigma^2\mathbb{E}(\tilde{Z}_i)$. Define $\tilde{\theta}_i = \theta_i^{-1} = \tau^2 d_i^2$. Then by Lemma A.3, $(\partial/\partial\tilde{\theta}_i)\mathbb{E}(\tilde{Z}_i) = -\text{Cov}(\tilde{Z}_i, \tilde{W}_i) > 0$ on $0 < \tilde{\theta}_i < 1$. Therefore, $SURE_i$ is a monotone increasing function of $\tilde{\theta}_i$ on $0 < \tilde{\theta}_i < 1$, or equivalently a monotone decreasing function of θ_i on $\theta_i \in (1, \infty)$.

Thus, combining the two cases above, we get that SURE at $s_i = 0$ is a monotone decreasing function of θ_i for any $\theta_i \in (0, \infty)$, or equivalently, an increasing function of $\tau^2 d_i^2$. Since $0 \leq \tilde{Z}_i \leq 1$ almost surely, a natural upper bound on $SURE_i$ is $2\sigma^2$. However, it is possible to do better provided τ is chosen sufficiently small. Assume that $\tau^2 \leq d_i^{-2}$. Then, since $SURE_i$ is monotone increasing in θ_i, the upper bound on SURE is achieved when $\theta_i = (\tau^2 d_i^2)^{-1} = 1$. In this case, $\mathbb{E}(Z_i)$ has a particularly simple expression, given by

$$\begin{aligned}\mathbb{E}(Z_i) =& \frac{\int_0^1 z_i(1 - z_i)^{-\frac{1}{2}}\{\theta_i + (1 - \theta_i)z_i\}^{-1}dz_i}{\int_0^1 (1 - z_i)^{-\frac{1}{2}}\{\theta_i + (1 - \theta_i)z_i\}^{-1}dz_i} \\ =& \frac{\int_0^1 z_i(1 - z_i)^{-\frac{1}{2}}dz_i}{\int_0^1 (1 - z_i)^{-\frac{1}{2}}dz_i} = \frac{2}{3}. \end{aligned} \tag{A.8}$$

Thus, $\sup SURE_i = 2\sigma^2(1 - \mathbb{E}Z_i) = 2\sigma^2/3$, completing the proof of Part B.

To prove Part C, we first note that when $s_i = 1$ we have

$$SURE_i = 2\sigma^2[1 - \mathbb{E}(Z_i)|_{s_i=1} + 2\mathbb{E}(Z_i^2)|_{s_i=1} - \{\mathbb{E}(Z_i)|_{s_i=1}\}^2]$$

where $E(Z_i)$ and $E(Z_i^2)$ are evaluated at $s_i = 1$. Recall that when $\theta_i \geq 1$ and $z_i \in (0, 1)$ we have $\theta_i^{-1} \leq \{\theta_i + (1 - \theta_i)z_i\}^{-1} \leq 1$. Thus,

$$\begin{aligned}\mathbb{E}(Z_i^2)|_{s_i=1} =& \frac{\int_0^1 z_i^2(1 - z_i)^{-\frac{1}{2}}\{\theta_i + (1 - \theta_i)z_i\}^{-1}\exp(-z_i)dz_i}{\int_0^1 (1 - z_i)^{-\frac{1}{2}}\{\theta_i + (1 - \theta_i)z_i\}^{-1}\exp(-z_i)dz_i} \\ \leq& \frac{\int_0^1 z_i^2(1 - z_i)^{-\frac{1}{2}}\exp(-z_i)dz_i}{\theta_i^{-1}\int_0^1 (1 - z_i)^{-\frac{1}{2}}\exp(-z_i)dz_i} \approx \theta_i\frac{0.459}{1.076} = 0.43\theta_i, \end{aligned} \tag{A.9}$$

and

$$\begin{aligned}\mathbb{E}(Z_i)|_{s_i=1} =& \frac{\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\{\theta_i+(1-\theta_i)z_i\}^{-1}\exp(-z_i)dz_i}{\int_0^1(1-z_i)^{-\frac{1}{2}}\{\theta_i+(1-\theta_i)z_i\}^{-1}\exp(-z_i)dz_i},\\ &\geq \frac{\theta_i^{-1}\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-z_i)dz_i}{\int_0^1(1-z_i)^{-\frac{1}{2}}\exp(-z_i)dz_i} \approx \theta_i^{-1}\frac{0.614}{1.076} = 0.57\theta_i^{-1}. \qquad \text{(A.10)}\end{aligned}$$

Thus,

$$SURE_i \leq 2\sigma^2\left[1-\frac{0.57}{\theta_i}+0.86\theta_i-\left(\frac{0.57}{\theta_i}\right)^2\right].$$

When $\theta_i = 1$, it can be seen that $SURE_i \leq 1.93\sigma^2$.

A.5. Proof of Theorem 5.1

The proof of Theorem 5.1 makes use of technical lemmas in Appendix A.6.

Recall from Appendix A.1 that if we define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$ then the density of Z_i is given by

$$(Z_i \mid \hat{\alpha}_i, d_i, \tau, \sigma^2) \sim \text{CCH}\,(Z_i \mid 1, 1/2, 1, s_i, 1, \theta_i)\,. \qquad \text{(A.11)}$$

where $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$ and $\theta_i = (\tau^2 d_i^2)^{-1}$. Consider the case where $d_i = 1$ for all i and $\tau^2 = 1$, i.e., $\theta_i = 1$ for all i. From Equation (A.6), the risk estimate is $SURE = \sum_{i=1}^n SURE_i$ with

$$\begin{aligned}SURE_i &= 2\sigma^2[1-\mathbb{E}(Z_i)+s_i\mathbb{E}(Z_i^2)+s_i\text{Var}(Z_i)],\\ &\leq 2\sigma^2[1-\mathbb{E}(Z_i)+s_i+s_i\text{Var}(Z_i)] = \check{R}_i.\end{aligned}$$

We begin by showing that the upper bound $\check{R}_i = 2\sigma^2[1-\mathbb{E}(Z_i)+s_i+s_i\text{Var}(Z_i)]$ is convex in s_i when $s_i \in (0,1)$. It suffices to show $-\mathbb{E}(Z_i)$ and $s_i\text{Var}(Z_i)$ are separately convex. First, $(\partial^2/\partial^2 s_i)\mathbb{E}(Z_i) = \mathbb{E}\{(Z_i-\mu_i)^3\} \leq 0$, by Lemmas A.2 and A.4, proving $-\mathbb{E}(Z_i)$ is convex. Next,

$$\begin{aligned}\frac{\partial^2}{\partial s_i^2}\{s_i\text{Var}(Z_i)\} &= \frac{\partial}{\partial s_i}\left[\text{Var}(Z_i)+s_i\frac{\partial}{\partial s_i}\{\text{Var}(Z_i)\}\right]\\ &= 2\frac{\partial}{\partial s_i}\{\text{Var}(Z_i)\}+s_i\frac{\partial^2}{\partial s_i^2}\{\text{Var}(Z_i)\}\\ &= -2\mathbb{E}(Z_i-\mu_i)^3 - s_i\frac{\partial}{\partial s_i}\mathbb{E}(Z_i-\mu_i)^3 \quad \text{(by Lemma A.2)}\\ &= -2\mathbb{E}(Z_i-\mu_i)^3+s_i\mathbb{E}(Z_i-\mu_i)^4, \quad \text{(by Lemma A.5)}\\ &\geq 0,\end{aligned}$$

where the last inequality follows by Lemma A.4. Thus, since $\check{R}_i$ is convex, it lies entirely below the straight line joining the two end points for $s_i \in (0,1)$. But $\check{R}_i|_{s_i=0} \leq 2\sigma^2/3 = 0.67\sigma^2$ (by Equation (A.8)) and

$$\check{R}_i|_{s_i=1} \leq 2\sigma^2\left[1-0.57+1+0.43-(0.57)^2\right] = 3.07\sigma^2,$$

by Equations (A.9) and (A.10).Thus, by convexity

$$SURE_i \leq \check{R}_i \leq 0.67\sigma^2 + s_i(3.07 - 0.67)\sigma^2 = (0.67 + 2.4s_i)\sigma^2 \text{ for } s_i \in (0,1) \tag{A.12}$$

We remark here that our simulations suggest $SURE_i$ itself is convex, not just the upper bound $\check{R}_i$, although a proof seems elusive. Nevertheless, as we shall see below, the convexity of $\check{R}_i$ is sufficient for our purposes.

Next, consider the interval $s_i \in (1,3)$. Noting that both $\mathbb{E}(Z_i)$ and $\mathbb{E}(Z_i^2)$ are monotone decreasing functions of s_i we have

$$SURE_i \leq 2\sigma^2[1 - \mathbb{E}(Z_i)|_{s_i=3} + 2s_i\{\mathbb{E}(Z_i^2)|_{s_i=1}\} - s_i\{\mathbb{E}(Z_i)|_{s_i=3}\}^2]$$

But,

$$\mathbb{E}(Z_i)|_{s_i=3,\theta_i=1} = \frac{\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-3z_i)dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}}\exp(-3z_i)dz_i} = 0.35.$$

$\mathbb{E}(Z_i^2)|_{s_i=1} < 0.43$ from Equation (A.9). Thus,

$$SURE_i \leq 2\sigma^2[1 - 0.35 + 0.86s_i - s_i(0.35)^2\}^2] = 2\sigma^2(0.65 + 0.74s_i) \text{ for } s_i \in (1,3). \tag{A.13}$$

Using the upper bound from Theorem 4.2,

$$SURE_i \leq 11.55\sigma^2 \text{ for } s_i \geq 3. \tag{A.14}$$

When $\alpha_i = 0$, we have that $\hat{\alpha}_i \sim \mathcal{N}(0, \sigma^2 d_i^{-2})$. Thus, $\hat{\alpha}_i^2 d_i^2/\sigma^2 \sim \chi^2(1)$. Since $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$ we have that $p(s_i) = (\pi)^{-1/2} s_i^{-1/2}\exp(-s_i)$ for $s_i \in (0,\infty)$. Combining Equations (A.12), (A.13) and (A.14) we have

$$\begin{aligned}
\text{Risk}_i = \mathbb{E}(SURE_i) \leq & \int_0^1 \sigma^2(0.67 + 2.4s_i)\pi^{-1/2}s_i^{-1/2}\exp(-s_i)ds_i \\
& + \int_1^3 2\sigma^2(0.65 + 0.74s_i)\pi^{-1/2}s_i^{-1/2}\exp(-s_i)ds_i \\
& + \int_3^\infty 11.55\sigma^2\pi^{-1/2}s_i^{-1/2}\exp(-s_i)ds_i \\
& = 1.75\sigma^2.
\end{aligned}$$

A.6. Technical lemmas

Lemma A.1 *If $Z \sim \text{CCH}(p,q,r,s,\nu,\theta)$, then $(\partial/\partial s)\mathbb{E}(Z^k) = \mathbb{E}(Z)\mathbb{E}(Z^k) - \mathbb{E}(Z^{k+1})$.*

Lemma A.2 *If $Z \sim \text{CCH}(p,q,r,s,\nu,\theta)$, then $(\partial^2/\partial^2 s)\mathbb{E}(Z) = -(\partial/\partial s)\text{Var}(Z) = \mathbb{E}\{(Z - \mu)^3\}$, where $\mu = \mathbb{E}(Z)$.*

Lemma A.3 *If $Z \sim \text{CCH}(p,q,r,s,\nu,\theta)$, then $(\partial/\partial\theta)\mathbb{E}(Z) = -\text{Cov}(Z,W)$, for $W = (1 - \nu Z)\{\theta + (1-\theta)\nu Z\}^{-1}$. If $0 < \theta \leq 1$ then $(\partial/\partial\theta)\mathbb{E}(Z) > 0$.*

Lemma A.4 *If $Z \sim \text{CCH}(p,q,r,s,1,1)$ with $q > p$, then $\mathbb{E}(Z-\mu)^3 \leq 0$, where $\mu = \mathbb{E}(Z)$.*

Lemma A.5 *If $Z \sim \text{CCH}(p,q,r,s,\nu,\theta)$, then $(\partial/\partial s)\mathbb{E}(Z-\mu)^3 = -\mathbb{E}\{(Z-\mu)^4\}$, where $\mu = \mathbb{E}(Z)$.*

A.6.1. Proof of Lemma A.1

Let, $Z \sim \text{CCH}(p, q, r, s, \nu, \theta)$. Then for any integer k

$$\mathbb{E}(Z^k) = \frac{\int_0^{1/\nu} z^{k+p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}.$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial s}\mathbb{E}(Z^k) =& \frac{\int_0^{1/\nu} -z^{k+p}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz} \\ &- \left[\frac{\int_0^{1/\nu} z^{k+p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right. \\ &\left. \times \frac{\int_0^{1/\nu} -z^{p}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\ =& -\mathbb{E}(Z^{k+1}) + \mathbb{E}(Z)\mathbb{E}(Z^k).\end{aligned}$$

For an alternative proof directly using the $H(\cdot)$ functions, see Appendix D of Gordy (1998).

A.6.2. Proof of Lemma A.2

Let, $Z \sim \text{CCH}(p, q, r, s, \nu, \theta)$. From Lemma A.1, $(\partial/\partial s)\mathbb{E}(Z) = -\mathbb{E}(Z^2) + \{\mathbb{E}(Z)\}^2 = -\text{Var}(Z)$. Let $\mu = \mathbb{E}(Z)$. Then,

$$\begin{aligned}\frac{\partial^2}{\partial s^2}\mathbb{E}(Z) =& -\frac{\partial}{\partial s}\text{Var}(Z) \\ =& -\frac{\partial}{\partial s}\left[\frac{\int_0^{1/\nu}(z-\mu)^2 z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\ =& \frac{\int_0^{1/\nu}(z-\mu)^2 z^{p}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz} \\ &- \left[\frac{\int_0^{1/\nu}(z-\mu)^2 z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right. \\ &\left. \times \frac{\int_0^{1/\nu} z^{p}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\ =& \text{Cov}(Z, (Z-\mu)^2) \\ =& \mathbb{E}[(Z-\mu)\{(Z-\mu)^2 - \mathbb{E}(Z-\mu)^2\}] \\ =& \mathbb{E}\{(Z-\mu)^3\} - \text{Var}(Z)\mathbb{E}(Z-\mu) = \mathbb{E}\{(Z-\mu)^3\}.\end{aligned}$$

A.6.3. Proof of Lemma A.3

Let $Z \sim \text{CCH}(p, q, r, s, \nu, \theta)$ and $W = (1 - \nu Z)\{\theta + (1 - \theta)\nu Z\}^{-1}$. Then,

$$\begin{aligned}
\frac{\partial}{\partial \theta}\mathbb{E}(Z) =& - \frac{\int_0^{1/\nu} z^p(1-\nu z)^q\{\theta + (1-\theta)\nu z\}^{-(r+1)}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz} \\
&+ \left[\frac{\int_0^{1/\nu} z^p(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right. \\
&\left. \times \frac{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q}\{\theta + (1-\theta)\nu z\}^{-(r+1)}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\
=& - \mathbb{E}(ZW) + \mathbb{E}(Z)\mathbb{E}(W) = -\text{Cov}(Z, W).
\end{aligned}$$

When $0 < \theta \leq 1$, it is obvious that Z and W are negatively correlated, and thus $-\text{Cov}(Z, W) > 0$.

A.6.4. Proof of Lemma A.4

Let $Z \sim \text{CCH}(p, q, r, s, 1, 1)$. Then,

$$\mathbb{E}(Z - \mu)^3 = \frac{\int_0^1 (z-\mu)^3 z^{p-1}(1-z)^{q-1}\exp(-sz)dz}{\int_0^1 z^{p-1}(1-z)^{q-1}\exp(-sz)dz},$$

which can be seen to have the same sign as the third central moment, or skewness of a $\text{Beta}(p, q)$ random variable, which is negative when $q > p$.

A.6.5. Proof of Lemma A.5

Let, $Z \sim \text{CCH}(p, q, r, s, \nu, \theta)$. Let $\mu = \mathbb{E}(Z)$. Then,

$$\begin{aligned}
\frac{\partial}{\partial s}\mathbb{E}(Z-\mu)^3 =& - \frac{\int_0^{1/\nu}(z-\mu)^3 z^p(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz} \\
&+ \left[\frac{\int_0^{1/\nu}(z-\mu)^3 z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right. \\
&\left. \times \frac{\int_0^{1/\nu} z^{p}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\
=& - \text{Cov}(Z, (Z-\mu)^3) \\
=& - \mathbb{E}[(Z-\mu)\{(Z-\mu)^3 - \mathbb{E}(Z-\mu)^3\}] \\
=& - \mathbb{E}\{(Z-\mu)^4\} + \mathbb{E}(Z-\mu)^3\mathbb{E}(Z-\mu) = -\mathbb{E}\{(Z-\mu)^4\}.
\end{aligned}$$

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. ISSN 0018-9286. doi: 10.1109/TAC.1974.

1100705. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1100705.

Artin Armagan, Merlise Clyde, and David B Dunson. Generalized beta mixtures of Gaussians. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. N. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 523–531, 2011.

Artin Armagan, David B Dunson, and Jaeyong Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143, 2013.

O Barndorff-Nielsen, John Kent, and Michael Sørensen. Normal variance-mean mixtures and z distributions. *International Statistical Review*, 50:145–159, 1982.

Anindya Bhadra, Jyotishka Datta, Nicholas G Polson, and Brandon Willard. Default bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969, 2016.

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

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015. doi: 10.1080/01621459.2014.960967. URL http://dx.doi.org/10.1080/01621459.2014.960967.

Nicholas H Bingham, Charles M Goldie, and Jef L Teugels. *Regular variation*, volume 27 of *Encyclopedia of mathematics and its applications*. Cambridge University Press, Cambridge, 1989.

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

George Casella and Jiunn Tzon Hwang. Limit expressions for the risk of James–Stein estimators. *Canadian Journal of Statistics*, 10(4):305–309, 1982.

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.

Hyonho Chun and Sündüz Keles. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

Merlise Clyde, Heather Desimone, and Giovanni Parmigiani. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208, 1996.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

David G. T. Denison and Edward I. George. *Bayesian prediction with adaptive ridge estimators*, volume 8 of *IMS Collections*, pages 215–234. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2012. doi: 10.1214/11-IMSCOLL815. URL http://dx.doi.org/10.1214/11-IMSCOLL815.

Lee H Dicker. Variance estimation in high-dimensional linear models. *Biometrika*, 101(2): 269–284, 2014.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, to appear, 2017.

Bradley Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association*, 99(467):619–642, 2004.

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.

Dean P Foster and Edward I George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975, 1994.

Ildiko E. Frank and Jerome H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993. ISSN 00401706. URL http://www.jstor.org/stable/1269656.

Hong-Ye Gao and Andrew G Bruce. Waveshrink with firm shrinkage. *Statistica Sinica*, pages 855–874, 1997.

Edward I. George, Feng Liang, and Xinyi Xu. Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, 34(1):78–91, 02 2006. doi: 10.1214/009053606000000155. URL http://dx.doi.org/10.1214/009053606000000155.

Prasenjit Ghosh, Xueying Tang, Malay Ghosh, and Arijit Chakrabarti. Asymptotic properties of bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Analysis*, 11(3):753–796, 09 2016. doi: 10.1214/15-BA973. URL http://dx.doi.org/10.1214/15-BA973.

Michael B Gordy. A generalization of generalized beta distributions. In *Finance and Economics Discussion Series*. Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board, 1998.

Jim E Griffin and Philip J Brown. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, New York, 2nd edition, 2009. URL http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634.

Ian T. Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973. doi: 10.1080/00401706.1973.10489103. URL http://dx.doi.org/10.1080/00401706.1973.10489103.

Ryan Martin and Stephen G. Walker. Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206, 2014. doi: 10.1214/14-EJS949. URL http://dx.doi.org/10.1214/14-EJS949.

C. Masreliez. Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control*, 20(1):107–110, Feb 1975. ISSN 0018-9286. doi: 10.1109/TAC.1975.1100882.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495): 1125–1138, 2011.

LR Pericchi and AFM Smith. Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 54:793–804, 1992.

Nicholas G Polson and James G Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, James O. Berger, A. P. Dawid, David Heckerman, Adrian F. M. Smith, and Mike West, editors, *Bayesian Statistics 9*, Oxford, 2011. Oxford University Press.

Nicholas G Polson and James G Scott. Local shrinkage rules, Lévy processes, and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311, 2012a.

Nicholas G. Polson and James G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 12 2012b. doi: 10.1214/12-BA730. URL http://dx.doi.org/10.1214/12-BA730.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

Veronika Ročková and Edward I George. The spike-and-slab lasso. *Journal of the American Statistical Association*, to appear, 2016.

Veronika Ročková and Edward I. George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846, 2014. doi: 10.1080/01621459.2013.869223. URL http://dx.doi.org/10.1080/01621459.2013.869223.

Charles M Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 197–206, Berkeley, Calif., 1956. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200501656.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 11 1981. doi: 10.1214/aos/1176345632. URL http://dx.doi.org/10.1214/aos/1176345632.

Gergely Szakács, Jean-Philippe Annereau, Samir Lababidi, Uma Shankavaram, Angela Arciello, Kimberly J. Bussey, William Reinhold, Yanping Guo, Gary D. Kruh, Mark Reimers, John N. Weinstein, and Michael M. Gottesman. Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells. *Cancer Cell*, 6(2):129 – 137, 2004. ISSN 1535-6108. doi: http://dx.doi.org/10.1016/j.ccr.2004.06.026. URL http://www.sciencedirect.com/science/article/pii/S1535610804002065.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58:267–288, 1996.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 04 2012. doi: 10.1214/12-AOS1003. URL http://dx.doi.org/10.1214/12-AOS1003.

SL van der Pas, BJK Kleijn, and AW van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

S.L. van der Pas, J-B. Salomond, and J. Schmidt-Hieber. Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics*, 10(1): 976–1000, 2016.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier, North-Holland, 1986.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010.

Supplementary Material to
Prediction risk for the horseshoe regression

Anindya Bhadra
Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA.
bhadra@purdue.edu

Jyotishka Datta
Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA.
jd033@uark.edu

Yunfan Li
Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA.
li896@purdue.edu

Nicholas G. Polson and Brandon Willard
The University of Chicago Booth School of Business, 5807 S. Woodlawn Ave., Chicago, IL 60637, USA.
ngp@chicagobooth.edu, brandonwillard@gmail.com

S.1. Additional simulations

We provide additional simulation results, complementing the results in Table 2. For each simulation setting, we report SURE when a formula is available. We also report the average out of sample prediction SSE (standard deviation of SSE) computed based on one training set and 200 testing sets. For each setting, $n = 100$. The methods under consideration are ridge regression (RR), principal components regression (PCR), the lasso, the adaptive lasso (A_LASSO), the minimax concave penalty (MCP), the smoothly clipped absolute deviation (SCAD) penalty and the proposed horseshoe regression (HS). The method with the **lowest SSE** is in bold and that with *lowest SURE* is in italics for each setting. The features of these additional simulations include the following.

1. We explore a higher dimensional case ($p = 1000$) for each setting.

2. We incorporate two non-convex regression methods for comparisons. These are SCAD (Fan and Li, 2001) and MCP (Zhang, 2010).

3. We explore different choices of the design matrix X. These include three cases: (i) X is generated from a factor model, where it is relatively ill-conditioned (as in Table 2), (ii) X is generated from a standard normal, where it is well-conditioned and (iii) X is exactly orthogonal, with all singular values equal to 1. These are reported in corresponding table captions.

4. We explore different choices of true α. These include three cases: (i) Sparse-robust α, where most elements of α are close to zero and a few are large, (ii) null α, where all elements of α are zero and (iii) dense α, where all elements are non-zero. Exact settings and the value of $||\alpha||^2$ are reported in the table captions.

The major finding is that the horseshoe regression outperforms the other global shrinkage methods (ridge and PCR) when α is sparse-robust, which is consistent with the theoretical observation in Section 5. It also outperforms the other selection-based methods in this case. On the other hand, the dense α case is most often favorable to ridge regression, while the null α case is favorable to selection-based methods such as the lasso, adaptive lasso, MCP or SCAD, due to the ability of these methods to produce exact zero estimates. However, the selection-based methods perform considerably worse compared to both global and global-local shrinkage methods in the dense α case.

Table S.1: Sparse-robust α (five large coefficients equal to 10 and other coefficients equal to 0.5 or -0.5 randomly, giving $\sum_{i=1}^{n} \alpha_i^2 = 523.75$); X generated by a factor model with 4 factors, each factor follows a standard normal distribution; d_1/d_n is the ratio of largest and smallest singular values of X.

		RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	d_1/d_n	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	2360.43	165.45	159.83	163.80	161.62	122.78	145.07	132.25	127.07	127.85	*116.01*	**123.07**
			(22.02)		(21.28)		(19.39)	(17.57)	(16.71)	(17.19)		(16.43)
200	28.47	188.13	206.39	217.40	244.71	174.48	162.94	**148.41**	154.01	157.73	*160.89*	152.37
			(28.61)		(29.80)		(24.44)	(22.48)	(23.17)	(23.23)		(22.75)
300	22.76	192.35	212.05	266.84	280.25	*155.26*	190.09	175.46	172.18	176.29	157.50	**164.17**
			(28.50)		(32.62)		(26.20)	(22.17)	(21.55)	(22.19)		(22.85)
400	21.81	194.73	199.36	337.32	328.48	179.45	182.89	197.25	199.08	198.40	*172.63*	**165.15**
			(28.75)		(34.79)		(27.41)	(25.02)	(25.52)	(25.31)		(24.67)
500	18.18	196.03	180.12	410.82	379.03	158.07	173.82	223.21	224.91	226.76	*166.10*	**161.77**
			(27.16)		(39.41)		(26.50)	(27.98)	(29.65)	(29.26)		(24.22)
1000	15.20	197.91	184.86	669.69	736.69	196.83	205.28	345.26	344.04	344.04	*191.64*	**182.18**
			(26.42)		(56.58)		(29.56)	(36.60)	(37.34)	(37.34)		(25.43)

Table S.2: Null α ($\sum_{i=1}^{n} \alpha_i^2 = 0$); X is the same as in Table S.1.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	*88.23*	100.86	92.85	113.28	87.36	100.81	**100.70**	100.81	100.81	92.42	102.31
		(13.20)		(14.91)		(13.29)	(13.21)	(13.29)	(13.29)		(13.72)
200	121.30	107.68	128.83	115.65	*117.90*	105.77	**100.32**	104.39	101.78	122.29	111.39
		(15.70)		(16.28)		(15.06)	(14.80)	(14.93)	(14.89)		(16.12)
300	125.78	101.36	139.96	124.37	*108.85*	111.85	**101.30**	104.89	102.91	119.67	112.00
		(13.99)		(17.35)		(15.37)	(14.02)	(14.27)	(14.00)		(15.76)
400	113.00	99.50	113.50	99.41	*102.81*	111.92	114.62	**99.40**	110.30	113.42	107.20
		(13.12)		(13.09)		(15.51)	(15.80)	(13.20)	(15.20)		(14.90)
500	90.74	101.04	*88.26*	107.31	90.26	99.49	**99.06**	99.49	99.49	101.55	102.93
		(14.17)		(15.08)		(14.16)	(14.04)	(14.16)	(14.16)		(14.68)
1000	88.86	100.34	85.67	103.47	*82.51*	100.43	**99.52**	100.43	100.41	99.73	104.84
		(14.00)		(14.29)		(13.90)	(13.70)	(13.90)	(14.00)		(14.96)

Table S.3: Dense α (all coefficients equal to 2, giving $\sum_{i=1}^{n} \alpha_i^2 = 400$); X is the same as in Table S.1.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	*162.49*	**159.94**	177.47	175.19	194.86	203.89	504.55	491.67	491.67	185.46	173.89
		(21.60)		(22.36)		(28.11)	(46.13)	(45.31)	(45.31)		(23.63)
200	*183.75*	**200.92**	196.06	233.12	211.99	232.36	960.77	895.83	911.94	204.10	228.18
		(27.97)		(31.18)		(31.06)	(59.48)	(60.85)	(60.19)		(31.21)
300	*189.38*	**209.92**	200.39	225.92	216.01	524.84	1344.27	1298.80	1298.80	206.99	227.55
		(27.88)		(30.15)		(69.45)	(71.29)	(77.97)	(77.97)		(29.98)
400	*193.02*	**195.01**	197.74	217.68	218.16	306.15	1768.05	1675.73	1675.73	207.81	213.91
		(28.92)		(31.02)		(42.65)	(78.92)	(75.86)	(75.86)		(31.14)
500	*194.85*	**175.46**	208.87	201.18	220.34	743.40	2154.61	2082.54	2081.42	207.93	188.93
		(26.52)		(29.07)		(100.54)	(92.70)	(92.93)	(93.37)		(28.10)
1000	*197.37*	**181.65**	247.40	197.59	224.75	210.78	4280.80	4075.00	4075.00	203.47	186.48
		(26.50)		(27.76)		(29.70)	(145.28)	(138.72)	(138.72)		(26.95)

Table S.4: Sparse-robust α (five large coefficients equal to 10 and other coefficients equal to 0.5 or -0.5 randomly, giving $\sum_{i=1}^{n} \alpha_i^2 = 523.75$); X follows a standard normal distribution; d_1/d_n is the ratio of largest and smallest singular values of X.

		RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	d_1/d_n	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	351.2	196.72	188.67	228.78	231.34	207.63	425.67	2537.23	2573.27	2573.27	*195.22*	**188.52**
			(29.04)		(34.36)		(59.68)	(112.58)	(128.46)	(128.46)		(28.90)
200	5.73	*199.84*	**193.41**	221.35	206.25	218.26	1618.40	4849.94	4915.72	4964.26	201.91	194.14
			(28.36)		(28.28)		(211.54)	(146.61)	(186.74)	(186.55)		(28.45)
300	3.63	*199.91*	**217.43**	8538.46	8082.93	222.62	1926.13	13132.01	7316.38	7316.39	200.92	219.97
			(27.91)		(320.98)		(248.97)	(281.12)	(218.82)	(218.83)		(28.13)
400	2.89	*199.94*	**197.47**	228.38	223.43	224.53	2384.04	9593.41	9695.47	9695.47	200.31	197.69
			(27.43)		(31.13)		(299.58)	(210.42)	(323.72)	(323.72)		(27.46)
500	2.51	*199.95*	**193.86**	256.09	273.63	224.96	471.73	11980.15	11991.11	11991.11	200.15	194.17
			(27.26)		(33.97)		(60.52)	(235.77)	(272.80)	(272.80)		(27.24)
1000	1.88	199.98	185.85	605.64	560.45	222.18	5781.13	23866.06	24566.13	24566.13	*199.96*	**185.79**
			(27.17)		(45.94)		(759.08)	(326.06)	(941.16)	(941.16)		(27.17)

Table S.5: Null α ($\sum_{i=1}^{n} \alpha_i^2 = 0$); X is the same as in Table S.4.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	118.45	119.12	96.35	106.88	*92.06*	101.21	**100.52**	101.21	101.21	119.11	114.69
		(18.19)		(15.18)		(14.33)	(14.20)	(14.33)	(14.33)		(17.47)
200	136.93	135.02	96.49	100.14	*94.54*	100.15	**100.13**	100.39	102.06	126.19	126.34
		(21.74)		(14.77)		(14.69)	(14.70)	(14.88)	(15.26)		(20.13)
300	152.52	160.29	119.00	131.94	*118.15*	100.71	**100.49**	100.71	100.71	140.91	140.08
		(21.61)		(18.01)		(14.48)	(14.37)	(14.48)	(14.48)		(18.93)
400	158.64	159.13	100.88	104.06	*96.30*	103.07	**100.46**	103.03	103.03	138.62	132.14
		(23.15)		(15.59)		(15.11)	(14.82)	(15.04)	(15.04)		(19.62)
500	166.06	158.83	98.64	98.10	*94.30*	100.36	**97.99**	100.36	100.36	140.14	131.53
		(23.35)		(14.50)		(14.79)	(14.50)	(14.79)	(14.79)		(19.59)
1000	181.23	169.22	89.95	100.66	*87.79*	100.07	**99.80**	100.66	100.51	141.11	138.94
		(25.25)		(14.10)		(14.03)	(14.00)	(14.08)	(14.07)		(21.12)

Table S.6: Dense α (all coefficients equal to 2, giving $\sum_{i=1}^{n} \alpha_i^2 = 400$); X is the same as in Table S.4.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	*193.13*	**188.60**	206.31	200.53	222.52	210.23	40019.73	40063.42	40063.42	199.25	191.74
		(28.91)		(29.51)		(31.36)	(690.71)	(717.27)	(717.27)		(29.37)
200	*199.76*	**193.93**	392.38	349.52	224.73	316.05	80016.11	80187.49	80187.49	200.14	194.48
		(28.41)		(37.90)		(42.77)	(983.86)	(1102.52)	(1102.52)		(28.50)
300	*199.88*	**217.60**	400.63	445.11	222.50	16845.87	120071.46	123161.75	123161.75	200.02	217.75
		(27.93)		(43.70)		(2167.53)	(1191.24)	(3757.26)	(3757.26)		(27.92)
400	*199.92*	**196.61**	627.97	618.38	222.51	43325.17	159926.82	161662.45	161662.45	200.00	196.70
		(27.35)		(59.10)		(5447.99)	(1418.60)	(3304.65)	(3304.65)		(27.35)
500	*199.94*	**193.02**	794.03	823.27	225.01	6497.32	199982.59	200043.69	200043.69	200.00	193.33
		(27.16)		(62.69)		(824.72)	(1550.64)	(1647.47)	(1647.47)		(27.18)
1000	*199.97*	**185.98**	2116.68	2108.78	224.77	3145.06	399770.90	400168.82	400168.82	200.03	186.02
		(27.17)		(101.50)		(411.60)	(2359.10)	(2934.25)	(2934.25)		(27.17)

Table S.7: Sparse-robust α (five large coefficients equal to 10 and other coefficients equal to 0.5 or -0.5 randomly, giving $\sum_{i=1}^{n} \alpha_i^2 = 523.75$); X with all singular values equal to 1.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	183.50	179.99	291.45	275.49	139.29	139.39	129.30	126.70	126.19	*131.81*	**122.60**
		(25.31)		(32.44)		(20.36)	(19.20)	(18.79)	(18.83)		(18.72)
200	184.47	196.14	261.65	277.35	135.93	150.17	**128.76**	129.25	129.58	*129.15*	131.16
		(28.79)		(33.30)		(21.76)	(17.75)	(17.61)	(17.90)		(18.63)
300	182.50	192.24	267.96	269.04	126.35	146.72	132.26	132.05	132.17	*119.09*	**128.72**
		(25.37)		(30.40)		(18.96)	(17.85)	(17.87)	(17.73)		(17.34)
400	184.03	178.58	311.01	287.95	145.28	139.68	128.94	127.57	127.41	*130.13*	**123.83**
		(25.20)		(32.98)		(19.20)	(17.40)	(17.42)	(17.43)		(17.02)
500	183.81	173.44	278.35	268.85	147.54	139.74	126.65	127.19	126.30	*132.70*	**120.78**
		(24.08)		(30.78)		(19.98)	(18.36)	(18.31)	(18.17)		(17.29)
1000	185.36	166.39	280.59	262.54	124.52	130.61	128.83	129.78	129.47	*119.24*	**125.49**
		(23.16)		(30.01)		(18.02)	(17.50)	(17.70)	(17.62)		(17.37)

Table S.8: Null α ($\sum_{i=1}^{n} \alpha_i^2 = 0$); X with all singular values equal to 1.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	94.70	**100.13**	97.63	102.62	*94.54*	100.15	**100.13**	100.15	100.15	99.92	101.44
		(14.71)		(15.37)		(14.69)	(14.70)	(14.69)	(14.69)		(15.06)
200	115.52	103.43	111.09	118.11	*109.16*	122.81	100.49	112.22	**100.80**	116.62	106.72
		(14.80)		(16.91)		(17.79)	(14.37)	(16.20)	(14.53)		(15.47)
300	98.74	100.49	99.45	113.35	*96.40*	103.03	**100.46**	103.03	103.03	103.10	102.40
		(14.80)		(16.49)		(15.04)	(14.82)	(15.04)	(15.04)		(14.87)
400	96.78	**97.99**	103.88	102.24	*94.02*	103.08	**97.99**	101.71	103.17	99.64	101.01
		(14.49)		(15.12)		(14.96)	(14.50)	(14.81)	(14.97)		(14.78)
500	88.55	**99.97**	89.06	100.91	*87.74*	100.65	99.98	100.65	100.65	93.63	101.53
		(14.81)		(14.72)		(14.87)	(14.83)	(14.87)	(14.87)		(14.98)
1000	88.87	**100.94**	91.96	107.30	*88.45*	101.14	100.95	101.62	101.14	94.34	102.26
		(14.17)		(15.40)		(14.14)	(14.17)	(14.30)	(14.14)		(14.48)

Table S.9: Dense α (all coefficients equal to 2, giving $\sum_{i=1}^{n} \alpha_i^2 = 400$); X with all singular values equal to 1.

	RR		PCR		LASSO		A_LASSO	MCP	SCAD	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SSE	SSE	SURE	SSE
100	*177.89*	**183.69**	220.87	200.41	203.16	307.44	502.55	505.43	505.43	200.80	204.16
		(25.16)		(26.35)		(43.19)	(41.53)	(43.20)	(43.20)		(27.46)
200	*181.88*	**188.39**	207.49	239.95	214.11	255.88	499.88	498.84	498.84	205.16	217.46
		(27.10)		(33.45)		(35.61)	(41.69)	(42.90)	(42.90)		(30.81)
300	*176.64*	**193.53**	215.03	205.00	196.19	250.76	496.84	497.60	495.93	199.33	212.19
		(25.76)		(26.70)		(32.66)	(45.74)	(47.41)	(45.90)		(27.68)
400	*174.62*	**195.83**	248.90	249.51	209.80	221.41	495.21	494.17	494.17	198.84	206.49
		(26.36)		(32.46)		(29.85)	(40.50)	(40.89)	(40.89)		(28.41)
500	*179.13*	**173.13**	202.78	192.97	214.36	201.27	501.67	503.19	503.19	205.16	193.40
		(23.18)		(25.08)		(25.63)	(38.90)	(38.93)	(38.93)		(25.60)
1000	*179.32*	**173.71**	225.50	195.46	209.35	248.20	503.44	507.29	507.29	204.51	194.73
		(24.14)		(25.78)		(30.87)	(41.48)	(41.38)	(41.38)		(26.67)