High Dimensional Linear Regression via the R2-D2 Shrinkage Prior

Yan Zhang,† Brian J. Reich† and Howard D. Bondell‡

September 12, 2017

Abstract

We propose a new class of shrinkage priors for linear regression, the R-squared induced Dirichlet decomposition (R2-D2) prior. The prior is induced by a Beta prior on the coefficient of determination, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. We demonstrate both theoretically and empirically the advantages of the proposed prior over a number of common shrinkage priors, including the Horseshoe, Horseshoe+, generalized double Pareto, and Dirichlet-Laplace priors. Specifically, the proposed prior possesses an unbounded density around zero with polynomial order, and the heaviest tails among these common shrinkage priors. We demonstrate that this can lead to improved empirical estimation and prediction accuracy for simulated and real data applications. We show that the Bayes estimator of the proposed prior converges to the truth at a Kullback-Leibler super-efficient rate, attaining a sharper information theoretic bound than existing common shrinkage priors. We also demonstrate that our proposed prior yields a consistent posterior.

Keywords: Dirichlet-Laplace; Global-local shrinkage; Horseshoe; Horseshoe+; Kullback-Leibler efficiency; Linear regression.

†Department of Biostatistics, Johns Hopkins University, yzhan284@jhu.edu
‡Department of Statistics, North Carolina State University, brian.reich@ncsu.edu
‡Department of Statistics, North Carolina State University, bondell@stat.ncsu.edu
1 Introduction

Consider the linear regression model,

\[ Y_i = x_i^T \beta + \varepsilon_i, \ i = 1, \cdots, n, \]  

where \( Y_i \) is the \( i \)th response, \( x_i \) is the \( p \)-dimensional vector of covariates for \( i \)th observation, \( \beta = (\beta_1, \cdots, \beta_p)^T \) is the coefficient vector, and the \( \varepsilon_i \)'s are the error terms assumed be normal and independent with \( E(\varepsilon_i) = 0 \) and \( \text{var}(\varepsilon_i) = \sigma^2 \). High-dimensional data with \( p > n \) in this context is common in diverse application areas. It is well known that maximum likelihood estimation performs poorly in this setting, and this motivates a number of approaches in shrinkage estimation and variable selection. In the Bayesian framework, there are two main approaches to address such problems: two component discrete mixture prior (also referred as spike and slab prior) and continuous shrinkage priors. The discrete mixture priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ishwaran and Rao, 2005; Narisetty et al., 2014) put a point mass (spike) at \( \beta_j = 0 \) and a continuous prior (slab) for the terms with \( \beta_j \neq 0 \). Although these priors have an intuitive and appealing representation, they lead to computational issues due to the spread of posterior probability over the \( 2^p \) models formed by including subsets of the coefficients to zero. The spike-and-slab lasso proposed in Ročková and George (2016), which is a continuous version of the spike-and-slab prior with Laplace spike and slab, is also a class of spike-and-slab prior and the implementation requires applying the stochastic search variable selection strategy proposed in George and McCulloch (1993).

These issues with discrete mixture priors motivate continuous shrinkage priors. The shrinkage priors are essentially written as global-local scale mixture Gaussian family as summarized in Polson and Scott (2010), i.e.,

\[ \beta_j \mid \phi_j, \omega \sim N(0, \omega \phi_j), \ \phi_j \sim \pi(\phi_j), \ (\omega, \sigma^2) \sim \pi(\omega, \sigma^2), \]

where \( \omega \) represents the global shrinkage, while \( \phi_j \)'s are the local variance components. Current existing global-local priors exhibit desirable theoretic and empirical properties.
The priors are continuous but have high concentration at zero and heavy tails, which reflects the prior that many covariates are irrelevant while a few have large effect, without explicitly having prior probability at $\beta_j = 0$. Some examples include normal-gamma mixtures (Griffin et al., 2010), Horseshoe (Carvalho et al., 2009, 2010), generalized Beta (Armagan et al., 2011), generalized double Pareto (Armagan et al., 2013a), Dirichlet-Laplace (Bhattacharya et al., 2015), and Horseshoe+ (Bhadra et al., 2016). Global-local priors have substantial computational advantages over the discrete mixture priors.

While continuous global-local shrinkage priors exhibit desirable theoretical, computational and empirical properties, they also have their own challenges. Since the posterior probability mass on zero is always zero, unlike the discrete mixture priors which directly generate sparse estimates, shrinkage priors require additional steps to go from the continuous posterior distribution to a sparse estimate. There are several methods to deal with this. The most common method is to threshold to decide which predictor to be included. For example, Carvalho et al. (2010) described a simple rule for Horseshoe prior that yields a sparse estimate. In addition to thresholding, there are two major approaches: penalized variable selection based on posterior credible regions, and decoupling shrinkage and selection. Bondell and Reich (2012) proposed the penalized credible region variable selection method, which fits the full model under a continuous shrinkage prior, and then selects the sparsest solution within the posterior credible region. Hahn and Carvalho (2015) proposed the decoupling shrinkage and selection method, which uses a loss function combining a posterior summarizer with an explicit parsimony penalty to induce sparse estimator.

We propose a new global-local prior, which we term $R^2$-induced Dirichlet Decomposition (R2-D2) prior. The coefficient of determination $R^2$ is defined as the square of the correlation coefficient between the original dependent variable and the modeled value. The motivation comes from the fact that it is hard to specify a $p$-dimensional prior on $\beta$ with high dimensional data, however, it is more direct to construct a prior on the 1-dimensional $R^2$. The proposed new prior is induced by a Beta($a, b$) prior on $R^2$, and then the total prior variance of the regression coefficients is decomposed through a Dirichlet prior. Prior information about $R^2$ collected from previous experiments can be coerced into the hyper-
parameters, $a$ and $b$, and then reflected on the prior of $\beta$. We show that the class of
the proposed new prior with different kernels induces a number of existing priors as special cases, such as the normal-gamma (Griffin et al., 2010) and Horseshoe (Carvalho et al., 2009, 2010) priors. The proposed new prior has many appealing properties, such as strongly shrinking small coefficients due to a tight peak at zero, allowing for large coefficients due to the heavy tails, and a hierarchical representation that leads to Gibbs sampler. We also offer a theoretical framework to compare different global-local priors. The proposed method compares favorably to the other global-local shrinkage priors in terms of both concentration around the origin and tail behavior. We also demonstrate that in the orthogonal design setup, the proposed new prior guarantees that the Bayes estimator converges to the truth at a Kullback-Leibler super-efficient rate. In fact, our new proposed prior attains a sharper information theoretic bound than the existing global-local priors, such as the Horseshoe (Carvalho et al., 2009, 2010) and Horseshoe+ (Bhadra et al., 2016) prior.

In terms of posterior properties, Armagan et al. (2013b) investigates the asymptotic behavior of posterior distributions of regression coefficients as $p$ grows with $n$. They prove the posterior consistency for some shrinkage priors, including the double-exponential prior, Student’s $t$ prior, generalized double Pareto prior, and the Horseshoe-like priors. Under similar conditions, Zhang and Bondell (arXiv:1602.01160) demonstrate posterior consistency for the Dirichlet-Laplace prior. van der Pas et al. (2016) propose general conditions on the priors to ensure posterior contraction at a minimax rate. In this paper, we prove that our proposed R2-D2 prior leads to consistent posterior distributions.

2 A New Class of Global-Local Shrinkage Priors

2.1 Motivation

The primary goal is to estimate the vector $\beta$ and select important covariates. Common Bayesian methods assume a prior on $\beta$ directly. In this paper, we start by placing a prior on a univariate function of $\beta$ with practical meaning, and then induce a prior on the $p$-dimensional $\beta$. 

4
Suppose that the predictor vectors \( x_1, \ldots, x_n \sim H(\cdot) \) independently, with \( E(x_i) = \mu \) and \( \text{cov}(x_i) = \Sigma \). Assume that for each \( i = 1, \ldots, n \), \( x_i \) is independent of the \( n \)-vector of errors, \( \varepsilon \), and the marginal variance of \( Y_i \) is then \( \text{var}(x^T \beta) + \sigma^2 \). For simplicity, we assume that the response is centered and covariates are standardized so that there is no intercept term in \((1)\), and all diagonal elements of \( \Sigma \) are 1. The coefficient of determination, \( R^2 \), can be calculated as the square of the correlation coefficient between the dependent variable, \( Y \), and the modeled value, \( x^T \beta \), i.e.,

\[
R^2 = \frac{\text{cov}(Y, x^T \beta)}{\text{var}(Y) \text{var}(x^T \beta)} = \frac{\text{cov}(x^T \beta + \varepsilon, x^T \beta)}{\text{var}(x^T \beta + \varepsilon) \text{var}(x^T \beta)} = \frac{\text{var}(x^T \beta)}{\text{var}(x^T \beta) + \sigma^2}.
\]

Consider a prior for \( \beta \) satisfying \( E(\beta) = 0 \) and \( \text{cov}(\beta) = \sigma^2 \Lambda \), where \( \Lambda \) is a diagonal matrix with diagonal elements \( \lambda_1, \ldots, \lambda_p \). Then

\[
\text{var}(x^T \beta) = E_x \{ \text{var}_\beta(x^T \beta \mid x) \} + \text{var}_x \{ E_\beta(x^T \beta \mid x) \} = E_x (\sigma^2 x^T \Lambda x) + \text{var}_x(0) = \sigma^2 E_x \{ \text{tr}(\Lambda x^T x) \} = \sigma^2 \text{tr}(\Lambda) = \sigma^2 \sum_{j=1}^p \lambda_j.
\]

Then \( R^2 \) is represented as

\[
R^2 = \frac{\text{var}(x^T \beta)}{\text{var}(x^T \beta) + \sigma^2} = \frac{\sigma^2 \sum_{j=1}^p \lambda_j}{\sigma^2 \sum_{j=1}^p \lambda_j + \sigma^2} = \frac{\sum_{j=1}^p \lambda_j}{\sum_{j=1}^p \lambda_j + 1} \equiv \frac{W}{W + 1},
\]

where \( W = \sum_{j=1}^p \lambda_j \) is the sum of the prior variances scaled by \( \sigma^2 \).

Suppose \( R^2 \sim \text{Beta}(a, b) \), a Beta distribution with shape parameters \( a \) and \( b \), then the induced prior density for \( W = R^2/(1 - R^2) \) is a Beta Prime distribution (Johnson et al., 1995) denoted as \( \text{BP}(a, b) \), with probability density function

\[
\pi_W(x) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{x^{a-1}}{(1 + x)^{n+b}}, \quad (x > 0).
\]

Therefore \( W \sim \text{BP}(a, b) \) is equivalent to the prior \( R^2 \sim \text{Beta}(a, b) \). The following section will induce a prior on \( \beta \) based on the distribution of the sum of prior variances \( W \).
2.2 The R2-D2 prior

Any prior of the form $E(\beta) = 0$, $\text{cov}(\beta) = \sigma^2 \Lambda$ and $W = \sum_{j=1}^{p} \lambda_j \sim \text{BP}(a, b)$ induces a Beta($a, b$) prior on the $R^2$. To construct a prior with such properties, we follow the global-local prior framework and express $\lambda_j = \phi_j \omega$ with $\sum_{j=1}^{p} \phi_j = 1$. Then $W = \sum_{j=1}^{p} \phi_j \omega = \omega$ is the total prior variability, and $\phi_j$ is the proportion of total variance allocated to the $j$-th covariate. It is natural to assume that $\omega \sim \text{BP}(a, b)$ and the variances across covariates have a Dirichlet prior with concentration parameter $(a, \cdots, a, a)$, i.e., $\phi = (\phi_1, \cdots, \phi_p) \sim \text{Dir}(a, \cdots, a)$. Since $\sum_{j=1}^{p} \phi_j = 1$, $E(\phi_j) = 1/p$, and $\text{var}(\phi_j) = (p-1)/\{p^2(pa_a + 1)\}$, then smaller $a_a$ would lead to larger variance of $\phi_j$, $j = 1, \cdots, p$, thus more $\phi_j$ would be close to zero with only a small proportion of larger components; while larger $a_a$ would lead to smaller variance of $\phi_j$, $j = 1, \cdots, p$, thus producing a more uniform $\phi$, i.e., $\phi \approx (1/p, \cdots, 1/p)$.

So $a_a$ controls the sparsity.

Assume a prior $K(.)$ on each dimension of $\beta$, with $K(\delta)$ denotes a kernel (density) with mean zero and variance $\delta$. The prior is summarized as

$$
\beta_j \mid \sigma^2, \phi_j, \omega \sim K(\sigma^2 \phi_j \omega), \; \phi \sim \text{Dir}(a, \cdots, a), \; \omega \sim \text{BP}(a, b).
$$

(3)

Such prior is induced by a prior on $R^2$ and the total prior variance of $\beta$ is decomposed through a Dirichlet prior, therefore we refer to the prior as the $R^2$-induced Dirichlet Decomposition (R2-D2) prior.

**Proposition 1.** If $\omega \mid \xi \sim \text{Ga}(a, \xi)$ and $\xi \sim \text{Ga}(b, 1)$, then $\omega \sim \text{BP}(a, b)$, where $\text{Ga}(\mu, \nu)$ is the Gamma random variable with shape $\mu$ and rate $\nu$.

Hence (3) can also be written as

$$
\beta_j \mid \sigma^2, \phi_j, \omega \sim K(\sigma^2 \phi_j \omega), \; \phi \sim \text{Dir}(a, \cdots, a), \; \omega \mid \xi \sim \text{Ga}(a, \xi), \; \xi \sim \text{Ga}(b, 1).
$$

(3)

As shown in the next section, Proposition 1's representation of a Beta prime variable in terms of two Gamma variables reveals connections among other common shrinkage priors.
Proposition 2. If \( \omega \sim \text{Ga}(a, \xi) \), \( (\phi_1, \cdots, \phi_p) \sim \text{Dir}(a_\pi, \cdots, a_\pi) \), and \( a = pa_\pi \), then \( \phi_j \omega \sim \text{Ga}(a_\pi, \xi) \) independently for \( j = 1, \cdots, p \).

Now, using Proposition 2 reducing to the special case of \( a = pa_\pi \), (3) is equivalent to

\[
\beta_j | \sigma^2, \lambda_j \sim K(\sigma^2 \lambda_j), \quad \lambda_j | \xi \sim \text{Ga}(a_\pi, \xi), \quad \xi \sim \text{Ga}(b, 1),
\]

or by applying Proposition 1 again, it can also be represented as

\[
\beta_j | \sigma^2, \lambda_j \sim K(\sigma^2 \lambda_j), \quad \lambda_j \sim \text{BP}(a_\pi, b).
\]

2.3 Normal kernel

The class of R2-D2 priors relies on the kernel density \( K \). The R2-D2 prior with normal kernel and \( a = pa_\pi \) is

\[
\beta_j | \sigma^2, \lambda_j \sim N(0, \sigma^2 \lambda_j), \quad \lambda_j | \xi \sim \text{Ga}(a_\pi, \xi), \quad \xi \sim \text{Ga}(b, 1).
\]

This is a special case of the general normal-gamma priors as proposed in [Griffin et al. (2010)] by keeping the shape hyperparameter in the Gamma prior for the variance coefficients, \( a_\pi \), fixed, and the rate hyperparameter, \( \xi \), given a particular Gamma hyperprior.

Another equivalent form is

\[
\beta_j | \sigma^2, \lambda_j \sim N(0, \sigma^2 \lambda_j), \quad \lambda_j \sim \text{BP}(a_\pi, b),
\]

and the density of \( \lambda_j^{1/2} \) is

\[
\pi_{\lambda_j^{1/2}}(x) = \frac{2\Gamma(a_\pi + b)}{\Gamma(a_\pi)\Gamma(b)} \frac{x^{2a_\pi - 1}}{(1 + x^2)^{a_\pi + b}}.
\]

When \( a_\pi = a/p = b = 1/2 \), this is the standard half-Cauchy distribution, i.e., \( C^+(0,1) \), then the R2-D2 prior is written as

\[
\beta_j | \sigma^2, \tau_j \sim N(0, \sigma^2 \tau_j^2), \quad \tau_j \sim C^+(0, 1),
\]
which is a special case of the Horseshoe prior proposed in \cite{Carvalho2009} with global shrinkage parameter fixed at 1.

### 2.4 Double-exponential kernel

As shown in Section 2.3, the choice of normal kernel gives the special case of the normal-gamma family and Horseshoe prior. However, the double-exponential distribution has more mass around zero and heavier tails than the normal distribution. Thus, to encourage shrinkage, it is reasonable to replace the normal kernel with a double exponential kernel, i.e., \( \beta_j \mid \sigma^2, \phi_j, \omega \sim \text{DE}(\sigma(\phi_j\omega/2)^{1/2}) \) for \( j = 1, \ldots, p \), with \( \text{DE}(\delta) \) denoting a double-exponential distribution with mean 0 and variance \( 2\delta^2 \). The prior is then summarized as follows:

\[
\beta_j \mid \sigma^2, \phi_j, \omega \sim \text{DE}(\sigma(\phi_j\omega/2)^{1/2}), \quad \phi \sim \text{Dir}(a_{\pi}, \cdots, a_{\pi}), \quad \omega \sim \text{BP}(a, b). \tag{4}
\]

In this global-local shrinkage prior, \( \omega \) controls the global shrinkage degree through \( a \) and \( b \), while \( \phi_j \) controls the local shrinkage through \( a_{\pi} \). In particular, when \( a_{\pi} \) is small, the prior would lead to large variability between the proportions \( \phi_j \)'s, thus more shrinkage for the regression coefficients; while when \( a_{\pi} \) is large, less shrinkage is assumed.

Given \( a = pa_{\pi} \), by Proposition 1 and 2, the R2-D2 prior can also be equivalently written as:

\[
\beta_j \mid \sigma^2, \lambda_j \sim \text{DE}(\sigma(\lambda_j/2)^{1/2}), \quad \lambda_j \mid \xi \sim \text{Ga}(a_{\pi}, \xi), \quad \xi \sim \text{Ga}(b, 1), \tag{5}
\]

or

\[
\beta_j \mid \sigma^2, \lambda_j \sim \text{DE}(\sigma(\lambda_j/2)^{1/2}), \quad \lambda_j \sim \text{BP}(a_{\pi}, b). \tag{6}
\]

We focus on this double-exponential kernel-based prior for the remainder of the paper.
2.5 Posterior computation

For posterior computation, the following equivalent representation is useful. The R2-D2 prior (4) is equivalent to

\[
\beta_j \mid \sigma^2, \psi_j, \phi_j, \omega \sim N(0, \psi_j \phi_j \omega \sigma^2/2), \ \psi_j \sim \text{Exp}(1/2), \\
\phi \sim \text{Dir}(a_{\pi}, \cdots, a_{\pi}), \ \omega \mid \xi \sim \text{Ga}(a, \xi), \ \xi \sim \text{Ga}(b, 1),
\]

(7)

where \(\text{Exp}(\delta)\) denotes the exponential distribution with mean \(\delta^{-1}\). The Gibbs sampling procedure is based on (7) with \(a = pa_{\pi}\). Assume the variance has prior \(\sigma^2 \sim \text{IG}(a_1, b_1)\), an inverse Gamma distribution with shape and scale parameters \(a_1\) and \(b_1\) respectively. The details of Gibbs sampling procedures have been given in Appendix.

3 Theoretical Properties

3.1 Marginal density

In this section, a number of theoretical properties of the proposed R2-D2 prior with the double exponential kernel are established. The properties of the Horseshoe (Carvalho et al., 2009, 2010), Horseshoe+ (Bhadra et al., 2016), and Dirichlet-Laplace prior (Bhattacharya et al., 2015) are provided as a comparison. Proofs and technical details are given in the Appendix. For simplicity of comparison across approaches, the variance term \(\sigma^2\) is fixed at 1. For the R2-D2 prior, the Dirichlet concentration \(a_{\pi}\) is set to \(a = pa_{\pi}\), so we consider the new proposed prior represented as (5) or (6) in this section.

**Proposition 3.** Given the R2-D2 prior (5), the marginal density of \(\beta_j\) for any \(j = 1, \cdots, p\) is

\[
\pi_{R2-D2}(\beta_j) = \frac{1}{(2\pi)^{p/2} \Gamma(a_{\pi}) \Gamma(b)} \, G^{31}_{13} \left( \frac{\beta^2_j}{2} \left| \frac{1}{a_{\pi}} - \frac{1}{2}, 0, \frac{1}{2} \right. \right) = \frac{1}{(2\pi)^{p/2} \Gamma(a_{\pi}) \Gamma(b)} \, G^{13}_{31} \left( \frac{2}{\beta^2_j} \left| \frac{1}{2}, \frac{1}{2}, a_{\pi}, \frac{1}{2} \right. \right),
\]

where \(G^{m,n}_{p,q}(z \mid .)\) denotes the Meijer G-function.
The Horseshoe prior proposed in Carvalho et al. (2009, 2010) is

\[ \beta_j | \lambda_j \sim N(0, \lambda_j^2), \quad \lambda_j | \tau \sim C^+(0, \tau), \]

where \( C^+(0, \tau) \) denotes a half-Cauchy distribution with scale parameter \( \tau \), with density
\[
p(y | \tau) = \frac{2}{\pi \tau (1 + (y/\tau)^2)}.\]
The Horseshoe+ prior proposed in Bhadra et al. (2016) is

\[ \beta_j | \lambda_j \sim N(0, \lambda_j^2), \quad \lambda_j | \tau, \eta_j \sim C^+(0, \tau \eta_j), \quad \eta_j \sim C^+(0, 1). \]

The Dirichlet-Laplace prior proposed in Bhattacharya et al. (2015) is

\[ \beta_j | \psi_j \sim DE(\psi_j), \quad \psi_j \sim Ga(a_D, 1/2). \]

Figure 1 plots the marginal density function of the R2-D2 density along with the Horseshoe, Dirichlet-Laplace, and Cauchy distributions. In the figure, for visual comparison, the hyperparameter in the four priors, i.e., \( \tau \) in Horseshoe and Horseshoe+ prior, \( a_D \) in Dirichlet-Laplace prior, \((a_\pi, b)\) in the R2-D2 prior, are selected to ensure the interquartile range is approximately 1. Note that for comparable purpose, \( a_\pi \) in the new proposed prior is set as \( a_D/2 \), which is half of the hyperparameter in Dirichlet-Laplace prior. Another hyperparameter \( b \) in the new prior is then tuned to ensure the interquartile range is 1. From Figure 1, the R2-D2 prior density shows the most mass around zero and the heaviest tails; we formally investigate these asymptotic properties in the following sections.

### 3.2 Asymptotic tail behaviors

We examine the asymptotic behaviors of tails of the proposed R2-D2 prior in this section. A prior with heavy tails is desirable in high-dimensional regression to allow the posterior to estimate large values for important predictors.

**Theorem 1.** Given \( |\beta| \to \infty \), for any \( a_\pi > 0 \) and \( b > 0 \), the marginal density of the R2-D2 prior \( \pi_{R2-D2}(\beta) \) satisfies

\[
\pi_{R2-D2}(\beta) = O(1/|\beta|^{2b+1}).
\]

Furthermore, when \( 0 < b < 1/2 \),

\[
\lim_{|\beta| \to \infty} \frac{\pi_{R2-D2}(\beta)}{|\beta|^{-2}} = \infty,
\]

i.e., the R2-D2 prior has heavier tails than the Cauchy dis-
As a comparison, we also study the tail behavior of the Dirichlet-Laplace and double Pareto prior. The density of generalized double Pareto prior proposed in Armagan et al. (2013a) is
\[ \pi_{\text{GDP}}(\beta_j | \eta, \alpha) = (1 + |\beta_j|/\eta)^{-(\alpha+1)}/(2\eta/\alpha), \quad (\alpha, \eta > 0). \]

**Theorem 2.** Given $|\beta| \to \infty$, for any $\alpha > 0$, the marginal density of the generalized double Pareto prior satisfies $\pi_{\text{GDP}}(\beta) = O(1/|\beta|^\alpha)$. Furthermore, when $\alpha < 1$, $\lim_{|\beta| \to \infty} \pi_{\text{GDP}}(\beta)/\beta^{-2} = \infty$, i.e., the double Pareto prior has heavier tails than the Cauchy distribution.

**Theorem 3.** Given $|\beta| \to \infty$, for any $a_D > 0$, the marginal density of the Dirichlet-Laplace prior satisfies $\pi_{\text{DL}}(\beta) = O(\log |\beta|/\beta^2)$ as illustrated in the proof of Theorem 4.6 in Bhadra et al. (2016). Therefore, the double Pareto prior and the proposed R2-D2 prior lead to the heaviest tail, followed by Horseshoe+, then Horseshoe, and finally the Dirichlet-Laplace prior. With a polynomial tail heavier than Cauchy distribution, the new proposed prior attains a substantial improvement over a large class of global-local shrinkage priors.

### 3.3 Concentration properties

In this section, we study the concentration properties of the new proposed prior around the origin. The concentration properties of Dirichlet-Laplace, Horseshoe, and Horseshoe+ priors are also given. We prefer priors with high concentration near zero to reflect the prior that most of the covariates do not have a substantial effect on the response. We now show that the proposed R2-D2 prior has higher concentration at zero to go along with heavier tails than other global-local priors.
**Theorem 4.** As $|\beta| \to 0$, if $0 < a_\pi < 1/2$, the marginal density of the R2-D2 prior $\pi_{R2-D2}(\beta)$ satisfies $\pi_{R2-D2}(\beta) = O(1/|\beta|^{1-2a_\pi})$.

**Theorem 5.** As $|\beta| \to 0$, if $0 < a_D < 1$, the marginal density of the Dirichlet-Laplace prior satisfies $\pi_{DL}(\beta) = O(1/|\beta|^{1-a_D})$.

For the Horseshoe prior, as summarized in [Carvalho et al. (2010)](#), the marginal density $\pi_{HS}(\beta) = (2\pi^3)^{-1/2} \exp(\beta^2/2) E_1(\beta^2/2)$, where $E_1(z) = \int_1^\infty e^{-tz}/t \, dt$ is the exponential integral function. As $|\beta| \to 0$,

$$\frac{1}{2(2\pi^3)^{1/2}} \log(1 + 4/\beta^2) \leq \pi_{HS}(\beta) \leq \frac{1}{(2\pi^3)^{1/2}} \log(1 + 2/\beta^2).$$

Therefore around the origin, $\pi_{HS}(\beta) = O(\log(1/|\beta|))$. Also by the proof of Theorem 4.6 in [Bhadra et al. (2016)](#), as $|\beta| \to 0$, the marginal density of Horseshoe+ prior satisfies $\pi_{HS+}(\beta) = O(\log^2(1/|\beta|))$. It is clear that $2a_\pi$ in the R2-D2 prior plays the same role around origin as $a_D$ in the Dirichlet-Laplace prior. Accordingly, when $a_D = 2a_\pi \in (0,1)$, all these four priors possess unbounded density near the origin. However, the R2-D2 prior and Dirichlet-Laplace prior diverge to infinity with a polynomial order, much faster than the Horseshoe+ (with a squared logarithm order) and the Horseshoe prior (with a logarithm order). Although the double Pareto prior also has a polynomial order tail similar as our proposed R2-D2 prior, the double Pareto prior differs around the origin, as it remains bounded, while our new proposed prior is unbounded at the origin.

Now we see that the R2-D2 prior and Dirichlet-Laplace prior put more mass in a small neighborhood of zero compared to the Horseshoe and Horseshoe+ prior. [Polson and Scott (2010)](#) established that when the truth is zero, a prior with unbounded density near zero is super-efficient in terms of the Kullback-Leibler risk. As formalized below, the more mass the prior puts around the neighborhood of the origin, the more efficient. Then the four priors with unbounded density around zero are all super-efficient, with R2-D2 prior and Dirichlet-Laplace more efficient than the Horseshoe+, and followed by Horseshoe. Section 3.5 discusses it in detail.
3.4 Posterior consistency

In this section, we show that the proposed R2-D2 prior yields posterior consistency. Assume the true regression parameter is $\beta^0_n$, and the regression parameter $\beta_n$ is given some shrinkage prior. If the posterior of $\beta_n$ converges in probability towards $\beta^0_n$, i.e., for any $\epsilon > 0$, $\Pr(\|\beta_n - \beta^0_n\| > \epsilon | Y) \to 0$ as $p_n, n \to \infty$, we say the prior yields a consistent posterior.

Assume the following regularity conditions:

(A1) The number of predictors $p_n$ is $o(n)$;

(A2) Let $d_{\min}$ and $d_{\max}$ be the smallest and the largest singular values of $X^TX/n$ respectively. Assume $0 < d_{\min} < \lim_{n \to \infty} d_{\min} \leq \limsup_{n \to \infty} d_1 < d_{\max} < \infty$, where $d_{\min}$ and $d_{\max}$ are fixed and $X = (x_1^T, \ldots, x_n^T)^T$;

(A3) $\limsup_{n \to \infty} \max_{j=1, \ldots, p_n} |\beta_{0j}^0| < \infty$;

(A4) $q_n = o(n/\log n)$, in which $q_n$ is the number of nonzero components in $\beta^0_n$.

**Theorem 6.** Under assumptions (A1)–(A4), for any $b > 0$, given the linear regression model (1), and $a_{\pi} = C/(p_n^{b/2}n^{\rho b/2} \log n)$ for finite $\rho > 0$ and $C > 0$, the R2-D2 prior (5) yields a consistent posterior.

3.5 Predictive efficiency

In this section, we study the predictive efficiency of the shrinkage priors. We focus on the case when the design matrix $X = (x_1^T, \ldots, x_n^T)^T$ is orthogonal, i.e., $X^TX = I_p$. We also assume $\sigma^2$ to be known. Though this may not be a realistic setup in practice, it provides some insight and motivation for measuring the predictive efficiency. In this case, the sufficient statistic for $\beta$ is the ordinary least square estimate, i.e., $\hat{\beta} = X^TY$, with $\hat{\beta} \sim N(\beta, \sigma^2I_p)$. For simplicity of notation, without loss of generality, assume $\beta_0$ is the true parameter, we rewrite the sampling model as

$$y_i \sim N(\beta_0, \sigma^2)$$

(8)
we use the Kullback-Leibler divergence between the true model and the Bayes estimator (or the posterior mean estimator) of the density function to measure the predictive efficiency. Denote $\pi(y \mid \beta)$ as the sampling model (8), and $KL(\pi_1, \pi_2) = E_{\pi_1} \{ \log(\pi_1 / \pi_2) \}$ as the Kullback-Leibler divergence of $\pi_2$ from $\pi_1$. The results are based on the following Lemma.

**Lemma 1.** (Clarke and Barron, 1990) Define $A_\epsilon = \{ \beta : KL(\pi_{\beta_0}, \pi_\beta) \leq \epsilon \}$ as the Kullback-Leibler information set of measure $\epsilon$ centered at $\beta_0$. Let $\mu(d\beta)$ be the prior measure, the observed data $y(n) = (y_1, \cdots, y_n)$, the corresponding posterior distribution is $\mu_n(d\beta \mid y(n))$, and posterior predictive density $\hat{\pi}_n(y) = \int \pi(y \mid \beta) \mu_n(d\beta \mid y(n))$. Assume that $\mu(A_\epsilon) > 0$ for any $\epsilon > 0$, then at $\pi_{\beta_0}$, the prior $\mu(d\beta)$ is information dense. Then the Cesàro-average risk of the Bayes estimator $\hat{\pi}_n$, defined as $R_n = n^{-1} \sum_{j=1}^{n} KL(\pi_{\beta_0}, \hat{\pi}_j)$, satisfies $R_n \leq \epsilon - \log \mu(A_\epsilon) / n$.

Carvalho et al. (2010) proved that when the true parameter $\beta_0$ is zero, the upper bound of Cesàro-average risk of the maximum likelihood estimator is $R_n = O(n^{-1} \log n)$; while the Horseshoe estimator’s risk satisfies

$$R_n(\text{HS}) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - \log \log n + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{\log n} \right) \right).$$

In this sense, the Horseshoe estimator for the sampling density converges to the true model at a super-efficient rate. Bhadra et al. (2016) shows that the Horseshoe+ estimator slightly improves the rate with

$$R_n(\text{HS+}) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - 2 \log \log n + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{(\log n)^2} \right) \right).$$

In this section, we illustrate that our proposed new prior achieves a smaller risk. Our result is based on the following theorem.

**Theorem 7.** For $0 < a_\pi < 1/2$, when $\beta_0 = 0$, the Cesàro-average risk of the Bayes estimator under the $R2-D2$ prior (5) satisfies

$$R_n(\text{R2-D2}) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - \left( \frac{1}{2} - a_\pi \right) \log n + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{n^{1/2-a_\pi}} \right) \right).$$
As a complement, we also give the upper risk bounds for the DL prior:

**Theorem 8.** For $0 < a_D < 1$, when $\beta_0 = 0$, the Cesàro-average risk of the Bayes estimator under the Dirichlet-Laplace prior satisfies

$$R_n(DL) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - \left( \frac{1}{2} - \frac{a_D}{2} \right) \log n + O(1) \right) = O\left( \frac{1}{n \log \left( \frac{n}{n^{1/2-a_D/2}} \right)} \right).$$

Therefore the R2-D2 prior and Dirichlet-Laplace priors have smaller Kullback-Leibler risk bound than the Horseshoe and Horseshoe+ priors. Combining Theorems 1 and 4, our proposed prior achieves desirable behavior both around the origin and in the tails, as well as improved performance in prediction in the orthogonal case by Theorem 7. Table 1 provides the summary results of the above properties.

### 4 Simulation Study

To illustrate the performance of the proposed new prior, we conduct a simulation study with various number of predictors and effect size. In each setting, 200 datasets are simulated from the linear model (1) with $\sigma^2 = 1$, sample size $n$ fixed at 60 to match the real data example in Section 5, and the number of predictors $p$ varying in $p \in \{50, 100, 500\}$. The covariates $x_i$, $i = 1, \cdots, n$, are generated from multivariate normal distribution with mean zero, and correlation matrix of autoregressive (1) structure with correlation $\rho = 0.5$ or 0.9. For the regression coefficients $\beta$, we consider the following two setups.

**Setup 1:** $\beta = (0^T_{10}, B_1^T, 0^T_{30}, B_2^T, 0^T_{p-50})^T$ with $0_k$ representing the zero vector of length $k$, and $B_1$ and $B_2$ each of length 5 nonzero elements. The fractions of true coefficients with exactly zero values are 80%, 90% and 98% for $p \in \{50, 100, 500\}$, respectively, and the remaining 20%, 10% and 2% nonzero elements $B_1$ and $B_2$ were independently generated from a Student $t$ distribution with 3 degrees of freedom to give heavy tails. In this case, the theoretical $R^2$ as in equation (2) is 0.97.

**Setup 2:** $\beta = (0^T_{10}, s^* B_3^T, 0^T_{p-15})^T$ with $B_3$ also of length 5 with elements generated independently from the Student $t$ distribution with 3 degrees of freedom, with $s^* = 15^{-1/2}$ to ensure the total prior variance of $\beta$ is 1 and hence the theoretical $R^2$ as in equation (2) is
is 0.5.

We consider $p = 50, 100, 500$ for setup 1, and only $p = 100$ for setup 2 because other cases perform similarly. Setup 2 is designed to study the performance of the proposed R2-D2 prior with known $R^2$ information. For each simulated dataset, we use different shrinkage priors for $\beta$. The priors are Horseshoe, Horseshoe+, R2-D2$_{(0.5,0.5)}$ with $a = 0.5$, $b = 0.5$, $a_\pi = 1/(2p)$, R2-D2$_{(p/n,0.5)}$ with $a = p/n$, $b = 0.5$, $a_\pi = 1/n$, R2-D2$_{(p/n,0.1)}$ with $a = p/n$, $b = 0.1$, $a_\pi = 1/n$, R2-D2$_{(1,1)}$ with $a = 1$, $b = 1$, $a_\pi = 1/p$, DL$_{1/p}$ with $a_D = 1/p$, DL$_{2/n}$ with $a_D = 2/n$, and DL$_{1/n}$ with $a_D = 1/n$. For the Horseshoe and Horseshoe+, Markov chain Monte Carlo steps are implemented through Stan in R using the code provided by the author of Bhadra et al. (2016). For the R2-D2 and Dirichlet-Laplace, Gibbs samplers are implemented in R. 10,000 samples are collected with the first 5,000 samples discarded as burn-in.

The average value of the sum squared error corresponding to the posterior mean across the 200 replicates is provided in Table 2 with $\rho = 0.5$. Simulation setup 1 with $\rho = 0.9$ results are given in Table 3. Table 4 provides the results for simulation setup 2. Tables 2 and 4 give the total sum squared error as well as the sum squared error split into three pieces according to the value of the true $\beta$ at $\beta_j = 0$, $|\beta_j| \in (0,0.5]$, and $|\beta_j| > 0.5$, $j = 1, \cdots, p$. The averaged area under the Receiver-Operating Characteristic curve based on the posterior $t$-statistic, i.e., the ratio of the posterior mean and posterior standard deviation, is also given to offer further evaluation of the variable selection performance. We measure the reliability of the ordering of the magnitude of the posterior $t$-statistic through the area under the Receiver-Operating Characteristic curve, which is labeled as “ROC” in the tables.

Overall, the new proposed prior with $a = p/n$ or $a_\pi = 1/n$ has similar sum of squared error to the Horseshoe and Horseshoe+ prior, and smaller than the Dirichlet-Laplace prior. Horseshoe and Horseshoe+ yield good estimators with small sum of squared error. However, Horseshoe and Horseshoe+ generally have lower Receiver-Operating Characteristic area, worse than the R2-D2 prior and Dirichlet-Laplace priors. This may be explained by investigating the sum of squared error for zero, small coefficients and large coefficients. Although
Horseshoe and Horseshoe+ estimate the nonzero coefficients quite well, they estimate the zero coefficients poorly, which leads to more false positives and poor Receiver-Operating Characteristic performance. This corresponds with the poor concentration properties of the Horseshoe and Horseshoe+ as discussed in Section 3.3. In addition, we also conduct simulations for fixed $p$ with varying $n$, and the performance is similar.

Furthermore, the Dirichlet-Laplace priors exhibit excellent performance in estimating the zero coefficients, but poor estimates of large coefficients, leading to large total sum of squared error. This is due to the good concentration at the origin (Section 3.3) but light tails of the Dirichlet-Laplace priors (Section 3.2). However, inaccurate estimation at large coefficients does not greatly affect the Receiver-Operating Characteristic performance, which is comparable to the R2-D2 priors. For the R2-D2 prior, the value of $b$ slightly affects the estimation. By analogy of the R2-D2 prior with $a = p/n, b = 0.5$ and $a = p/n, b = 0.1$, we gain a key insight that a smaller value of $b$ results in slightly smaller total sum of squared error due to the better estimation at large coefficients. This coincides with the fact that $b$ controls the tail behavior (Section 3.2), with smaller $b$ giving heavier tails. The results show that $a = 1$ or $a_\pi = 1/p$ and $a = 0.5$ or $a_\pi = 1/(2p)$ lead to smaller sum of squared error at zero coefficients than $a = p/n$ or $a_\pi = 1/n$ when $p > n$, which again, matches the concentration properties of R2-D2 prior as described in Section 3.3. There is no significant difference in variable selection performance for the four parametrizations in “R2-D2” prior. In all, the R2-D2 prior with $a_\pi = 1/n$ and $b < 0.5$ achieves success in both estimation and variable selection, demonstrating distinguishable performance from the Horseshoe, Horseshoe+ and Dirichlet-Laplace priors.

5 PCR Data Analysis

We now analyze the mouse gene expression data collected by Lan et al. (2006), consisting of the expression levels of 22,575 genes for $n = 60$ (31 female and 29 male) mice. Real-time PCR was used to measure some physiological phenotypes, including number of phosphoenolpyruvate carboxykinase (PEPCK), glycerol-3-phosphate acyltransferase (GPAT), and stearoyl-CoA desaturase 1 (SCD1). We build regression models for the three phe-
notypes using gender and genetic covariates as predictors. The data can be found at http://www.ncbi.nlm.nih.gov/geo; accession number GSE3330.

To evaluate the performance of different priors, the 60 observations are randomly split into a training set of size 55 and testing set of size 5. Then in each split, the 22,575 genes were first screened down to 999 genes based on the ordering of the magnitude of the marginal correlation with the response only on the training data set with sample size 55. Then for each of the 3 regressions, the data set contains $p = 1,000$ predictors (999 genes and gender) and $n = 55$ observations. After screening, we performed linear regression using each of the global-local shrinkage prior. The convergence diagnostics plots of R2-D2 prior are shown in Figure 2 and 3. Table 5 gives the mean squared prediction error based on 100 random splits of the data. The posterior mean of the regression coefficients from the training set served as the estimate of $\beta$ to make prediction for the testing set. Overall, the results agree with the simulation studies. Our proposed R2-D2 prior performs better than other priors for this data set, and changing the value of $b$ has little effect on the results.

In addition, we also compared the agreement between methods in terms of variable selection. For each regression, we apply different shrinkage priors on the full data set, then posterior samples of $\beta$ are collected. For each $\beta_j$ ($j = 1, \cdots, p$), the posterior $t$-statistic is calculated by dividing the mean with the standard deviation of those posterior samples. The predictors are ordered by the magnitude of the posterior $t$-statistics from the largest to the smallest. Ideally, the important predictors will be in the beginning of the ordering. Figure 4 plots the agreement of the orders between various priors when fit to the full data set for PEPCK. The figures for SCD1 and GPAT are similar. Again, it shows that changing the value of $b$ does not result in too much variation of the agreement. In general, the difference of the agreement with different hyperparameter values in the R2-D2 priors is smaller than that of Dirichlet-Laplace prior, and the difference between Horseshoe and Horseshoe+ prior. For this data set, our proposed R2-D2 prior appears less sensitive to different hyperparameter values than the Dirichlet-Laplace prior.
References

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In Advances in neural information processing systems, pages 523–531.

Armagan, A., Dunson, D. B., and Lee, J. (2013a). Generalized double Pareto shrinkage. Statistica Sinica, 23(1):119.

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013b). Posterior consistency in linear models under shrinkage priors. Biometrika, 100(4):1011–1018.

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. (2016). The horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. Journal of the American Statistical Association, 110(512):1479–1490.

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. Journal of the American Statistical Association, 107(500):1610–1624.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In International Conference on Artificial Intelligence and Statistics, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The Horseshoe estimator for sparse signals. Biometrika, page asq017.

Clarke, B. S. and Barron, A. R. (1990). Information-theoretic asymptotics of Bayes methods. Information Theory, IEEE Transactions on, 36(3):453–471.

DLMF (2015). NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.10 of 2015-08-07. Online companion to [Olver et al. (2010)].

Fields, J. L. (1972). The asymptotic expansion of the Meijer G-function. Mathematics of Computation, pages 757–765.
George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

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

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

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, pages 730–773.

Johnson, N., Kotz, S., and Balakrishnan, N. (1995). Continuous univariate distributions, volume 2. John Wiley & Sons. Inc., 75.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E., Flowers, M. T., Schueler, K. L., Manly, K. F., et al. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet*, 2(1):e6.

Miller, P. D. (2006). *Applied asymptotic analysis*, volume 75. American Mathematical Soc.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Narisetty, N. N., He, X., et al. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W., editors (2010). *NIST Handbook of Mathematical Functions*. Cambridge University Press, New York, NY. Print companion to [DLMF](https://dlmf.nist.gov/) (2015).

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538.
Ročková, V. and George, E. I. (2016). The spike-and-slab lasso. *Journal of the American Statistical Association*, (just-accepted).

Seshadri, V. (1997). Halphen’s laws. *Encyclopedia of statistical sciences*.

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

Zhou, M. and Carin, L. (2015). Negative binomial process count and mixture modeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):307–320.

Zwillinger, D. (2014). *Table of integrals, series, and products*. Elsevier.
A Appendix

A.1 Technical details

Proof of Proposition 1. The proposition follows from

\[
\pi(\omega) = \int_0^\infty \pi(\omega | \xi) \pi(\xi) d\xi = \int_0^\infty \frac{\xi^a}{\Gamma(a)} \omega^{a-1} e^{-\xi \omega} \frac{1}{\Gamma(b)} \xi^{b-1} e^{-\xi} d\xi
\]

\[
= \frac{1}{\Gamma(a)\Gamma(b)} \omega^{a-1} \int_0^\infty \xi^{a+b-1} e^{-(1+\omega)\xi} d\xi
\]

\[
= \frac{1}{\Gamma(a)\Gamma(b)} \omega^{a-1} \frac{\Gamma(a+b)}{(1+\omega)^{a+b}}
\]

\[
= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \omega^a (\omega > 0).
\]

\[
\]
\[ \pi_{\text{R2-D2}}(\beta) = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \int_0^\infty \exp(-\mu x) x^{2\nu-1} (x^2 + u^2)^{\rho-1} \, dx \]
\[ = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \frac{u^{2\nu-2\rho-2}}{2\pi^{1/2}(1 - \rho)} G^{31}_{13} \left( \frac{\mu^2 u^2}{4} \left| 1 - \rho - \nu, 0, \frac{1}{2} \right. \right) (3.389.2 \text{ in Zwillinger (2014)}) \]
\[ = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \frac{2^{1/2-a_\pi}}{2\pi^{1/2}\Gamma(a_\pi + b)} G^{31}_{13} \left( \frac{\beta^2}{2} \left| \frac{1}{2} - b, a_\pi - \frac{1}{2}, 0, \frac{1}{2} \right. \right) \]
\[ = \frac{1}{(2\pi)^{1/2}\Gamma(a_\pi) \Gamma(b)} G^{31}_{13} \left( \frac{\beta^2}{2} \left| \frac{1}{2} - a_\pi, 0, \frac{1}{2} \right. \right) \]

where \( G(.) \) denotes the Meijer G-Function, and the last equality follows from 16.19.1 in DLMF (2015). Proposition 3 follows.

\[ \square \]

**Proof of Theorem 7** For the proof of Theorem 7, we will use the following lemma found in Miller (2006).

**Lemma 2. (Watson’s Lemma)** Suppose \( F(s) = \int_0^\infty e^{-st} f(t) \, dt, f(t) = t^\alpha g(t) \) where \( g(t) \) has an infinite number of derivatives in the neighborhood of \( t = 0 \), with \( g(0) \neq 0 \), and \( \alpha > -1 \). Suppose \(|f(t)| < Ke^{ct}\) for any \( t \in (0, \infty) \), where \( K \) and \( c \) are independent of \( t \). Then, for \( s > 0 \) and \( s \to \infty \),

\[ F(s) = \sum_{k=0}^n \frac{g^{(k)}(0)}{k!} \frac{\Gamma(\alpha + k + 1)}{s^{\alpha+k+1}} + O\left(\frac{1}{s^{\alpha+n+2}}\right). \]

According to equation (9) in the proof of Proposition 3, we have

\[ \pi_{\text{R2-D2}}(\beta) = \frac{2^a \Gamma(a_\pi + b)}{\Gamma(a_\pi) \Gamma(b)} \int_0^\infty \exp(-|\beta| x) \frac{x^{2b}}{(x^2 + 2)^{a_\pi+b}} \, dx = \int_0^\infty e^{-|\beta|x} f(x) \, dx \equiv F(|\beta|), \]

where \( f(t) = C^* t^{2b} / (t^2 + 2)^{a_\pi+b} \equiv t^{2b} g(t) \), \( C^* = 2^a \Gamma(a_\pi + b) / \{\Gamma(a_\pi) \Gamma(b)\} \), and \( g(t) = C^* (t^2 + 2)^{-a_\pi-b} \) with \( g(t) \) has an infinite number of derivatives in the neighborhood of \( t = 0 \), with \( g(0) \neq 0 \). So the marginal density of R2-D2 prior is the Laplace transforms of \( f(.) \). By Watson’s Lemma, since \(|f(t)| < Ke^{ct}\) for any \( t \in (0, \infty) \), where \( K \) and \( c \) are
independent of \( t \), then as \( |\beta| \to \infty \),

\[
F(|\beta|) = \sum_{k=0}^{n} \frac{g^{(k)}(0) \Gamma(2b + k + 1)}{|\beta|^{2b+k+1}} + O\left( \frac{1}{|\beta|^{2b+n+2}} \right),
\]

and setting \( n = 2 \) gives

\[
F(|\beta|) = C^* \left\{ 2^{-a_{\pi}-b} \frac{\Gamma(2b + 1)}{|\beta|^{2b+1}} + 0 \frac{\Gamma(2b + 2)}{|\beta|^{2b+2}} + (-a_{\pi} - b) 2^{-a_{\pi}-b} \frac{\Gamma(2b + 3)}{|\beta|^{2b+3}} \right\} + O\left( \frac{1}{|\beta|^{2b+4}} \right)
\]

\[
= C^* 2^{-a_{\pi}-b} \left\{ \frac{\Gamma(2b + 1)}{|\beta|^{2b+1}} - (a_{\pi} + b) \frac{\Gamma(2b + 3)}{|\beta|^{2b+3}} \right\} + O\left( \frac{1}{|\beta|^{2b+4}} \right)
\]

\[
= O\left( \frac{1}{|\beta|^{2b+4}} \right).
\]

Hence, when \( b < 1/2 \), as \( |\beta| \to \infty \), we have

\[
\frac{\pi_{R2D2}(\beta)}{\beta^2} = C^* 2^{-a_{\pi}-b} \left\{ \frac{\Gamma(2b + 1)}{|\beta|^{2b}} - (a_{\pi} + b) \frac{\Gamma(2b + 3)}{|\beta|^{2b+1}} \right\} \to \infty.
\]

**Proof of Theorem 2** It is obvious based on the marginal density of the generalized double Pareto prior.

**Proof of Theorem 3** According to 10.25.3 in [DLMF (2015)](https://dlmf.nist.gov/10.25), when both \( \nu \) and \( z \) are real, if \( z \to \infty \), then \( K_{\nu}(z) \approx \pi^{1/2}(2z)^{-1/2}e^{-z} \). Then as \( |\beta| \to \infty \), the marginal density of the Dirichlet-Laplace prior given in [Bhattacharya et al. (2015)](https://www.researchgate.net/publication/281407500) satisfies

\[
\pi_{DL}(\beta) = \frac{1}{2^{(1+a_D)/2}\Gamma(a_D)} |\beta|^{a_D-1/2}K_{1-a_D}((2|\beta|)^{1/2})
\]

\[
= \frac{1}{2^{(1+a_D)/2}\Gamma(a_D)} |\beta|^{a_D-1/2}2^{1/2}2^{-3/4}|\beta|^{-1/4} \exp\{-2|\beta|\}^{1/2}
\]

\[
= C_0 |\beta|^{a_D/2-3/4} \exp\{-2|\beta|\}^{1/2} = O\left( \frac{|\beta|^{a_D/2-3/4}}{\exp\{(2|\beta|)^{1/2}\}} \right),
\]

where \( C_0 = \pi^{1/2}2^{-3/4}/\{2^{(1+a_D)/2}\Gamma(a_D)\} \) is a constant value. Furthermore, as \( |\beta| \to \infty \),

\[
\frac{\pi_{DL}(\beta)}{1/\beta^2} \approx C_0 |\beta|^{a_D/2+5/4} \exp\{-2|\beta|\}^{1/2} \to 0.
\]
Proof of Theorem 4. For the proof of Theorem 4 we use the following lemma from Fields (1972). Some useful notations used in the below proof: Denote \( a_P = (a_1, \ldots, a_p) \), as a vector, similarly, \( b_Q = (b_1, \ldots, b_q) \), \( c_M = (c_1, \ldots, c_m) \), and so on. Let \( \Gamma_n(c_P - t) = \prod_{k=1}^n \Gamma(c_k - t) \), with \( \Gamma_n(c_P - t) = 1 \) when \( n = p \), \( \Gamma(c_M - t) = \prod_{k=1}^m \Gamma(c_k - t) \), \( \Gamma^*(a_i - a_N) = \prod_{k=1; k \neq i}^n \Gamma(a_i - a_k) \), and

\[
pFq \left( \frac{a_P}{b_Q} \mid w \right) = \sum_{k=0}^{\infty} \frac{\Gamma(a_P + k) \Gamma(b_Q) w^k}{\Gamma(b_Q + k) \Gamma(a_P) k!} = \sum_{k=0}^{\infty} \frac{\prod_{j=1}^p \Gamma(a_j + k) \prod_{j=1}^q \Gamma(b_j) w^k}{\prod_{j=1}^q \Gamma(b_j + k) \prod_{j=1}^p \Gamma(a_j) k!}.
\]

Lemma 3. (Theorem 1 in Fields (1972)) Given (i) \( 0 \leq m \leq q, 0 \leq n \leq p \); (ii) \( a_i - b_k \) is not a positive integer for \( j = 1, \ldots, p \) and \( k = 1, \ldots, q \); (iii) \( a_i - a_k \) is not an integer for \( i, k = 1, \ldots, p \), and \( i \neq k \); and (iv) \( q < p \) or \( q = p \) and \( |z| > 1 \), we have

\[
C_{p,q}^{m,n} \left( z \mid a_1, \ldots, a_p \right) = \sum_{i=1}^n \frac{\Gamma^*(a_i - a_N) \Gamma(1 + b_M - a_i) \Gamma_n(1 + a_P - a_i) \Gamma_m(a_i - b_Q) z^{-1+a_i}}{\Gamma(a_i - a) \Gamma(1+a_P-a_i) \Gamma_m(a_i - b_Q)} \left( 1+q-m-n \right) F_p^{m,n} \left( \frac{1}{1+a_P-a_i} \mid \frac{-1}{z} \right).
\]

Now to prove Theorem 4. we have from Proposition 3 that, the marginal density of the R2-D2 prior has \( \pi_{R2-D2}(\beta_j) = (2\pi)^{-1/2} \left\{ \Gamma(a_\pi) \Gamma(b) \right\}^{-1} C_{p,q}^{m,n} \left( z \mid \cdot \right) \) with \( m = 1, n = 3, p = 3, q = 1, a_1 = 3/2 - a_\pi, a_2 = 1, a_3 = 1/2, b_1 = 1/2 + b, \) and \( z = 2/\beta^2 \). Conditions (i)-(iv) in Lemma 3 are satisfied for \(|\beta| \) near 0, since \( 0 < a_\pi < 1/2 \). Denote

\[
C_1^* = (2\pi)^{-1/2} \left( \Gamma(a_\pi) \Gamma(b) \right)^{-1} \left( 1 - a_\pi \right) \Gamma(1/2) \Gamma(1/2) > 0
\]

\[
C_2^* = (2\pi)^{-1/2} \left( \Gamma(a_\pi) \Gamma(b) \right)^{-1} \left( a_\pi - 1/2 \right) \Gamma(1/2) \Gamma(1/2) \Gamma(3/2 - a_\pi) < 0
\]

\[
C_3^* = (2\pi)^{-1/2} \left( \Gamma(a_\pi) \Gamma(b) \right)^{-1} \left( a_\pi - 1 \right) \Gamma(-1/2) \Gamma(3/2) \Gamma(2 - a_\pi) > 0
\]

25
\[ U_1(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(a_{\pi} + b + k)}{\Gamma(\frac{1}{2} + a_{\pi} + k)} \frac{(-1)^k (\beta^2)^{k+a_{\pi}-1/2}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_1(k, \beta^2) \]

\[ U_2(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(1/2 + b + k)}{\Gamma(\frac{3}{2} - a_{\pi} + k)} \frac{(-\beta^2)^{k}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_2(k, \beta^2) \]

\[ U_3(\beta^2) = \sum_{k=0}^{\infty} \frac{\Gamma(1 + b + k)}{\Gamma(2 - a_{\pi} + k)} \frac{(-1)^k (\beta^2)^{k+1/2}}{k!} \equiv \sum_{k=0}^{\infty} (-1)^k u_3(k, \beta^2), \]

then

\[ \pi_{R2,D2}(\beta) = \frac{1}{(2\pi)^{1/2} \Gamma(a_{\pi}) \Gamma(b)} C_{\beta}^2 \left( \frac{2}{\beta^2} \right)^{1/2} \frac{\beta^2}{\pi} \]

\[ = \frac{1}{(2\pi)^{1/2} \Gamma(a_{\pi}) \Gamma(b)} \sum_{i=1}^{3} \left( \sum_{k=1}^{3} \frac{\prod_{k=1}^{3} \Gamma(1+b_{\pi}-a_{\pi})}{\prod_{k=1}^{3} \Gamma(1+a_{\pi}-b_{\pi})} \frac{2}{\beta^2} \right) \frac{\Gamma(1+b_{\pi}-a_{\pi})}{\Gamma(1+b_{\pi})} \frac{1}{\Gamma(1+a_{\pi})} \frac{1}{\Gamma(1+b_{\pi}-a_{\pi})} \frac{(-\beta^2)^{k}}{k!} \]

\[ = \frac{1}{(2\pi)^{1/2} \Gamma(a_{\pi}) \Gamma(b)} \left\{ \left( \frac{1}{2} - a_{\pi} \right) \Gamma(1-a_{\pi}) \Gamma(a_{\pi}+b) \left( \frac{2}{\beta^2} \right)^{1/2} \sum_{k=0}^{\infty} \frac{\Gamma(1+k) \Gamma(1+a_{\pi}+b) \Gamma(1+a_{\pi}) \Gamma(1+b) \Gamma(1+b-a_{\pi}) \Gamma(1+b)}{\Gamma(1+a_{\pi}+b) \Gamma(1+b-a_{\pi}) \Gamma(1+b)} \frac{(-\beta^2)^{k}}{k!} \right\} \]

\[ + \Gamma(a_{\pi} - \frac{1}{2}) \Gamma(1-a_{\pi}) \Gamma(1+b) \left( \frac{2}{\beta^2} \right)^{1/2} \sum_{k=0}^{\infty} \frac{\Gamma(1+k) \Gamma(1+b-a_{\pi}) \Gamma(1+b)}{\Gamma(1+a_{\pi}+b) \Gamma(1+b-a_{\pi}) \Gamma(1+b)} \frac{(-\beta^2)^{k}}{k!} \]

\[ = \frac{1}{(2\pi)^{1/2} \Gamma(a_{\pi}) \Gamma(b)} \left\{ \left( \frac{1}{2} - a_{\pi} \right) \Gamma(1-a_{\pi}) \Gamma(a_{\pi}+b) \left( \frac{2}{\beta^2} \right)^{1/2} \sum_{k=0}^{\infty} \frac{\Gamma(1+k) \Gamma(1+a_{\pi}+b) \Gamma(1+a_{\pi}) \Gamma(1+b) \Gamma(1+b-a_{\pi}) \Gamma(1+b)}{\Gamma(1+a_{\pi}+b) \Gamma(1+b-a_{\pi}) \Gamma(1+b)} \frac{(-\beta^2)^{k}}{k!} \right\} \]

For fixed \( \beta \) near the neighborhood of zero, \( u_1(k, \beta^2) \), \( u_2(k, \beta^2) \), and \( u_3(k, \beta^2) \) are all monotone decreasing, and converge to zero as \( k \to \infty \). Thus, by alternating series test, \( U_1(\beta^2) \),
$U_2(\beta^2)$, and $U_3(\beta^2)$ all converge. Also, we have

$$C_0|\beta|^{2a\pi-1} - C_1|\beta|^{2a\pi+1} = u_1(0, \beta^2) - u_1(1, \beta^2) \leq U_1(\beta^2) \leq u_1(0, \beta^2) = C_0|\beta|^{2a\pi-1}$$

$$C_2 - C_3|\beta|^2 = u_2(0, \beta^2) - u_2(1, \beta^2) \leq U_2(\beta^2) \leq u_2(0, \beta^2) = C_2$$

$$C_4|\beta| - C_5|\beta|^3 = u_3(0, \beta^2) - u_3(1, \beta^2) \leq U_3(\beta^2) \leq u_3(0, \beta^2) = C_4|\beta|,$$

where $C_0$, $C_1$, $C_2$, $C_3$, and $C_4$ are all positive constants. So given that $|\beta|$ in the neighborhood of zero and $a\pi \in (0, \frac{1}{2})$,

$$C_1^*(C_0|\beta|^{2a\pi-1} - C_1|\beta|^{2a\pi+1}) + C_2^*C_2 + C_3^*(C_4|\beta| - C_5|\beta|^3) \leq \pi_{R2-D2}(\beta) \leq C_1^*C_0|\beta|^{2a\pi-1} + C_2^*(C_2-C_3|\beta|^2) + C_3^*C_4|\beta|,$$

then $\pi_{R2-D2}(\beta) = O(|\beta|^{2a\pi-1}).$}

\textit{Proof of Theorem 5.} According to 10.30.2 in DLMF (2015), when $\nu > 0$, $z \to 0$ and $z$ is real, $\bar{K}_{\nu}(z) \approx \Gamma(\nu)(z/2)^{-\nu/2}$. So given $0 < a_D < 1$ and $|\beta| \to 0$,

$$\pi_{DL}(\beta) = \frac{|\beta|^{(a_D-1)/2}K_{1-a_D}((2|\beta|)^{1/2})}{2^{1+aD}/2\Gamma(a_D)} \approx \frac{|\beta|^{(a_D-1)/2}1/2\Gamma(1 - a_D)((2|\beta|^{1/2})^{a_D-1}}{2^{1+aD}/2\Gamma(a_D)} = C|\beta|^{a_D-1},$$

where $C = \Gamma(1 - a_D)/2^{1+aD}\Gamma(a_D)$ is a constant value. Theorem 5 follows then.

\textit{Proof of Theorem 6.} Denote the estimated set of non-zero coefficients is $A_n = \{j : \beta_{nj} \neq 0, j = 1, \cdots, p_n\}$. Also $\sigma^2$ is fixed at 1. Given the $R^2$-induced Dirichlet decomposition
prior, the probability assigned to the region \((\mathbf{\beta}_n : ||\mathbf{\beta}_n - \mathbf{\beta}_n^0|| < t_n)\) is

\[
\Pr(\mathbf{\beta}_n : ||\mathbf{\beta}_n - \mathbf{\beta}_n^0|| < t_n) = \Pr\left\{ \mathbf{\beta}_n : \sum_{j \in A_n} (\mathbf{\beta}_n - \mathbf{\beta}_n^0)_j^2 + \sum_{j \notin A_n} \beta_n^2 < t_n^2 \right\}
\]

\[
\geq \prod_{j \in A_n} \left\{ \Pr\left( \mathbf{\beta}_n^0 - \frac{t_n}{p_n^{1/2}} < \beta_n^0 < \mathbf{\beta}_n^0 + \frac{t_n}{p_n^{1/2}} \right) \right\} \times \left[ 1 - \left\{ \Pr\left( \mathbf{\beta}_n^0 : \beta_n^0 \geq \frac{t_n^2}{p_n} \right) \right\} \right]^{p_n - q_n}
\]

\[
\geq \prod_{j \in A_n} \left\{ \Pr\left( -\sup_{j \in A_n} |\mathbf{\beta}_n^0| - \frac{t_n}{p_n^{1/2}} < \beta_n^0 < \sup_{j \in A_n} |\mathbf{\beta}_n^0| + \frac{t_n}{p_n^{1/2}} \right) \right\} \times \left[ 1 - \left\{ \Pr\left( \mathbf{\beta}_n^0 : |\mathbf{\beta}_n^0| \geq \frac{t_n^2}{p_n} \right) \right\} \right]^{p_n - q_n}
\]

\[
\geq \left\{ \frac{2}{p_n^{1/2}} \pi \left( \sup_{j \in A_n} |\mathbf{\beta}_n^0| + \frac{t_n}{p_n^{1/2}} \right)^{q_n} \right\} \times \left[ 1 - \left\{ \frac{b^{1/2} E(|\mathbf{\beta}_n^0|)}{t_n} \right\}^{p_n - q_n} \right],
\]

where \(\pi\) is the marginal density function of \(\beta_j\), symmetric and decreasing when the support is positive, and the last but one \(\geq\) is directly got from Markov’s inequality.

Also, based on prior \((5)\), for any \(b > 0\), conditional expectations give

\[
E(|\beta_j|^b) = E\left[ E\{E(|\beta_j|^b \mid \lambda_j) \mid \xi \} \right] = E_\xi \left[ E_{\lambda_j \mid \xi} \left\{ \frac{\Gamma(b + 1)}{(2/\lambda_j)^{b/2}} \mid \xi \right\} \right] = \frac{b^{1/2} \Gamma\left( \frac{b}{2} \right) \Gamma(b + \frac{b}{2})}{2^{b/2} \Gamma(a)}.
\]

Now assume \(t_n = \Delta/n^{3/2}\) and assumptions \([A1] - [A4]\) are satisfied. Then since \(\lim \sup_{j=1, \ldots, n} |\mathbf{\beta}_n^0| < \infty\), there exists a sequence \(k_n = o(n)\) such that \(\sup_{j=1, \ldots, n} |\mathbf{\beta}_n^0| < k_n\) and \(k_n \to \infty\) as \(n \to \infty\). For the R2-D2 prior, based on equation (9), the marginal density is a decreasing function on the positive supports. Then together with the tail approximation of the marginal density as in the proof of Theorem \([1]\) i.e., equation (9), we have

\[
\pi\left( \sup_{j \in A_n} |\mathbf{\beta}_n^0| + \frac{t_n}{p_n^{1/2}} \right) \geq \pi(k_n + \frac{t_n}{p_n^{1/2}}) \geq \frac{\Gamma(a + b)}{\Gamma(a) \Gamma(b)} \left( k_n + \frac{\Delta}{n^{3/2}/p_n^{1/2}} \right)^{2b+1}.
\]

Considering the fact that \(\Gamma(a) = a^{-1} - \gamma_0 + O(a)\) for \(a\) close to zero with \(\gamma_0\) the Euler-Mascheroni constant (see http://functions.wolfram.com/GammaBetaErf/Gamma/06/ShowAll.html),
now we have

\[ \operatorname{pr}(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) \]

\[
\geq \left\{ \frac{2\Delta}{n^{\rho/2} p_n^{1/2}} \frac{\Gamma(a_x + b)}{\Gamma(b)} \left( \frac{\Gamma(2b + 1)}{k_n + \frac{\Delta}{n^{\rho/2} p_n^{1/2}}} \right)^{2b+1} \right\}^{q_n} \times \left[ 1 - \left\{ \frac{p_n^{b/2} n^{\rho b/2} b b/2 \Gamma(b) \Gamma(a_x + b/2)}{\Delta^{b/2} \Gamma(a_x)} \right\}^{p_n - q_n} \right] 
\]

Taking the negative logarithm of both sides of the above formula, and letting \( a_x = C/(p_n^{b/2} n^{\rho b/2} \log n) \), we have

\[
- \log \operatorname{pr}(\beta_n : ||\beta_n - \beta_n^0|| < \frac{\Delta}{n^{\rho/2}}) \leq -q_n \log \left\{ \frac{2\Delta CT \Gamma(a_x + b) 2^{-b} \Gamma(2b + 1)}{n^{\rho/2} p_n^{1/2} p_n^{b/2} n^{\rho b/2} \log n \Gamma(b)} \right\} + q_n (2b + 1) \log \left( k_n + \frac{\Delta}{n^{\rho/2} p_n^{1/2}} \right) 
\]

\[
- q_n \log \left[ 1 - \left\{ \frac{p_n^{b/2} n^{\rho b/2} b \Gamma(b) \Gamma(a_x + b/2)}{\Delta^{b/2} \Gamma(a_x)} C \right\}^{p_n - q_n} \right] 
\]

\[
= -q_n \log \left\{ \frac{2\Delta CT (a_x + b) 2^{-b} \Gamma(2b + 1)}{\Gamma(b)} \right\} - q_n \log \left[ 1 - \left\{ \frac{b \Gamma(b) \Gamma(a_x + b/2) C}{\Delta^{b/2} \log n} \right\}^{p_n - q_n} \right] 
\]

\[
+ q_n (2b + 1) \log \left( k_n + \frac{\Delta}{n^{\rho/2} p_n^{1/2}} \right) + q_n \log \log n + \frac{b + 1}{2} q_n \log p_n + \frac{b + 1}{2} q_n n^{\rho} \log n 
\]

The dominating term is the last one, and if \( q_n = o(n/\log n) \), \(- \log \operatorname{pr}(\beta_n : ||\beta_n - \beta_n^0|| < \Delta/n^{\rho/2} < dn \) for all \( d > 0 \), so \( \operatorname{pr}(\beta_n : ||\beta_n - \beta_n^0|| < \Delta/n^{\rho/2} > \exp(-dn) \). The posterior consistency is completed by applying Theorem 1 in Armagan et al. (2013b). \( \square \)

**Proof of Theorem 7** As shown in Theorem 4 when \( |\beta| \) is close to zero and \( 0 < a_x < 1/2 \), \( \pi_{R2-D2}(\beta) \approx C_1 + C_2/\beta^{1-2a_x} \), where \( C_1 \) and \( C_2 \) are some constants, so

\[
\int_0^{n^{-1/2}} \pi_{R2-D2}(\beta) \, d\beta \approx \int_0^{n^{-1/2}} \left( C_1 + \frac{C_2}{\beta^{1-2a_x}} \right) \, d\beta = n^{-1/2} C_2 \left( n^{1-2a_x} + C_1 \frac{2a_x}{C_2} \right). 
\]
By applying Lemma 1, we have
\[
R_n(R2-D2) \leq \frac{1}{n} \left( \frac{1}{\sqrt{n}} \frac{C_2}{2a} \left( n^{1/2-a} + C_1 \frac{2a}{C_2} \right) \right) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - \log(n^{1/2-a}) + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{n^{1/2-a}} \right) \right),
\]
much smaller than the risk of the Horseshoe and Horseshoe+ prior, i.e., \( O \left( \log \left( n/(\log n)^{b_0} \right) /n \right) \), where \( b_0 \) is some constant value (note: \( b_0 \) is different for Horseshoe and Horseshoe+ prior).

Proof of Theorem 8. As shown in the proof of theorem 5, when \( |\beta| \) is close to zero and \( 0 < a_D < 1 \), \( \pi_{DL}(\beta) \approx C|\beta|^{a_D-1} \), where \( C = \Gamma(1-a_D)/(2^{1+a_D}\Gamma(a_D)) \), so
\[
\int_0^{n^{-1/2}} \pi_{DL}(\beta) d\beta \approx C \int_0^{n^{-1/2}} |\beta|^{a_D-1} d\beta = \frac{C}{a_D} n^{-a_D/2} = \frac{C}{a_D} n^{-1/2} n^{1/2-a_D/2}.
\]
By applying Lemma 1, we have
\[
R_n(DL) \leq \frac{1}{n} \left( \frac{1}{\sqrt{n}} \frac{C_2}{2a} \left( n^{1/2-a} + C_1 \frac{2a}{C_2} \right) \right) \leq \frac{1}{n} \left( 1 + \frac{\log n}{2} - \log(n^{1/2-a}) + O(1) \right) = O \left( \frac{1}{n} \log \left( \frac{n}{n^{1/2-a}} \right) \right),
\]
much smaller than the risk of the Horseshoe and Horseshoe+ prior, i.e., \( O \left( \log \left( n/(\log n)^{b_0} \right) /n \right) \), where \( b_0 \) is some constant value (note: \( b_0 \) is different for Horseshoe and Horseshoe+ prior).

A.2 Gibbs Sampling Procedures

Denote \( Z \sim \text{InvGaussian}(\mu, \lambda) \), if \( \pi(z) = \lambda^{1/2}(2\pi z^3)^{-1/2}\exp\{-\lambda(z - \mu)^2/(2\mu^2z)\} \). Denote \( Z \sim \text{giG}(\chi, \rho, \lambda_0) \), the generalized inverse Gaussian distribution (Seshadri, 1997), if \( \pi(z) \propto z^{\lambda_0-1}\exp\{-\rho z + \chi/z\}/2 \).

The Gibbs sampling procedure is as follows:

(a) Sample \( \beta|\psi, \phi, \omega, \sigma^2, Y \sim N(\mu, \sigma^2V) \), where \( \mu = VX^TY = (X^TX + S^{-1})^{-1}(X^TY) \),
\[ V = (X^T X + S^{-1})^{-1}, \ S = \text{diag}\{\psi_1\varphi_1\omega/2, \cdots, \psi_p\varphi_p\omega/2\}, \ X = (x_1, \cdots, x_n)^T, \text{ and } Y = (Y_1, \cdots, Y_n)^T. \]

(b) Sample \( \sigma^2|\beta, \psi, \phi, \omega, Y \sim \text{IG}(a_1 + (n+p)/2, b_1 + (\beta^T S^{-1} \beta + (Y - X\beta)^T (Y - X\beta))/2). \)

(c) Sample \( \psi|\beta, \phi, \omega, \sigma^2. \) One can draw \( \psi_j^{-1} \sim \text{InvGaussian}(\mu_j = \sqrt{\sigma^2 \phi_j \omega/2}/|\beta_j|, \lambda = 1), \text{ then take the reciprocal to get } \psi_j. \)

(d) Sample \( \omega|\beta, \psi, \phi, \xi, \sigma^2 \sim \text{giG}(\chi = \sum_{j=1}^p \beta_j^2/(\sigma^2 \psi_j \phi_j), \rho = 2\xi, \lambda_0 = a - p/2). \)

(e) Sample \( \xi|\omega \sim \text{Ga}(a + b, 1 + \omega). \)

(f) Sample \( \phi|\beta, \psi, \xi, \sigma^2. \) Motivated by Bhattacharya et al. (2015), if \( a = pa_\pi, \) one can draw \( T_1, \cdots, T_p \) independently with \( T_j \sim \text{giG}(\chi = 2\beta_j^2/(\sigma^2 \psi_j), \rho = 2\xi, \lambda_0 = a_\pi - 1/2). \)
Then set \( \phi_j = T_j/T \) with \( T = \sum_{j=1}^p T_j. \)
B Figures and Tables

Figure 1: Marginal density of the R2-D2 (R2-D2), Dirichlet-Laplace (DL), Horseshoe, Horseshoe+ prior and Cauchy distribution. In all cases, the hyperparameters are selected to ensure the inter quartile range is 1 for visual comparison.

Figure 2: Convergence plots for posterior samples, fitted on the PEPCK data using different priors. Posterior samples of $\beta_1$ and $\beta_{999}$. 
Figure 3: Autocorrelation plots for posterior samples, fitted on the PEPCK data using different priors. Posterior samples of $\beta_1$ and $\beta_{999}$.

Figure 4: The agreement of the variable selection between two different priors, based on the induced ordering through the magnitude of posterior $t$-statistics for PEPCK ($n = 60$ and $p = 1000$). The point $(x, y)$ on the line means in the first ordered $x$-length variable list by using the prior on $x$-axis, there are $y$ variables matching with such list by using the prior on $y$-axis.
Table 1: Asymptotic properties for Horseshoe, Horseshoe+, R2-D2 (R2-D2) and Dirichlet-Laplace prior as discussed in Section 3

|         | Tail Decay          | Concentration at zero | Cesàro-average Risk Bound |
|---------|---------------------|-----------------------|---------------------------|
| Horseshoe | $O\left(\frac{1}{|\beta|^2}\right)$ | $O\left(\log(\frac{1}{|\beta|})\right)$ | $O\left(\frac{1}{n} \log\left(\frac{n}{\log n}\right)\right)$ |
| Horseshoe+ | $O\left(\frac{\log |\beta|}{|\beta|^2}\right)$ | $O\left(\log^2(\frac{1}{|\beta|})\right)$ | $O\left(\frac{1}{n} \log\left(\frac{n}{(\log n)^2}\right)\right)$ |
| R2-D2    | $O\left(\frac{1}{|\beta|^{1+2\theta}}\right)$ | $O\left(\frac{1}{|\beta|^{1-2\alpha \pi}}\right)$ | $O\left(\frac{1}{n} \log\left(\frac{n}{n^{1/2-a\pi}}\right)\right)$ |
| Dirichlet-Laplace | $O\left(\frac{1}{|\beta|^{a_D/2-3/4}} \exp\left(\frac{2}{|\beta|^{1+2}}\right)\right)$ | $O\left(\frac{1}{|\beta|^{1-8\theta}}\right)$ | $O\left(\frac{1}{n} \log\left(\frac{n}{n^{1/2-a_D/2}}\right)\right)$ |
Table 2: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| \in (0,0.5]$, and $|\beta_j| > 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the Receiver-Operating Characteristic (ROC) curve, with “ROC” means the area under the ROC curve, based on 200 datasets generated by Setup 1, with $n = 60$ and $\rho = 0.5$, standard errors in parentheses, all values multiplied by 100.

| $p$   | Prior            | SSE(= 0) | SSE((0,0.5]) | SSE(>0.5) | SSE(Total) | ROC     |
|-------|------------------|----------|--------------|-----------|------------|---------|
| 50    | Horseshoe        | 19 (0.9) | 12 (0.7)     | 48 (2.9)  | 78 (3.0)   | 89 (0.5) |
|       | Horseshoe+       | 15 (1.7) | 13 (0.7)     | 46 (2.8)  | 74 (3.4)   | 89 (0.5) |
|       | R2-D2(0.5,0.5)   | 5 (0.3)  | 14 (0.8)     | 60 (3.7)  | 79 (3.7)   | 90 (0.5) |
|       | R2-D2(p/n,0.5)   | 5 (0.3)  | 14 (0.7)     | 55 (3.4)  | 74 (3.4)   | 90 (0.5) |
|       | R2-D2(p/n,0.1)   | 5 (0.3)  | 14 (0.7)     | 58 (3.5)  | 77 (3.5)   | 90 (0.5) |
|       | R2-D2(1,1)       | 6 (0.3)  | 14 (0.7)     | 54 (3.3)  | 74 (3.3)   | 90 (0.5) |
|       | DL$_{1/p}$       | 1 (0.2)  | 19 (0.9)     | 171 (10.5)| 191 (10.5) | 89 (0.5) |
|       | DL$_{2/n}$       | 2 (0.4)  | 18 (0.9)     | 152 (10.6)| 172 (10.7) | 89 (0.5) |
|       | DL$_{1/n}$       | 1 (0.3)  | 19 (0.9)     | 178 (11.8)| 199 (11.8) | 89 (0.5) |
| 100   | Horseshoe        | 32 (3.2) | 16 (0.9)     | 65 (3.8)  | 113 (5.5)  | 87 (0.6) |
|       | Horseshoe+       | 22 (2.0) | 16 (1.0)     | 63 (4.1)  | 102 (5.2)  | 88 (0.6) |
|       | R2-D2(0.5,0.5)   | 6 (0.4)  | 17 (1.0)     | 90 (5.7)  | 113 (5.7)  | 89 (0.6) |
|       | R2-D2(p/n,0.5)   | 11 (0.6) | 16 (0.9)     | 70 (4.4)  | 96 (4.5)   | 89 (0.6) |
|       | R2-D2(p/n,0.1)   | 11 (0.6) | 16 (0.9)     | 69 (4.5)  | 96 (4.6)   | 89 (0.6) |
|       | R2-D2(1,1)       | 8 (0.5)  | 17 (1.0)     | 77 (5.0)  | 102 (5.1)  | 89 (0.6) |
|       | DL$_{1/p}$       | 1 (0.1)  | 21 (1.0)     | 257 (17.0)| 279 (16.9) | 88 (0.6) |
|       | DL$_{2/n}$       | 3 (0.2)  | 20 (1.0)     | 175 (10.3)| 197 (10.3) | 88 (0.6) |
|       | DL$_{1/n}$       | 2 (0.1)  | 21 (1.0)     | 222 (13.4)| 244 (13.4) | 88 (0.6) |
| 500   | Horseshoe        | 79 (5.4) | 21 (1.4)     | 122 (7.7) | 222 (11.2) | 82 (0.6) |
|       | Horseshoe+       | 63 (4.7) | 21 (1.1)     | 114 (7.4) | 199 (9.6)  | 83 (0.7) |
|       | R2-D2(0.5,0.5)   | 4 (0.2)  | 23 (1.1)     | 376 (30.0)| 402 (29.9) | 85 (0.6) |
|       | R2-D2(p/n,0.5)   | 17 (0.8) | 20 (1.0)     | 156 (9.5) | 193 (10.0) | 87 (0.5) |
|       | R2-D2(p/n,0.1)   | 19 (0.8) | 20 (1.0)     | 146 (8.8) | 184 (9.3)  | 87 (0.5) |
|       | R2-D2(1,1)       | 5 (0.3)  | 22 (1.1)     | 281 (23.7)| 308 (23.7) | 86 (0.6) |
|       | DL$_{1/p}$       | 3 (0.7)  | 24 (1.2)     | 561 (31.9)| 587 (31.8) | 84 (0.6) |
|       | DL$_{2/n}$       | 11 (0.7) | 21 (1.0)     | 235 (14.2)| 268 (14.5) | 86 (0.6) |
|       | DL$_{1/n}$       | 18 (9.5) | 22 (1.1)     | 316 (18.8)| 357 (21.7) | 86 (0.6) |
Table 3: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| \in (0,0.5]$, and $|\beta_j| > 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the Receiver-Operating Characteristic (ROC) curve, with “ROC” means the area under the ROC curve, based on 200 datasets generated by Setup 1, with $n = 60$ and $\rho = 0.9$, standard errors in parentheses, and all values multiplied by 100.

| $p$ | Prior | SSE($= 0$) | SSE(($0,0.5]$) | SSE($> 0.5$) | SSE(Total) | ROC |
|-----|-------|------------|---------------|--------------|------------|-----|
| 50  | Horseshoe | 31 (3.6) | 26 (1.9) | 292 (17.7) | 349 (19.5) | 81 (0.6) |
|     | Horseshoe+ | 24 (3.0) | 27 (2.1) | 290 (17.3) | 341 (18.8) | 81 (0.7) |
|     | R2-D2($0.5,0.5$) | 8 (1.2) | 26 (2.1) | 341 (19.4) | 375 (20.0) | 83 (0.6) |
|     | R2-D2($p/n,0.5$) | 10 (1.6) | 26 (2.2) | 322 (18.5) | 358 (19.4) | 83 (0.6) |
|     | R2-D2($p/n,0.1$) | 10 (1.5) | 26 (2.2) | 323 (18.5) | 359 (19.3) | 83 (0.6) |
|     | R2-D2($1,1$) | 10 (1.5) | 26 (2.2) | 325 (18.5) | 361 (19.5) | 83 (0.6) |
|     | DL$_{1/p}$ | 3 (0.4) | 27 (1.8) | 491 (25.6) | 521 (25.6) | 84 (0.7) |
|     | DL$_{2/n}$ | 4 (0.5) | 26 (1.7) | 467 (24.7) | 497 (24.7) | 84 (0.7) |
|     | DL$_{1/n}$ | 4 (0.4) | 27 (2.1) | 507 (26.4) | 538 (26.4) | 83 (0.7) |
| 100 | Horseshoe | 18 (1.6) | 29 (2.6) | 342 (17.0) | 389 (17.5) | 80 (0.6) |
|     | Horseshoe+ | 18 (2.2) | 29 (2.9) | 342 (17.3) | 389 (18.4) | 80 (0.7) |
|     | R2-D2($0.5,0.5$) | 8 (0.8) | 27 (2.2) | 397 (20.2) | 432 (20.6) | 83 (0.6) |
|     | R2-D2($p/n,0.5$) | 11 (0.9) | 27 (2.3) | 352 (18.0) | 390 (18.5) | 83 (0.6) |
|     | R2-D2($p/n,0.1$) | 12 (1.0) | 27 (2.3) | 347 (17.8) | 386 (18.4) | 83 (0.6) |
|     | R2-D2($1,1$) | 9 (0.8) | 27 (2.2) | 375 (19.2) | 411 (19.6) | 83 (0.6) |
|     | DL$_{1/p}$ | 4 (0.7) | 28 (2.5) | 565 (28.8) | 598 (29.0) | 84 (0.6) |
|     | DL$_{2/n}$ | 5 (0.6) | 27 (1.9) | 478 (24.3) | 510 (24.3) | 84 (0.6) |
|     | DL$_{1/n}$ | 5 (1.1) | 29 (2.5) | 547 (27.9) | 581 (28.1) | 83 (0.6) |
| 500 | Horseshoe | 23 (3.8) | 31 (2.3) | 518 (26.4) | 572 (26.9) | 75 (0.7) |
|     | Horseshoe+ | 20 (3.0) | 30 (2.2) | 526 (30.0) | 577 (30.3) | 75 (0.8) |
|     | R2-D2($0.5,0.5$) | 4 (0.3) | 27 (1.8) | 703 (39.2) | 734 (39.1) | 83 (0.7) |
|     | R2-D2($p/n,0.5$) | 16 (0.9) | 27 (1.7) | 562 (30.1) | 604 (30.1) | 81 (0.6) |
|     | R2-D2($p/n,0.1$) | 17 (0.8) | 27 (1.7) | 552 (29.6) | 596 (29.6) | 81 (0.6) |
|     | R2-D2($1,1$) | 5 (0.4) | 27 (1.8) | 661 (37.3) | 693 (37.2) | 82 (0.7) |
|     | DL$_{1/p}$ | 4 (0.8) | 31 (2.5) | 796 (43.6) | 830 (43.6) | 85 (0.7) |
|     | DL$_{2/n}$ | 15 (2.4) | 26 (1.7) | 570 (29.9) | 611 (29.8) | 84 (0.7) |
|     | DL$_{1/n}$ | 6 (0.4) | 27 (1.7) | 639 (32.8) | 671 (32.7) | 85 (0.6) |
Table 4: The average of the sum of squared error (SSE) for $\beta_j$ is given separately for $\beta_j = 0$, $|\beta_j| < 0.5$, and $|\beta_j| \geq 0.5$, as well as the sum over all $\beta_j$, i.e., SSE(Total), together with the average area under the Receiver-Operating Characteristic (ROC) curve, with “ROC” means the area under the ROC curve, based on 200 datasets generated by Setup 2, with $n = 60$ and $p = 100$, standard errors in parentheses, and all values multiplied by 100

| $\rho$ | Prior       | SSE(0) | SSE((0, 0.5]) | SSE(>0.5) | SSE(Total) | ROC |
|--------|-------------|--------|---------------|-----------|------------|-----|
| 0.5    | Horseshoe   | 4 (0.4)| 19 (0.8)      | 8 (1.0)   | 27 (1.2)   | 72  (1.0)|
|        | Horseshoe+  | 4 (0.4)| 19 (0.8)      | 8 (1.0)   | 27 (1.2)   | 72  (1.0)|
|        | R2-D2(0.5,0.5) | 1 (0.1)| 18 (0.8)      | 13 (1.6)  | 30 (1.7)   | 74  (1.1)|
|        | R2-D2(p/n,0.5) | 3 (0.2)| 18 (0.7)      | 9 (1.1)   | 26 (1.2)   | 73  (1.0)|
|        | R2-D2(p/n,0.1) | 3 (0.2)| 18 (0.7)      | 9 (1.1)   | 26 (1.2)   | 73  (1.0)|
|        | R2-D2(1,1)  | 2 (0.1)| 17 (0.7)      | 10 (1.3)  | 28 (1.4)   | 73  (1.0)|
|        | DL1/p       | 0 (0.0)| 18 (0.8)      | 14 (1.8)  | 32 (1.8)   | 74  (1.0)|
|        | DL2/n       | 1 (0.1)| 17 (0.7)      | 11 (1.4)  | 28 (1.5)   | 74  (1.1)|
|        | DL1/n       | 1 (0.1)| 17 (0.7)      | 13 (1.7)  | 31 (1.7)   | 74  (1.0)|
| 0.9    | Horseshoe   | 2 (0.4)| 23 (1.1)      | 23 (2.5)  | 46 (2.7)   | 73  (1.2)|
|        | Horseshoe+  | 3 (0.4)| 24 (1.1)      | 23 (2.5)  | 47 (2.8)   | 73  (1.1)|
|        | R2-D2(0.5,0.5) | 1 (0.1)| 21 (0.9)      | 24 (2.6)  | 44 (2.7)   | 78  (1.3)|
|        | R2-D2(p/n,0.5) | 2 (0.3)| 22 (0.9)      | 22 (2.4)  | 44 (2.7)   | 76  (1.2)|
|        | R2-D2(p/n,0.1) | 2 (0.3)| 22 (0.9)      | 22 (2.4)  | 43 (2.6)   | 75  (1.2)|
|        | R2-D2(1,1)  | 1 (0.2)| 21 (0.9)      | 23 (2.5)  | 44 (2.7)   | 77  (1.3)|
|        | DL1/p       | 1 (0.1)| 21 (0.9)      | 24 (2.6)  | 45 (2.9)   | 79  (1.3)|
|        | DL2/n       | 2 (0.2)| 20 (0.8)      | 22 (2.5)  | 42 (2.6)   | 77  (1.2)|
|        | DL1/n       | 1 (0.1)| 21 (1.0)      | 24 (2.7)  | 45 (2.9)   | 79  (1.3)|

Table 5: Mean squared prediction error, with standard errors in parenthesis, based on 100 random splits of the real data

|            | PEPCK   | GPAT    | SCD1    |
|------------|---------|---------|---------|
| Horseshoe  | 0.51 (0.027) | 1.01 (0.079) | 0.56 (0.045) |
| Horseshoe+ | 0.51 (0.028) | 1.10 (0.083) | 0.61 (0.060) |
| R2-D2(0.5,0.5) | 0.50 (0.028) | 0.82 (0.064) | 0.81 (0.078) |
| R2-D2(p/n,0.5) | 0.49 (0.026) | 0.90 (0.073) | 0.58 (0.056) |
| R2-D2(p/n,0.1) | 0.49 (0.026) | 0.90 (0.073) | 0.57 (0.055) |
| R2-D2(1,1)   | 0.50 (0.028) | 0.83 (0.065) | 0.71 (0.067) |
| DL1/p       | 0.52 (0.030) | 0.87 (0.073) | 0.94 (0.201) |
| DL1/n       | 0.49 (0.028) | 0.88 (0.074) | 0.62 (0.056) |