A Laplace Mixture Representation of the Horseshoe and Some Implications

Ksheera Sagar and Anindya Bhadra

***Abstract*—The horseshoe prior, defined as a half Cauchy scale mixture of normal, provides a state of the art approach to Bayesian sparse signal recovery. We provide a new representation of the horseshoe density as a scale mixture of the Laplace density, explicitly identifying the mixing measure. Using the celebrated Bernstein–Widder theorem and a result due to Bochner, our representation immediately establishes the complete monotonicity of the horseshoe density and strong concavity of the corresponding penalty. Consequently, the equivalence between local linear approximation and expectation–maximization algorithms for finding the posterior mode under the horseshoe penalized regression is established. Further, the resultant estimate is shown to be sparse.**

***Index Terms*—Bayesian sparse signal recovery, Bernstein–Widder, completely monotone function, local linear approximation**

I. Introduction

FOR response y, covariates Φ and regression coefficients x, consider the high-dimensional linear regression model:

$$y = \Phi x + \epsilon, \tag{1}$$

where $y = \{y_j\} \in \mathbb{R}^n, \Phi = \{\phi_{ji}\} \in \mathbb{R}^{n\times p}, x = \{x_i\} \in \mathbb{R}^p$ with $p \gg n$ and $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$. The horseshoe prior hierarchy [1] is defined for $i = i, \ldots, p$ as a half Cauchy scale mixture of normal as:

$$x_i \mid \lambda_i, \tau \overset{ind}{\sim} \mathcal{N}(0, \lambda_i^2\tau^2),\ \lambda_i \overset{ind}{\sim} \mathcal{C}^+(0,1),\ \tau > 0, \tag{2}$$

where $\mathcal{C}^+$ denotes a standard half Cauchy random variable with density $p(\lambda_i) = (2/\pi)(1+\lambda_i^2)^{-1},\ \lambda_i > 0$ and τ is a hyperparameter. This prior has found numerous applications in sparse signal recovery [1]–[9]. Although there is no closed form to the marginal horseshoe density, obtained by integrating out the latent λ_is, the main benefits of the horseshoe stem from two crucial properties of the marginal prior. The first is infinite prior mass at the origin, allowing strong shrinkage towards zero in a sparse regime; and the second is heavy, polynomially decaying tails, aiding detection of signals [1].

Since the tails of the horseshoe prior follow a quadratic rate of decay, it is conceivable to expect the prior density to admit a mixture representation with respect to the double exponential (Laplace) density as well. Yet, to our knowledge, no such explicit representation is available in the literature. We close this gap by showing the horseshoe density can also be represented as a mixture of double exponential with respect to a random variable whose density is proportional to the Dawson function. This representation, in turn, allows us to appeal to the celebrated Bernstein–Widder theorem [10] and a result due to Bochner [11] to establish complete monotonicity of the horseshoe density and strong concavity of the horseshoe penalty, both new results. Further, appealing to the same mixture representation, we are able to establish an equivalence between local linear approximation or LLA [12], [13] and expectation–maximization or EM [14] schemes for finding the maximum a posteriori (MAP) estimate under the horseshoe prior for sparse signal recovery.

Submitted on 11/28/2022. Bhadra is supported by US National Science Foundation Grant DMS-2014371.

The authors are with the Department of Statistics, Purdue University, IN 47906, United States of America (e-mails: kkeralap@purdue.edu; bhadra@purdue.edu).

II. A Laplace Mixture Representation of the Horseshoe

Proposition 1. *The marginal horseshoe density for a scalar random variable ($p = 1$) admits the following representation as a Laplace mixture:*

$$p_{HS}(x) \;=\; \frac{2}{\pi\sqrt{\pi}\tau}\int_0^\infty \exp\left(-u\frac{|x|}{\tau}\right) D_+\left(\frac{u}{\sqrt{2}}\right) du,$$

where $D_+(z) = \exp(-z^2)\int_0^z \exp(t^2)dt,\ \ z > 0$, is the Dawson function.

Proof. We have using the hierarchy of Equation (2),

$$p_{HS}(x) = \int_0^\infty \frac{1}{\sqrt{2\pi\lambda^2\tau^2}}\exp\left(-\frac{x^2}{2\lambda^2\tau^2}\right)\frac{2}{\pi}\frac{1}{1+\lambda^2}d\lambda$$

$$= \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{x^2}{2\lambda^2\tau^2}\right)\int_0^\infty \exp(-(\lambda^2+1)t)dtd\lambda$$

$$= \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty\int_0^\infty \frac{1}{\lambda}\exp\left(-\frac{x^2}{2\lambda^2\tau^2}-\lambda^2 t\right)\exp(-t)dtd\lambda$$

$$= \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty \left\{\int_0^\infty \exp\left(-\frac{x^2}{2\omega\tau^2}-\omega t\right)\frac{1}{2}\omega^{-1}d\omega\right\}\exp(-t)dt$$

(substituting $\lambda^2 = \omega$, and using Fubini's theorem)

$$= \frac{1}{\pi\sqrt{2\pi}\tau}\int_0^\infty 2K_0\left(\sqrt{2t}\frac{|x|}{\tau}\right)\exp(-t)dt$$

(generalized inverse Gaussian integral [15];

K_0 = modified Bessel function of the second kind of order 0)

$$= \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty \left\{\int_0^\infty \exp\left(-\frac{|x|}{\tau}\sqrt{2t}\cosh\zeta\right)d\zeta\right\}\exp(-t)dt$$

(using identity 9.6.24 in [16])

$$= \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty\int_0^\infty \exp\left(-\frac{|x|}{\tau}\sqrt{2t}\cosh\zeta - t\right)d\zeta dt.$$

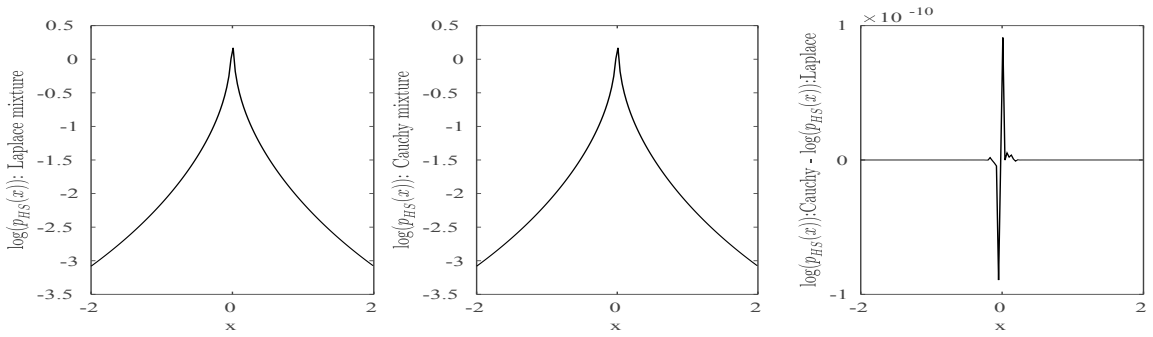

Fig. 1. Numerical validation of Proposition 1. Left panel: logarithm of horseshoe density, $\log p_{HS}(x)$, evaluated numerically, using the Laplace scale mixture in Proposition (1). Middle panel: logarithm of horseshoe density, $\log p_{HS}(x)$, evaluated numerically, using the normal scale mixture in Equation (2). Right panel: Difference between the logarithms of densities evaluated using normal and Laplace mixtures.

Let $\sqrt{2t}\cosh\zeta = u$ and $t = v$. The inverse transformations are, $\cosh^{-1}(u/\sqrt{2v}) = \zeta$ and $v = t$; with Jacobian:

$$J = \begin{vmatrix} \frac{\partial\zeta}{\partial u} & \frac{\partial\zeta}{\partial v} \\ \frac{\partial t}{\partial u} & \frac{\partial t}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{\partial\zeta}{\partial u} & \frac{\partial\zeta}{\partial v} \\ 0 & 1 \end{vmatrix} = \left|\frac{\partial\zeta}{\partial u}\right| = \frac{1}{\sqrt{u^2-2v}}.$$

Thus, $p_{HS}(x)$ equals,

$$\begin{aligned} & \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty\int_0^{u^2/2}\exp\left(-\frac{|x|}{\tau}u - v\right)\frac{1}{\sqrt{u^2-2v}}dvdu \\ = & \frac{\sqrt{2}}{\pi\sqrt{\pi}\tau}\int_0^\infty\exp\left(-\frac{|x|}{\tau}u\right)\left\{\int_0^{u^2/2}\frac{\exp(-v)}{\sqrt{u^2-2v}}dv\right\}du \\ = & \frac{2}{\pi\sqrt{\pi}\tau}\int_0^\infty\exp\left(-\frac{|x|}{\tau}u\right)\left\{e^{-u^2/2}\int_0^{\frac{u}{\sqrt{2}}}e^{t^2}dt\right\}du \\ & \text{(substituting } u^2-2v = 2t^2\text{)} \\ = & \frac{2}{\pi\sqrt{\pi}\tau}\int_0^\infty\exp\left(-\frac{|x|}{\tau}u\right)D_+\left(\frac{u}{\sqrt{2}}\right)du. \qquad \square \end{aligned}$$

Proposition 1 is our main result and the remainder of the paper discusses its implications in sparse signal recovery. Numerical validation of Proposition 1 is presented in Fig. 1 by evaluating the horseshoe density using both the usual normal scale mixture (cf. Equation (2)) and the newly derived Laplace scale mixture representations over $[-2, 2]$. The pointwise maximum difference in density evaluation (in logarithmic scale) between the two representations is less than 10^{-10}.

III. Main Implications of Proposition 1: Complete Monotonicity of the Horseshoe Density and Strong Concavity of the Penalty

We begin by stating the definition of a completely monotone function and the celebrated Bernstein–Widder theorem [10, Chapter 4].

Definition 2. *A function $g(\cdot)$ is said to be completely monotone if $(-1)^k g^{(k)}(x) \geq 0$ for $x \geq 0,\ k = 0, 1, \ldots,$ with superscripts (k) denoting the order of derivatives and $g^{(0)} \equiv g$.*

Theorem 3. *(Bernstein–Widder). The function $g(\cdot)$ is completely monotone if and only if it is the Laplace transform of a positive measure γ, i.e., admits the representation: $g(x) = \int_0^\infty \exp(-xu)d\gamma(u),\ x \geq 0$.*

Since the Dawson function is positive over $(0, \infty)$, and the horseshoe density is symmetric around zero, Proposition 1 in conjunction with Theorem 3 immediately establishes the complete monotonicity of $p_{HS}(|x|)$.

In a penalized likelihood setting, it is customary to view the negative log likelihood as the loss function and the negative log prior density as the penalty in an optimization problem. The following result, due to Bochner [11, Theorem 4.1.5] establishes the strong concavity of the horseshoe penalty.

Theorem 4. *(Bochner). The function $g(x) = \exp(-a\phi(x)), x \geq 0$, is completely monotone for every $a > 0$ if and only if $\phi'(x) = \phi^{(1)}(x)$ is completely monotone.*

Since $p_{HS}(|x|)$ is completely monotone, Theorem 4 implies the corresponding penalty function $\mathrm{pen}_{HS}(|x|) = -\log p_{HS}(|x|)$ is a Bernstein function (integral of a completely monotone function) and thus strongly concave. In other words, $\mathrm{pen}'_{HS}(|x|) = -(p'_{HS}(|x|))/(p_{HS}(|x|))$ is completely monotone. This property of the penalty has important implications on the convergence of optimization algorithms for penalized likelihood approaches, which we discuss next.

IV. Sparse Signal Recovery via LLA

A. Local Linear Approximation of Non-convex Penalties

Consider the penalized estimation problem:

$$\underset{x}{\mathrm{argmin}}\left(\ell(x; y) + \mathrm{pen}(|x|)\right), \tag{3}$$

where $\ell(x; y) = -\log\mathcal{L}(x; y)$ is a loss function providing some measure of model fit to the data y, usually a negative log likelihood function and $\mathrm{pen}(|x|)$ is a penalty symmetric around zero, with $|\cdot|$ denoting the vector ℓ_1 norm. Although convex penalties are simpler to study, both theoretically and computationally, concave penalties enjoy several desirable statistical properties such as near-unbiasedness, minimax optimal properties and sparsity [17], [18]. Nevertheless, the optimization problem becomes much more challenging, and usually, it is only possible to find a local optimum. Required conditions for global optimum under concave penalties are discussed by [19].

The local linear approximation (LLA) algorithm of [12] provides one approach for solving the problem in Equation (3) for concave penalties. LLA iterations proceed via a first order Taylor approximation of the penalty function, as follows

$$\text{pen}(|x|) \approx \text{pen}(|x_{(k)}|) + \text{pen}'(|x_{(k)}|)(|x| - |x_{(k)}|),$$

where $x_{(k)}$ is the current parameter value. With this approximation, and assuming the loss (but not necessarily the penalty) is convex, the non-convex optimization problem in Equation (3) reduces to the convex problem

$$x_{(k+1)} = \underset{x}{\text{argmin}} \left(\ell(x; y) + \lambda_{(k)}|x| \right), \tag{4}$$

where $\lambda_{(k)} = \text{pen}'(|x_{(k)}|)$. With the common squared error loss function, $\ell(x; y) = |x - y|^2$, Equation (4) defines a lasso problem [20], which can be solved very efficiently. In addition, the resultant solution is naturally sparse.

B. Implications of Complete Monotonicity on LLA

Numerical convergence and statistical properties of LLA estimates are non-trivial objects of study for general non-convex penalties and properties of LLA for general penalties are largely unknown. Nevertheless, in the special case when the penalty under consideration is a Bernstein function, certain simplifications ensue due to the following result of [12].

Theorem 5. *(Zou and Li). Consider a penalty* $\text{pen}(|x|)$ *such that* $\exp(-\text{pen}(|x|))$ *is completely monotone. Then, under a Gaussian likelihood, the LLA algorithm is equivalent to an EM algorithm for MAP estimation under a prior, not necessarily proper, with density* $p_X(x) \propto \exp(-\text{pen}(x))$.

Thus, an application of Proposition 1 immediately establishes the LLA algorithm as an EM algorithm for horseshoe penalties. This result is of practical significance, since the convergence properties of LLA can then be studied using the same tools for EM, which are better understood [21]. This equivalence is also established by [13], who provide a variational interpretation. Thus, it is established that the MAP estimate under the horseshoe prior is sparse; due to the soft thresholding operator in LLA resulting in exact zeros. Further, to our knowledge, this is the first viable algorithm for finding the MAP estimate under the horseshoe.

C. Implementing LLA under the Horseshoe Penalty

Consider now the practicalities of LLA under the horseshoe penalty. For implementation, one needs to compute $\lambda_{(k)} = \text{pen}'_{HS}(|x_{(k)}|)$. Differentiating the expression of Proposition 1 under the integral sign, we have:

$$\lambda_{(k)} = \text{pen}'_{HS}(|x_{(k)}|) = \frac{\int_0^\infty \frac{u}{\tau} \exp\left(-u\frac{|x_{(k)}|}{\tau}\right) D_+\left(\frac{u}{\sqrt{2}}\right) du}{\int_0^\infty \exp\left(-u\frac{|x_{(k)}|}{\tau}\right) D_+\left(\frac{u}{\sqrt{2}}\right) du}. \tag{5}$$

We provide numerical details of evaluating $\lambda_{(k)}$ in the Appendix. Subsequent to this, we use efficient solvers based on pathwise coordinate optimization [22], for solving the lasso problem in Equation (4). We note here that using the expression from Equation (5) is not necessary for evaluating pen'_{HS}, strictly speaking, and other alternative numerical methods can be used. However, in our experience, Equation (5) is numerically very stable.

V. Numerical Experiments

A. Sparse Normal Means Model

Revisiting the regression model in Equation (1), and setting $p = n$, $\Phi = \mathbf{I}_n$, we recover the normal means model $y_i \overset{ind}{\sim} \mathcal{N}(x_i, \sigma^2)$. In our experiments, we set $n = 50$, $\sigma^2 = 1$ and the components in x as, $x_1 = x_2 = 3$, $x_3, \ldots, x_{50} = 0$; y_i is sampled from $\mathcal{N}(x_i, \sigma^2)$. With the generated data, optimizing Equation (4) with $\ell(x; y) = \frac{1}{2}\sum_{i=1}^n (y_i - x_i)^2$, with pathwise coordinate optimization [22], we have :

$$x_{(k+1)} = \underset{x}{\text{argmin}} \left(\frac{1}{2}\sum_{i=1}^n (y_i - x_i)^2 + \sum_{i=1}^n \lambda_{i(k)}|x_i| \right),$$

yielding, $x_{i(k+1)} = S(y_i, \lambda_{i(k)})$, where,

$$S(\beta, \gamma) = \begin{cases} \beta - \gamma, & \text{if } \beta > 0 \text{ and } \gamma < |\beta|, \\ \beta + \gamma, & \text{if } \beta < 0 \text{ and } \gamma < |\beta|, \\ 0, & \text{if } \gamma \geq |\beta|, \end{cases} \tag{6}$$

is the soft thresholding operator. In the above, $x_{i(k+1)}$ denotes the value of i^{th} component of x, at $(k+1)^{th}$ iteration. Similarly, $\lambda_{i(k)}$ denotes the derivative of the penalty in Equation (5), evaluated at $x_{i(k)}$. We estimate the normal means, $\hat{x} = \{\hat{x}_i\}$, under the horseshoe penalty using LLA; and compare with the posterior mean of x under the horseshoe prior obtained by Markov chain Monte Carlo (HS-MCMC), the minimax concave penalty or MCP [23], the smoothly clipped absolute deviation penalty or SCAD [17] and the lasso [20] penalties, using the `R` packages `horseshoe` [24], `sparsenet` [19], `ncvreg` [25] and `glmnet` [26] respectively. MCP are SCAD are non-convex penalties, while lasso is convex. Under all the point estimation methods, the respective tuning parameters are chosen using 10-fold cross validation; whereas for HS-MCMC, we consider $\tau \sim \mathcal{C}^+(0, 1)$, a total of 1.5×10^4 posterior samples with 5000 samples as burnin, and 50% credible intervals for variable selection [3]. For the LLA procedure, we start with $x_{i(0)} = 1, \forall i$ and set the stopping criterion as $\sum_{i=1}^n \left(\hat{x}_{i(k+1)} - \hat{x}_{i(k)}\right)^2 < 10^{-6}$. To compare the estimates, we compute the in sample sum of squared errors or SSE $= \sum_{i=1}^n (\hat{x}_i - x_i)^2$, out of sample prediction SSE (pSSE) $= \|y_{\text{out}} - \Phi\hat{x}\|^2$; for y_{out} generated similarly as y, true negative (positive) rates, for zeros (non-zeros) in x identified correctly as TNR, TPR respectively and the computational time in seconds. We estimate $\hat{x}$ for 50 replications, and present the mean and standard deviation (in parentheses) of the performance measures in Table I. We observe that HS-LLA has the best pSSE and its SSE and TNR are competitive with the resultant MCMC estimate. The good statistical performance of the posterior mean estimate obtained by MCMC is not surprising, since it is the optimal Bayes estimate under ℓ_2 loss. However, the MCMC estimate is about 3 times slower than than HS-LLA, for negligible gain in statistical performance.

TABLE I
COMPARISON OF RESULTS FOR COMPETING PROCEDURES, IN A SPARSE NORMAL MEANS MODEL.

	HS-LLA	HS-MCMC	SCAD	MCP	lasso
SSE	10.051 (6.721)	9.757 (3.803)	26.22 (18.629)	32.724 (17.794)	28.121 (18.177)
pSSE	59.238 (16.771)	61.381 (12.506)	75.641 (25.609)	80.358 (22.146)	76.082 (23.228)
TNR	0.982 (0.023)	0.992 (0.013)	0.633 (0.336)	0.65 (0.325)	0.464 (0.366)
TPR	0.64 (0.351)	0.53 (0.409)	0.86 (0.268)	0.84 (0.31)	0.93 (0.202)
Time (s)	0.351	1.07	0.268	0.31	0.202

B. Sparse Linear Regression

We set $n = 50, p = 100$ and the components x as $x_1 = 3, x_2 = 1.5, x_3 = 2, x_4, \ldots, x_{100} = 0$. The entries in columns of Φ, where $\Phi = [\Phi_1, \ldots, \Phi_p]$ are generated from a multivariate Gaussian, with zero mean, such that the covariance between Φ_k and Φ_l, for $1 \leq k, l \leq p$ is $0.5^{|k-l|}$. The observations y_j are generated as $y_j \sim \mathcal{N}(\Phi x, 1)$, for $j \in \{1, \ldots, n\}$. With the generated data, optimizing (4) with $\ell(x; y) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \Phi x)^2$, with pathwise coordinate optimization [22], we have :

$$x_{(k+1)} = \underset{x}{\text{argmin}} \left(\frac{1}{2}\sum_{j=1}^{n}(y_i - \Phi x)^2 + \sum_{i=1}^{p}\lambda_{i(k)}|x_i| \right),$$

$$\text{yielding, } x_{i(k+1)} = \frac{1}{\sum_{j=1}^{n}\phi_{ji}^2} S\left(\sum_{j=1}^{n}\phi_{ji}\left(y_j - \widetilde{y}_{ji}\right), \lambda_{i(k)} \right),$$

$$\text{where, } \widetilde{y}_{ji} = \sum_{l \neq i}\phi_{jl}\widetilde{x}_l \,,\; \widetilde{x}_l = \begin{cases} x_{l(k+1)}, & \text{if } i > l, \\ x_{l(k)}, & \text{else.} \end{cases}$$

and the definition of $S(\cdot, \cdot)$ follows from Equation (6). For the LLA procedure, we start with $x_{i(0)} = 0.1$, $\forall i$ and set the stopping criterion as mentioned in Section V-A. We estimate $\hat{x}$ for 50 replications, and present the mean and standard deviation (in parentheses) of the performance measures in Table II. From these results, we observe that HS-MCMC has the best TNR, followed by HS-LLA and other measures under HS-LLA are comparable to the best performer, which is SCAD for this setting.

We again observe from the results presented in Table II that the performance measures are comparable for HS-MCMC and HS-LLA. However, HS-MCMC is drastically slower and also requires a choice for the credible interval width for variable selection; which is not the case under HS-LLA. This underlines the scalability of our approach for Bayesian sparse signal recovery.

TABLE II
COMPARISON OF RESULTS FOR COMPETING PROCEDURES, IN SPARSE LINEAR REGRESSION.

	HS-LLA	HS-MCMC	SCAD	MCP	lasso
SSE	0.207 (0.2)	0.183 (0.175)	0.118 (0.108)	0.182 (0.166)	0.281 (0.165)
pSSE	59.612 (12.918)	58.038 (11.45)	56.607 (11.753)	59.233 (14.665)	65.817 (16.688)
TNR	0.987 (0.019)	1 (0)	0.975 (0.026)	0.975 (0.064)	0.92 (0.077)
TPR	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)
Time (s)	0.26	4.23	0.05	0.29	0.1

VI. CONCLUDING REMARKS

We provide a new Laplace mixture representation of the popular horseshoe prior. While this can be used as an alternative generative hierarchy for the horseshoe, similar to the usual representation of half Cauchy scale mixture of normals, our results show horseshoe prior enjoys some particularly nice properties due to the resultant complete monotonicity of the density function. This is relevant because not all normal scale mixtures are also automatically Laplace mixtures. For example, the class of power exponential densities, $p_X(x) \propto \exp(-|x|^\alpha)$, can be written as a normal scale mixture with respect to a positive $\alpha/2$ stable variable for $\alpha \in (1, 2)$ [27], but these densities are not Laplace mixtures. In addition, our results imply a numerically stable LLA approach for penalized likelihood estimation problem under the horseshoe, and establishes LLA as special case of EM, thus proving LLA in this case results in the MAP estimate, which is sparse, due to the soft thresholding step in LLA. Deeper studies of these connections should be considered future work.

APPENDIX

The hyperparameter τ. Proposition 1 is presented for a fixed $\tau > 0$. In both fully Bayesian and penalized likelihood settings, τ needs to be tuned for practical implementation. We used cross validation for our numerical results in Section V. Fully Bayesian approaches require a hyperprior on τ, popular among them the half-Cauchy [28]. Other techniques for choosing τ include Akaike information criterion or AIC [29], marginal likelihood [30] and effective model size [31]. For a detailed survey, see [5, Section 5].

Fast computation of $\lambda_{(k)}$. Computing the ratio of integrals in Equation (5), at every iteration of the optimization procedure might be perceived as a potential drawback, for two possible reasons: (a) closed form of the integrals do not exist and (b) numerical integration might be a time intensive exercise. Luckily, near-minimax rational approximation of the Dawson function exists [32], and is given by:

$$\frac{D_+(u)}{u} \approx G(u) = \frac{1 + \frac{33}{232}u^2 + \frac{19}{632}u^4 + \frac{23}{1471}u^6}{1 + \frac{517}{646}u^2 + \frac{58}{173}u^4 + \frac{11}{262}u^6 + \frac{46}{1471}u^8}.$$

This approximation has a maximum relative error of 6.1×10^{-4} for $|u| < \infty$, is exact at $u = 0$ and has the asymptotic property, $uG(u) \sim D_+(u)$, as $u \to \infty$ [32]. Therefore, having the values of $D_+(u)$ pre-computed and stored on a large enough and fine grid of u, a simple Riemann sum can be used to evaluate the expressions in Equation (5). Lipschitz continuity of the rational approximation can also be established with some further algebra, aiding a selection of grid size.

REFERENCES

[1] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, pp. 465–480, 2010.

[2] ——, "Handling sparsity via the horseshoe," in *Artificial Intelligence and Statistics*. PMLR, 2009, pp. 73–80.

[3] Y. Li, B. A. Craig, and A. Bhadra, "The graphical horseshoe estimator for inverse covariance matrices," *Journal of Computational and Graphical Statistics*, vol. 28, no. 3, pp. 747–757, 2019.

[4] H. Yu, L. Xin, and J. Dauwels, "Variational Wishart approximation for graphical model selection: Monoscale and multiscale models," *IEEE Transactions on Signal Processing*, vol. 67, no. 24, pp. 6468–6482, 2019.

[5] A. Bhadra, J. Datta, N. G. Polson, and B. T. Willard, "Lasso meets horseshoe: A survey," *Statistical Science*, vol. 34, no. 3, pp. 405–427, 2019.

[6] A. Bhadra, J. Datta, Y. Li, and N. Polson, "Horseshoe regularisation for machine learning in complex and deep models," *International Statistical Review*, vol. 88, pp. 302–320, 2020.

[7] J. Lee, J. Son, S. Zhou, and Y. Chen, "Variation source identification in manufacturing processes using Bayesian approach with sparse variance components prior," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 3, pp. 1469–1485, 2020.

[8] A. Bhadra, J. Datta, N. G. Polson, and B. T. Willard, "The horseshoe-like regularization for feature subset selection," *Sankhya B*, vol. 83, no. 1, pp. 185–214, 2021.

[9] Y. Li, J. Datta, B. A. Craig, and A. Bhadra, "Joint mean–covariance estimation via the horseshoe," *Journal of Multivariate Analysis*, vol. 183, p. 104716, 2021.

[10] D. V. Widder, *The Laplace Transform*. Princeton University Press, 1946.

[11] S. Bochner, *Harmonic analysis and the theory of probability*. University of California Press, 1955.

[12] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[13] N. G. Polson and J. G. Scott, "Mixtures, envelopes and hierarchical duality," *Journal of the Royal Statistical Society: Series B*, vol. 78, no. 4, pp. 701–727, 2016.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, 1977.

[15] O. E. Barndorff-Nielsen, "Exponentially decreasing distributions for the logarithm of particle size," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 353, no. 1674, pp. 401–419, 1977.

[16] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. US Government printing office, 1964, vol. 55.

[17] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, 2001.

[18] J. Fan, L. Xue, and H. Zou, "Strong oracle optimality of folded concave penalized estimation," *Annals of Statistics*, vol. 42, no. 3, pp. 819–849, 2014.

[19] R. Mazumder, J. H. Friedman, and T. Hastie, "Sparsenet: Coordinate descent with nonconvex penalties," *Journal of the American Statistical Association*, vol. 106, no. 495, pp. 1125–1138, 2011.

[20] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[21] C. F. J. Wu, "On the convergence properties of the EM algorithm," *The Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[22] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[23] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *The Annals of statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[24] S. van der Pas, J. Scott, A. Chakraborty, and A. Bhattacharya, "horseshoe: Implementation of the horseshoe prior," *R package version 0.1. 0*, vol. 12, 2016.

[25] P. Breheny and J. Huang, "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, vol. 5, no. 1, pp. 232–253, 2011.

[26] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010. [Online]. Available: https://www.jstatsoft.org/v33/i01/

[27] M. West, "On scale mixtures of normal distributions," *Biometrika*, vol. 74, no. 3, pp. 646–648, 1987.

[28] N. G. Polson and J. G. Scott, "On the half-cauchy prior for a global scale parameter," *Bayesian Analysis*, vol. 7, no. 4, pp. 887–902, 2012.

[29] C. Lingjærde, B. P. Fairfax, S. Richardson, and H. Ruffieux, "Scalable multiple network inference with the joint graphical horseshoe," *arXiv preprint arXiv:2206.11820*, 2022.

[30] A. Bhadra, K. Sagar, S. Banerjee, and J. Datta, "Graphical evidence," *arXiv preprint arXiv:2205.01016*, 2022.

[31] J. Piironen and A. Vehtari, "Sparsity information and regularization in the horseshoe and other shrinkage priors," *Electronic Journal of Statistics*, vol. 11, no. 2, pp. 5018–5051, 2017.

[32] F. G. Lether, "Constrained near-minimax rational approximations to Dawson's integral," *Applied Mathematics and Computation*, vol. 88, no. 2-3, pp. 267–274, 1997.