Title

NON-LOCAL PRIORS FOR HIGH-DIMENSIONAL ESTIMATION.

Journal

Journal of the American Statistical Association, 112(517)

ISSN

0162-1459

Authors

Rossell, David

Telesca, Donatello

Publication Date

2017

DOI

10.1080/01621459.2015.1130634

Peer reviewed

HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2018 June 05.

Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(517): 254–265. doi:10.1080/01621459.2015.1130634.

NON-LOCAL PRIORS FOR HIGH-DIMENSIONAL ESTIMATION

DAVID ROSSELL[1] and **DONATELLO TELESCA**[2]

[1]UNIVERSITY OF WARWICK, DEPARTMENT OF STATISTICS

[2]UCLA, DEPARTMENT OF BIOSTATISTICS

Abstract

Jointly achieving parsimony and good predictive power in high dimensions is a main challenge in statistics. Non-local priors (NLPs) possess appealing properties for model choice, but their use for estimation has not been studied in detail. We show that for regular models NLP-based Bayesian model averaging (BMA) shrink spurious parameters either at fast polynomial or quasi-exponential rates as the sample size n increases, while non-spurious parameter estimates are not shrunk. We extend some results to linear models with dimension p growing with n. Coupled with our theoretical investigations, we outline the constructive representation of NLPs as mixtures of truncated distributions that enables simple posterior sampling and extending NLPs beyond previous proposals. Our results show notable high-dimensional estimation for linear models with $p \gg n$ at low computational cost. NLPs provided lower estimation error than benchmark and hyper-g priors, SCAD and LASSO in simulations, and in gene expression data achieved higher cross-validated R^2 with less predictors. Remarkably, these results were obtained without pre-screening variables. Our findings contribute to the debate of whether different priors should be used for estimation and model selection, showing that selection priors may actually be desirable for high-dimensional estimation.

Keywords

Model Selection; MCMC; Non Local Priors; Bayesian Model Averaging; Shrinkage

1. Introduction

Developing high-dimensional methods to balance parsimony and predictive power is a main challenge in statistics. Non-local priors (NLPs) are appealing for Bayesian model selection. Relative to local priors (LPs), NLPs discard spurious covariates faster as the sample size n grows, but preserve exponential rates to detect non-zero coefficients (Johnson and Rossell, 2010). When combined with Bayesian model averaging (BMA), this regularization has important consequences for estimation.

Denote the observations by $\mathbf{y}_n \in \mathcal{Y}_n$, where $\mathcal{Y}_n$ is the sample space. We entertain a collection of models M_k for $k = 1, \ldots, K$ with densities $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \boldsymbol{\phi}_k)$, where $\boldsymbol{\theta}_k \in \Theta_k \subseteq \Theta$ are parameters of interest and $\boldsymbol{\phi}_k \in \Phi$ is a fixed-dimension nuisance parameter. Let $p_k = \dim(\Theta_k)$ and without loss of generality let M_K be the full model within which $M_1, \ldots, M_{K-1}$ are nested ($\Theta_k \subset \Theta_K = \Theta$). To ease notation let $(\boldsymbol{\theta}, \boldsymbol{\phi}) = (\boldsymbol{\theta}_K, \boldsymbol{\phi}_K) \in \Theta \times \Phi$ be the parameters under M_K and $p = p_K = \dim(\Theta)$. A prior density $\pi(\boldsymbol{\theta}_k \mid M_k)$ for $\boldsymbol{\theta}_k \in \Theta_k$ under M_k is a NLP if it

converges to 0 as $\boldsymbol{\theta}_k$ approaches any value $\boldsymbol{\theta}_0$ consistent with a sub-model $M_{k'}$ (and a LP otherwise).

Definition 1

Let $\boldsymbol{\theta}_k \in \Theta_k$, an absolutely continuous measure with density $\pi(\boldsymbol{\theta}_k \mid M_k)$ is a non-local prior if $\lim_{\boldsymbol{\theta}_k \to \boldsymbol{\theta}_0} \pi(\boldsymbol{\theta}_k \mid M_k) = 0$ for any $\boldsymbol{\theta}_0 \in \Theta_k, \subset \Theta_k$, $k' \neq k$.

For precision we assume that intersections $\Theta_k \cap \Theta_{k'}$ have 0 Lebesgue measure and are included in some $M_{k''}$, $k'' \in \{1, \ldots, K\}$. As an example consider a Normal linear model $\mathbf{y}_n \sim N(X_n\boldsymbol{\theta}, \phi I)$ where X_n is an $n \times p$ matrix with p predictors, $\boldsymbol{\theta} \in \Theta = \mathbb{R}^p$ and $\phi \in \Phi = \mathbb{R}^+$. As we do not know which columns in X_n truly predict $\mathbf{y}_n$ we consider $K = 2^p$ models by setting elements in $\boldsymbol{\theta}$ to 0, *i.e.* $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k) = N(\mathbf{y}_n; X_{k,n}\boldsymbol{\theta}_k, \phi_k I)$ where $X_{k,n}$ is a subset of columns of X_n. We develop our analysis considering the following NLPs

$$\pi_M(\boldsymbol{\theta} \mid \phi_k, M_k) = \prod_{i \in M_k} \frac{\theta_i^2}{\tau\phi_k} N(\theta_i; 0, \tau\phi_k) \quad (1)$$

$$\pi_I(\boldsymbol{\theta} \mid \phi_k, M_k) = \prod_{i \in M_k} \frac{(\tau\phi_k)^{\frac{1}{2}}}{\sqrt{\pi}\theta_i^2} \exp\left\{-\frac{\tau\phi_k}{\theta_i^2}\right\} \quad (2)$$

$$\pi_E(\boldsymbol{\theta} \mid \phi_k, M_k) = \prod_{i \in M_k} \exp\left\{\sqrt{2} - \frac{\tau\phi_k}{\theta_i^2}\right\} N(\theta_i; 0, \tau\phi_k), \quad (3)$$

where $i \in M_k$ are the non-zero coefficients and π_M, π_I and π_E are called the product MOM, iMOM and eMOM priors (pMOM, piMOM and peMOM).

A motivation for considering K models is to learn which parameters are truly needed to improve estimation. Consider the usual BMA estimate

$$E(\boldsymbol{\theta} \mid \mathbf{y}_n) = \sum_{k=1}^{K} E(\boldsymbol{\theta} \mid M_k, \mathbf{y}_n) P(M_k \mid \mathbf{y}_n) \quad (4)$$

where $P(M_k \mid \mathbf{y}_n) \propto m_k(\mathbf{y}_n)P(M_k)$ and $m_k(\mathbf{y}_n) = \int\int f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k)\pi(\boldsymbol{\theta} \mid \phi_k, M_k)\pi(\phi_k \mid M_k)d\boldsymbol{\theta}_k d\phi_k$ is the integrated likelihood under M_k. BMA shrinks estimates by assigning small $P(M_k \mid \mathbf{y}_n)$ to unnecessarily complex models. The intuition is that NLPs assign even smaller weights. Let M_t be the smallest model such that $f_t(\mathbf{y}_n \mid \boldsymbol{\theta}_t, \phi_t)$ minimizes Kullback-Leibler

divergence (KL) to the data-generating density $f^*(\mathbf{y}_n)$ amongst all $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in \Theta \times \Phi$. For instance, in Normal linear regression this means minimizing the expected quadratic error $E((\mathbf{y}_n - X_n\boldsymbol{\theta})'(\mathbf{y}_n - X_n\boldsymbol{\theta}))$ with respect to $f^*(\mathbf{y}_n)$ (which may not be a linear model and include X_n when it is random). Under regular models with fixed $P(M_k)$ and p, if $\pi(\boldsymbol{\theta}_k \mid M_k)$ is a LP and $M_t \subset M_k$ then $P(M_k|\mathbf{y}_n) = O_p(n^{-\frac{1}{2}(p_k - p_t)})$ (Dawid, 1999). Models with spurious parameters are hence regularized at a slow polynomial rate, which we shall see implies $E(\theta_i \mid \mathbf{y}_n) = O_p(n^{-1})r$ (Section 2), where r depends on model prior probabilities. Any LP can be transformed into a NLP to achieve faster shrinkage, *e.g.* $E(\theta_i \mid \mathbf{y}_n) = O_p(n^{-2})r$ (pMOM) or $E(\theta_i|\mathbf{y}_n) = O_p(e^{-\sqrt{n}})r$ (peMOM, piMOM). We note that another strategy is to shrink via r, *e.g.* Castillo and Van der Vaart (2012) and Castillo et al. (2014) show that $P(M_k)$ decreasing fast enough with p_k achieve good posterior concentration. Martin and Walker (2013) propose a related empirical Bayes strategy. Yet another option is to consider the single model M_K and specify absolutely continuous shrinkage priors that induce posterior concentration (Bhattacharya et al., 2012). For a related review on penalized-likelihood strategies see Fan and Lv (2010).

In contrast our strategy is based upon faster $m_k(\mathbf{y}_n)$ rates, a data-dependent quantity. For Normal linear models with bounded $P(M_k)/P(M_t)$ Johnson and Rossell (2012) and Shin et al. (2015) showed that when $p = O(n^\alpha)$ or $p = O(e^{n^\alpha})$ (respectively) with $\alpha < 1$ and certain regularity conditions pertain one obtains $P(M_t|\mathbf{y}_n) \xrightarrow{P} 1$ when using certain NLPs and to 0 when using any LP, which from (4) implies the strong oracle property $E(\boldsymbol{\theta}|\mathbf{y}_n) \xrightarrow{P} E(\boldsymbol{\theta}|\mathbf{y}_n, M_t)$. We note that when sparse unbounded $P(M_k)/P(M_t)$ are used, consistency of $P(M_t \mid \mathbf{y}_n)$ may still be achieved with LPs, *e.g.* setting prior inclusion probabilities $O(p_K^{-\gamma})$ for $\gamma > 0$ as in Liang et al. (2013) or Narisetty and He (2014).

Our main contribution is considering parameter estimation under NLPs, as previous work focused on model selection. We characterize complexity penalties and BMA shrinkage for certain linear and asymptotically Normal models (Section 2). We also provide a fully general NLP representation from latent truncations (Section 3) that justifies NLPs intuitively and adds flexibility in prior choice. Suppose we wish to both estimate $\theta \in \mathbb{R}$ and test $M_1 : \theta = 0$ *vs.* $M_2 : \theta \neq 0$. Figure 1 (grey) shows a Cauchy(0, 0.25) prior expressing confidence that θ is close to 0, *e.g.* $P(|\theta| > 0.25) = 0.5$. Under this prior $P(\theta = 0 \mid \mathbf{y}_n) = 0$ and hence there is no BMA shrinkage. Instead we set $P(\theta = 0) = 0.5$ and, conditional on $\theta \neq 0$, a Cauchy(0,0.25) truncated to exclude $(-\lambda, \lambda)$, where λ is a practical significance threshold (Figure 1(top)). Truncated priors have been discussed before, *e.g.* Verdinelli and Wasserman (1996), Rousseau (2007). They encourage coherence between estimation and testing, but they cannot detect small but non-zero coefficients. Suppose that we set $\lambda \sim G(2.5, 10)$ to express our uncertainty about λ. Figure 1 (bottom) shows the marginal prior on θ after integrating out λ. It is a smooth version of the truncated Cauchy that goes to 0 as $\theta \to 0$, *i.e.* a NLP. Section 4 exploits this construction for posterior sampling. Finally, Section 5 studies finite-sample performance in simulations and gene expression data, in particular finding that BMA achieves lower quadratic error than the posterior modes used in Johnson and Rossell (2012).

2. Data-dependent shrinkage

We now show that NLPs induce a strong data-dependent shrinkage. To see why, note that any NLP can be written as $\pi(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid M_k) \propto d_k(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid M_k)$, where $d_k(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k) \to 0$ as $\boldsymbol{\theta}_k \to \boldsymbol{\theta}_0$ for any $\boldsymbol{\theta}_0 \in \Theta_{k'} \subset \Theta_k$ and $\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)$ is LP. NLPs are often expressed in this form but the representation is always possible since $\pi(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k|M_k) = \frac{\pi(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k|M_k)}{\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k|M_k)}\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k|M_k) = d_k(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k|M_k)$. Intuitively, $d_k(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)$ adds a penalty term that improves both selection and shrinkage via (4). The theorems below make the intuition rigorous. Proposition 1 shows that NLPs modify the marginal likelihood by a data-dependent term that converges to 0 for certain models containing spurious parameters. The result does not provide precise rates, but shows that under very general situations NLPs improve Bayesian regularization. Proposition 2 gives rates for posterior means and modes under a given M_k for finite p asymptotically Normal models and growing p linear models, whereas gives Proposition 3 Bayes factor and BMA rates.

We first discuss the needed regularity assumptions. Throughout we assume that $\pi(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k \mid M_k)$ is proper, $\pi(\boldsymbol{\phi}_k \mid M_k)$ is continuous and bounded for all $\boldsymbol{\phi}_k \in \Phi$, denote by $m_k(\mathbf{y}_n)$ the integrated likelihood under $\pi(\boldsymbol{\theta}_k \mid \boldsymbol{\phi}_k, M_k) = d_k(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)\pi^L(\boldsymbol{\theta}_k, \boldsymbol{\phi}_k)$ and by $m_k^L(\mathbf{y}_n) = \iint f_k(\mathbf{y}_n|\boldsymbol{\theta}_k, \phi_k)\pi^L(\boldsymbol{\theta}_k, \phi_k|M_k)d\boldsymbol{\theta}_k d\phi_k$ that under the corresponding LP. Assumptions A1–A5, B1–B4 are from Walker (1969) (W69, Supplementary Section 1) and guarantee asymptotic MLE normality and validity of second order log-likelihood expansions, *e.g.* including generalized linear models with finite p. A second set of assumptions for finite p models follows.

Conditions on finite-dimensional models

C1 Let $A \subset \Theta_k \times \Phi$ be such that $f_k(\mathbf{y}_n|\boldsymbol{\theta}_k^*, \phi_k^*)$ for any $(\boldsymbol{\theta}_k^*, \phi_k^*) \in A$ minimizes KL to $f^*(\mathbf{y}_n)$. For any $(\widetilde{\boldsymbol{\theta}}_k, \widetilde{\phi}_k) \notin A$ as $n \to \infty$

$$\frac{f_k(\mathbf{y}_n|\boldsymbol{\theta}_k^*, \phi_k^*)}{f_k(\mathbf{y}_n|\widetilde{\boldsymbol{\theta}}_k, \widetilde{\phi}_k)} \xrightarrow{\text{a.s.}} \infty.$$

C2 Let $\pi^L_{k,\tau}(\boldsymbol{\theta}_k, \phi_k) = N(\mathbf{0}; \tau\phi_k I)$. The ratio of marginal likelihoods $m^L_{k,\tau(1+\varepsilon)}(\mathbf{y}_n)/m^L_{k,\tau}(\mathbf{y}_n) \xrightarrow{\text{a.s.}} c \in (0, \infty)$ as $n \to \infty$, $\varepsilon \in (0, 1)$.

C3 Let $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ minimize KL$(f^*(\mathbf{y}_n), f_K(\boldsymbol{\theta}, \boldsymbol{\phi}))$ for $(\boldsymbol{\theta}, \boldsymbol{\phi}) \in (\Theta, \Phi)$. There is a unique M_t with smallest p_t such that $f_t(\mathbf{y}_n|\boldsymbol{\theta}_t^*, \phi_t^*) = f_K(\mathbf{y}_n|\boldsymbol{\theta}^*, \phi^*)$ and KL$(f_t(\mathbf{y}_n|\boldsymbol{\theta}_t^*, \phi_t^*), f_k(\mathbf{y}_n|\boldsymbol{\theta}_k, \phi_k)) > 0$, for any k such that $M_k \not\subset M_t$.

C4 In C3 $\boldsymbol{\phi}^*$ is fixed and $\theta_i^* = \theta_{0i}^* a_n$ for fixed θ_{0i}^* where either $a_n = 1$ or $\lim_{n\to\infty} a_n = 0$ with $a_n \gg n^{-1/2}$ (pMOM) or $a_n \gg n^{-1/4}$ (peMOM, piMOM).

C1 essentially gives MLE consistency and C2 a boundedness condition that guarantees $P(\boldsymbol{\theta}_k \in N(A) | \mathbf{y}_n . M_k) \xrightarrow{P} 1$ under a pMOM for a certain neighbourhood $N(A)$ of the KL-optimal parameter values, the key to ensure that $d_k(\boldsymbol{\theta}_k, \phi_k)$ acts as a penalty term. Redner (1981) gives general conditions for C1 that include even certain non-identifiable models. C2 is equivalent to the ratio of posterior densities under τ and $\tau(1 + \varepsilon)$ at an arbitrary $(\boldsymbol{\theta}_k, \phi_k)$ and converging to a constant, which holds under W69 or Conditions D1–D2 below (see proof of Proposition 1 for details). C3 assumes a unique smallest model $f_t(\mathbf{y}_n | \boldsymbol{\theta}_t^*, \phi_t^*)$ minimizing KL to $f^*(\mathbf{y}_n)$ and that there is no equivalent model $M_k \not\supset M_t$, *e.g.* for linear models no $M_k \not\supset M_t$ can have $p_k = p_t$ variables being perfectly collinear with $X_{t,n}$. C4 allows $\boldsymbol{\theta}^*$ to be either fixed or to vanishes at rates slower than $n^{-1/2}$ (pMOM) or $n^{-1/4}$ (peMOM, piMOM), to characterize the ability to estimate small signals. Finally, for linear models we consider the following.

Conditions on linear models of growing dimension

D1 Suppose $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k) = N(\mathbf{y}_n; X_{k,n}\boldsymbol{\theta}_k, \phi_k I)$, $\boldsymbol{\theta}_k \in \Theta_k$, $p_k = \dim(\boldsymbol{\theta}_k) = O(n^a)$ and $a < 1$.

D2 There are fixed $a, b, n_0 > 0$ such that $a < \frac{1}{n} l_1\left(X'_{k,n} X_{k,n}\right) < \frac{1}{n} l_k\left(X'_{k,n} X_{k,n}\right) < b$ for all $n > n_0$, where l_1, l_k are the smallest and largest eigenvalues of $X'_{k,n} X_{k,n}$.

D1 reflects the common practice that although $p \gg n$ one does not consider models with $p_k \geq n$, which lead to data interpolation. D2 guarantees strong MLE consistency (Lai et al., 1979) and implies that no considered model has perfectly collinear covariates, aligning with applied practice. For further discussion on eigenvalues see Chen and Chen (2008) and Narisetty and He (2014). We now state our first result. All proofs are in the Supplementary Material.

Proposition 1—Let $m_k(\mathbf{y}_n)$, $m_k^L(\mathbf{y}_n)$ be as above.

i. We have: $m_k(\mathbf{y}_n) = m_k^L(\mathbf{y}_n) g_k(\mathbf{y}_n)$, where

$$g_k(\mathbf{y}_n) = \iint d_k(\boldsymbol{\theta}_k, \phi_k) \pi^L(\boldsymbol{\theta}_k, \phi_k | \mathbf{y}_n) d\boldsymbol{\theta}_k d\phi_k$$

ii. Assume $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k)$ with finite p_k satisfies C1 under a peMOM or piMOM prior or C2 under a pMOM prior for some A. If $A = \left\{(\boldsymbol{\theta}_k^*, \phi_k^*)\right\}$ is a singleton (identifiable models), then $g_k(\mathbf{y}_n) \xrightarrow{P} d_k(\boldsymbol{\theta}_k^*, \phi_k^*)$. For any A, if

$$f^*(\mathbf{y}_n) = f_t(\mathbf{y}_n | \boldsymbol{\theta}_t^*, \phi_t^*)$$

for some $t \in \{1, \ldots, K\}$, then $g_k(\mathbf{y}_n) \xrightarrow{P} 0$ when $M_t \subset M_k$, $k \neq t$ and

$$g_k(\mathbf{y}_n) \xrightarrow{P} c > 0$$

when $M_k \subseteq M_t$.

iii. Let $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k) = N(\mathbf{y}_n; X_{n,k}\boldsymbol{\theta}_k, \phi_k I)$, with growing p_k, satisfy D1–D2. Let $(\boldsymbol{\theta}_k^*, \phi_k^*)$ minimize KL to $f^*(\mathbf{y}_n)$ with $Var(\mathbf{y}_n - X_{k,n}\boldsymbol{\theta}_k^*) = \phi_k^* < \infty$ and $\pi(\phi_k^* | M_k) > 0$. Then $g_k(\mathbf{y}_n) \xrightarrow{P} d_k(\boldsymbol{\theta}_k^*, \phi_k^*)$ and $d_k(\mathbf{m}_{k,n}, \phi_k^*) \xrightarrow{\text{a.s.}} d_k(\boldsymbol{\theta}_k^*, \phi_k^*)$, where $\mathbf{m}_{k,n} = S_{k,n}^{-1} X_{k,n}' \mathbf{y}_n$, $S_{k,n} = X_{k,n}' X_{k,n} + \tau^{-1} I$. Further, if $f^*(\mathbf{y}_n) = N(\mathbf{y}_n; X_{t,n}\boldsymbol{\theta}_t^*, \phi_t^*)$ then $g_k(\mathbf{y}_n) \xrightarrow{P} c$ with $c = 0$ when either $M_t \subset M_k$ or $M_t \not\subset M_k$ but a column in $(X_{k,n}', X_{k,n})^{-1} X_{k,n}' X_{t,n}$ converges to zero. Else, $c > 0$.

That is, even when the data-generating $f^*(\mathbf{y}_n)$ does not belong to the set of considered models, $g_k(\mathbf{y}_n)$ converges to 0 for certain M_k containing spurious parameters, *e.g.* for linear models when either $M_t \subset M_k$ or $M_t \not\subset M_k$ but some columns in $X_{k,n}$ are uncorrelated with $X_{t,n}$ given $X_{k,n} \cap X_{t,n}$. Propositions 2–3 give rates for the case when $f^*(\mathbf{y}_n) = f_t(\mathbf{y}_n | \boldsymbol{\theta}_t^*, \phi_t^*)$.

Proposition 2—Let $(\hat{\boldsymbol{\theta}}_k, \hat{\phi}_k)$ be the unique MLE and $f_k(\mathbf{y}_n | \boldsymbol{\theta}_k^*, \phi_k^*)$ minimize KL to the data-generating $f_t(\mathbf{y}_n | \boldsymbol{\theta}_t^*, \phi_t^*)$ for $(\boldsymbol{\theta}_k^*, \phi_k^*) \in \Theta_k \times \Phi$. Assume C3–C4 are satisfied.

i. Let $f_k(\mathbf{y} \mid \boldsymbol{\theta}_k, \phi_k)$ with fixed p_k satisfy W69 and $\tilde{\boldsymbol{\theta}}_k$ be the posterior mode, with $sign(\tilde{\theta}_{ki}) = sign(\hat{\theta}_{ki})$ for $i = 1, \ldots, p_k$ under a pMOM, peMOM or piMOM prior. If $\theta_{ki}^* \neq 0$ is fixed then $n(\tilde{\theta}_{ki} - \hat{\theta}_{ki}) \xrightarrow{P} c$ for some $0 < c < \infty$. If $\theta_{ki}^* = \theta_{0i}^* a_n \neq 0$ with $a_n \to 0$ as in C4 then $\tilde{\theta}_i - \hat{\theta}_{ki} = O_p(1/na_n))$ for pMOM and $\tilde{\theta}_i - \hat{\theta}_{ki} = O_p(1/na_n^3))$ for peMOM, piMOM. If $\theta_{ki}^* = 0$ then $n^2(\tilde{\theta}_{ki} - \hat{\theta}_{ki})^2 \xrightarrow{P} c$ for pMOM and $n\tilde{\theta}_{ki}^4 \xrightarrow{P} c$ for peMOM, piMOM with $0 < c < \infty$. Further, any other posterior mode is $O_p(n^{-1/2})$ (pMOM) or $O_p(n^{-1/4})$ (peMOM, piMOM).

ii. Under the conditions in (i) $E(\theta_{ki} | M_k, \mathbf{y}_n) = \hat{\theta}_{ki} + O_p(n^{-1/2}) = \theta_{ki}^* + O_p(n^{-1/2})$ for pMOM and $\hat{\theta}_{ki} + O_p(n^{-1/4}) = \theta_{ki}^* + O_p(n^{-1/4})$ for peMOM/piMOM.

iii. Let $f_k(\mathbf{y}_n \mid \boldsymbol{\theta}_k, \phi_k) = N(\mathbf{y}_n; X_{n,k}\boldsymbol{\theta}_k, \phi_k I)$ satisfy D1–D2 with diagonal $X_{n,k}' X_{n,k}$. Then the rates in (i)–(ii) remain valid.

We note that given that there is a prior mode in each of the 2^{pk} quadrants (combination of signs of θ_{ki}) there always exists a posterior mode $\tilde{\boldsymbol{\theta}}_k$ satisfying the sign conditions in (i). Further, for elliptical log-likelihoods given that the pMOM, peMOM and piMOM priors have independent symmetric components the global posterior mode is guaranteed to occur in the same quadrant as $\hat{\boldsymbol{\theta}}_k$. Part (i) first characterizes the behaviour of this dominant mode and subsequently the behaviour of all other modes. Conditional on M_k, spurious parameter

estimates converge to 0 at $n^{-1/2}$ (pMOM) or $n^{-1/4}$ (peMOM,piMOM). Vanishing $\theta_i^* \neq 0$ are captured as long as $\theta_i^* \gg n^{-1/2}$ (pMOM) or $\theta_i^* \gg n^{-1/4}$ (peMOM, piMOM). This holds for fixed p_k or linear models with growing p_k and diagonal $X'_{n,k}X_{n,k}$. We leave further extensions as future work.

Proposition 3 shows that weighting these estimates with $P(M_k \mid \mathbf{y}_n)$ gives a strong selective shrinkage. We denote $SSR_0 = \sum_{\theta_i^*=0}(E(\theta_i|\mathbf{y}_n) - \theta_i^*)^2$, $SSR_1 = \sum_{\theta_i^* \neq 0}(E(\boldsymbol{\theta}_i|\mathbf{y}_n) - \theta_i^*)^2$, $p_0 = \sum_{i=1}^{p} I(\theta_i^* = 0)$, $p_1 = p - p_0$ and let $E_{\boldsymbol{\theta}^*}(SSR_0) = \int SSR_0 f(\mathbf{y}_n|\boldsymbol{\theta}^*, \phi^*)d\mathbf{y}_n$ be the mean under the data-generating $f(\mathbf{y}_n \mid \boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$.

Proposition 3—Let $E(\theta_i \mid \mathbf{y}_n)$ be as in (4), M_t the data-generating model, $BF_{kt} = m_k(\mathbf{y})/m_t(\mathbf{y})$ and a_n as in C4. Assume that $P(M_k)/P(M_t) = o(n^{(pk-pt)})$ for $M_t \subset M_k$.

i. Let all M_k satisfy W69, C3 and p be fixed. If $M_t \not\subset M_k$, then $BF_{kt} = O_p(e^{-n})$ under a pMOM, peMOM or piMOM prior if $\theta_{ti}^* \neq 0$ are fixed and $BF_{kt} = O_p(e^{-a_n^2 n})$ if $\theta_{ti}^* = \theta_{0t}^* a_n$. If $M_t \subset M_k$ then $BF_{kt} = O_p(n^{-\frac{3}{2}(p_k - p_t)})$ under a pMOM prior and $BF_{kt} = O_p(e^{-\sqrt{n}})$ under peMOM or piMOM.

ii. Under the conditions in (i) let a_n be as in C4 and $r = max_k P(M_k)/P(M_t)$ where $p_k = p_t + 1$, $M_t \subset M_k$. Then the posterior means and sums of squared errors satisfy

	pMOM		peMOM-piMOM	
	$E(\boldsymbol{\theta}_i \mid \mathbf{y}_n)$	SSR	$E(\boldsymbol{\theta}_i \mid \mathbf{y}_n)$	SSR
$\theta_i^* \neq 0$	$\theta_i^* + O_p(n^{-1/2})$	$O_p(p_1 n^{-1})$	$\theta_i^* + O_p(n^{-1/2})$	$O_p(p_1 n^{-1})$
$\theta_i^* = \theta_{0t}^* a_n$	$\theta_i^* + O_p(n^{-1/2})$	$O_p(p_1 n^{-1})$	$\theta_i^* + O_p(n^{-1/4})$	$O_p(p_1 n^{-1/2})$
$\theta_i^* = 0$	$rO_p(n^{-2})$	$O_p(p_0 r^2 n^{-4})$	$rO_p(e^{-\sqrt{n}})$	$O_p(p_0 r^2 e^{-\sqrt{n}})$

iii. Let $\mathbf{y}_n \sim N(X_{n,k}\boldsymbol{\theta}_k, \boldsymbol{\phi}_k I)$ satisfy D1–D2 with diagonal $X'_n X_n$ and known $\boldsymbol{\phi}$. Let ε, $\tilde{\varepsilon} > 0$ be arbitrarily small constants and assume that $P(\theta_1 \neq 0, \ldots, \theta_p \neq 0)$ is exchangeable with $r = P(\delta_i = 1)/P(\delta_i = 0)$. Then

	pMOM		peMOM-piMOM	
	$E(\theta_i \mid \mathbf{y}_n, \phi)$	$E_{\theta^*}(SSR)$	$E(\theta_i \mid \mathbf{y}_n, \phi)$	$E_{\theta^*}(SSR)$
$\theta_i^* \neq 0$	$\theta_i^* + O_p(n^{-1/2})$	$O(p_1/n^{1-\varepsilon})$	$\theta_i^* + O_p(n^{-1/2})$	$O_p(p_1/n^{1-\varepsilon})$
$\theta_i^* = \theta_{0i}^* a_n$	$\theta_i^* + O_p(n^{-1/2})$	$O(p_1/n^{1-\varepsilon})$	$\theta_i^* + O_p(n^{-1/4})$	$O(p_1/n^{\frac{1}{2}-\varepsilon})$
$\theta_i^* = 0$	$rO_p(n^{-2})$	$O_p(p_0 r^2/n^{4-\varepsilon})$	$rO_p(e^{-\sqrt{n}})$	$O(p_0 r^2 e^{-n^{1/2-\varepsilon}})$

where the results for $\theta_i^* \neq 0$ and $\theta_i^* = \theta_{0i}^* a_n$ hold as long as $r \gg e^{-n^{\tilde{\varepsilon}}}$ and the result for $\theta_i^* = 0$ holds for any r.

BMA estimates for active coefficients are $O_p(1/\sqrt{n})$ of their true value ($O_p(n^{-1/4})$ for vanishing θ_i^* under peMOM or piMOM), but inactive coefficients estimates are shrunk at $rO_p(n^{-2})$ or $rO_p(e^{-\sqrt{n}})$ (to be compared with $rO_p(n^{-1})$ under the corresponding LPs) where r are the prior inclusion odds. The condition $P(M_k)/P(M_t) = o(n^{pk-pt})$ for $M_t \subset M_k$ ensures that complex models are not favoured a priori (usually $P(M_k)/P(M_t) = O(1)$). The condition $r \gg e^{-n^{\tilde{\varepsilon}}}$ in Part (iii) prevents the prior from favouring overly sparse solutions. For instance, a Beta-Binomial(1, l) prior on the model size gives $r = 1/l$, hence any fixed finite l satisfies $r \gg e^{-n^{\tilde{\varepsilon}}}$. Suppose that we set $l = p$, then $r \gg e^{-n^{\tilde{\varepsilon}}}$ is satisfied as long as $p = O(e^{n^{\alpha}})$ for some $\alpha < 1$.

3. Non-local priors as truncation mixtures

We establish a correspondence between NLPs and truncation mixtures. Our discussion is conditional on M_k, hence for simplicity we omit ϕ and denote $\pi(\theta) = \pi(\theta \mid M_k)$, $p = \dim(\Theta_k)$.

3.1. Equivalence between NLPs and truncation mixtures

We show that truncation mixtures define valid NLPs, and subsequently that any NLP may be represented in this manner. Given that the representation is not unique, we give two constructions and discuss their merits. Let $\pi^L(\theta)$ be an arbitrary LP and $\lambda \in \mathbb{R}^+$ a latent truncation.

Proposition 4—Define $\pi(\theta \mid \lambda) \propto \pi^L(\theta) I(d(\theta) > \lambda)$, where $\lim_{\theta \to \theta_0} d(\theta) = 0$ for any $\theta_0 \in \Theta_k{}' \subset \Theta_k$, and $\pi^L(\theta)$ is bounded in a neighborhood of θ_0. Let $\pi(\lambda)$ be a marginal prior for λ placing no probability mass at $\lambda = 0$. Then $\pi(\theta) = \int \pi(\theta \mid \lambda)\pi(\lambda)d\lambda$ defines a NLP.

Corollary 5—Assume that $d(\boldsymbol{\theta}) = \prod_{i=1}^{p} d_i(\theta_i)$. Let $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) \propto \pi^L(\boldsymbol{\theta}) \prod_{i=1}^{p} I(d_i(\theta_i) > \lambda_i)$ where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ have an absolutely continuous prior $\pi(\boldsymbol{\lambda})$. Then $\int \pi(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}) d\boldsymbol{\lambda}$ is a NLP.

Example 1—Consider $\mathbf{y}_n \sim N(X\boldsymbol{\theta}, \phi I)$, where $\boldsymbol{\theta} \in \mathbb{R}^p$, ϕ is known and I is the $n \times n$ identity matrix. We define a NLP for $\boldsymbol{\theta}$ with a single truncation point with $\pi(\boldsymbol{\theta}|\lambda) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) I(\prod_{i=1}^{p} \theta_i^2 > \lambda)$ and some $\pi(\lambda)$, e.g. Gamma or Inverse Gamma. Obviously, the choice of $\pi(\lambda)$ affects $\pi(\boldsymbol{\theta})$ (Section 3.2). An alternative prior is

$$\pi(\boldsymbol{\theta}|\lambda_1, \ldots, \lambda_p) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) \prod_{i=1}^{p} I(\theta_i^2 > \lambda_i),$$

giving marginal independence when $\pi(\lambda_1, \ldots, \lambda_p)$ has independent components.

We address the reverse question: given any NLP, a truncation representation is always possible.

Proposition 6—Let $\pi(\boldsymbol{\theta}) \propto d(\boldsymbol{\theta}) \pi^L(\boldsymbol{\theta})$ be a NLP and denote $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda)$, where $P_u(\cdot)$ is the probability under $\pi^L(\boldsymbol{\theta})$. Then $\pi(\boldsymbol{\theta})$ is the marginal prior associated to $\pi(\boldsymbol{\theta} \mid \lambda) \propto \pi^L(\boldsymbol{\theta}) I(d(\boldsymbol{\theta}) > \lambda)$ and $\pi(\lambda) = h(\lambda)/E_u(d(\boldsymbol{\theta})) \propto h(\lambda)$, where $E_u(\cdot)$ is the expectation with respect to $\pi^L(\boldsymbol{\theta})$.

Corollary 7—Let $\pi(\boldsymbol{\theta}) \propto \pi^L(\boldsymbol{\theta}) \prod_{i=1}^{p} d_i(\theta_i)$ be a NLP,

$$h(\boldsymbol{\lambda}) = P_u(d_1(\theta_1) > \lambda_1, \ldots, d_p(\theta_p) > \lambda_p)$$

and assume that $\int h(\boldsymbol{\lambda}) d\boldsymbol{\lambda} < \infty$. Then $\pi(\boldsymbol{\theta})$ is the marginal prior associated to $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda}) \propto \pi^L(\boldsymbol{\theta}) \prod_{i=1}^{p} I(\theta_i > \lambda_i)$ and $\pi(\boldsymbol{\lambda}) \propto h(\boldsymbol{\lambda})$.

Corollary 7 adds latent variables but greatly facilitates sampling. The condition $\int h(\boldsymbol{\lambda}) d\boldsymbol{\lambda} < \infty$ is guaranteed when $\pi^L(\boldsymbol{\theta})$ has independent components (apply Proposition 6 to each θ_i).

Example 2—The pMOM prior with $d(\boldsymbol{\theta}) = \prod_{i=1}^{p} \theta_i^2$, $\pi^L(\boldsymbol{\theta}) = N(\boldsymbol{\theta}, \mathbf{0}, \tau I)$ can be represented as $\pi(\boldsymbol{\theta}|\lambda) \propto N(\boldsymbol{\theta}; \mathbf{0}, \tau I) I(\prod_{i=1}^{p} \theta_i^2 > \lambda)$ and

$$\pi(\lambda) = \frac{P(\prod_{i=1}^{p} \theta_i^2/\tau > \lambda/\tau^p)}{E_u(\prod_{i=1}^{p} \theta_i^2)} = \frac{h(\lambda/\tau^p)}{\tau^p},$$

where h(·) is the survival function for a product of independent chi-square random variables with 1 degree of freedom (Springer and Thompson, 1970). Prior draws are obtained by

1. Draw $u \sim Unif(0, 1)$. Set $\lambda = P^{-1}(u)$, where $P(u) = P_{\pi}(\lambda \leq u)$ is the cdf associated to $\pi(\lambda)$.
2. Draw $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau I) I(d(\boldsymbol{\theta}) > \lambda)$.

As drawbacks, $P(u)$ requires Meijer G-functions and is cumbersome to evaluate for large p and sampling from a multivariate Normal with truncation region $\prod_{i=1}^{p} \theta_i^2 > \lambda$ is nontrivial. Corollary 7 gives an alternative. Let $P(u) = P(\lambda < u)$ be the cdf associated to $\pi(\lambda) = \frac{h(\lambda/\tau)}{\tau}$ where $h(\cdot)$ is the survival of a χ_1^2. For $i = 1, \ldots, p$, draw $u_i \sim Unif(0, 1)$, set $\lambda_i = P^{-1}(u_i)$ and draw $\theta_i \sim N(0, \tau) I(\theta_i > |\lambda_i|)$. The function $P^{-1}(\cdot)$ can be tabulated and quickly evaluated, rendering efficient computations. Supplementary Figure 1 shows 100,000 draws from pMOM priors with $\tau = 5$.

3.2. Deriving NLP properties for a given mixture

We show how two important characteristics of a NLP functional form, the penalty and tails, depend on the chosen truncation. We distinguish whether a single or multiple truncation variables are used.

Proposition 8—Let $\pi(\boldsymbol{\theta})$ be the marginal of $\pi(\boldsymbol{\theta}, \lambda) = \frac{\pi^L(\boldsymbol{\theta})}{h(\lambda)} \pi(\lambda) \prod_{i=1}^{p} I(d(\theta_i) > \lambda)$, where $h(\lambda) = P_u(d(\theta_1) > \lambda, \ldots, d(\theta_p) > \lambda)$ and $\lambda \in \mathbb{R}^+$ with $P(\lambda = 0) = 0$. Let $d_{min}(\boldsymbol{\theta}) = min\{d(\theta_1), \ldots, d(\theta_p)\}$.

i. Consider any sequence $\{\boldsymbol{\theta}^{(m)}\}_{m \geq 1}$ such that $\lim_{m \to \infty} d_{min}(\boldsymbol{\theta}^{(m)}) = 0$ Then

$$\lim_{m \to \infty} \frac{\pi(\boldsymbol{\theta}^{(m)})}{\pi^L(\boldsymbol{\theta}^{(m)}) d_{min}(\boldsymbol{\theta}^{(m)}) \pi(\lambda^{(m)})} = 1,$$

for some $\lambda^{(m)} \in (0, d_{min}(\boldsymbol{\theta}^{(m)}))$. If $\pi(\lambda) = ch(\lambda)$ then $\lim_{m \to \infty} \pi(\lambda^{(m)}) = c \in (0, \infty)$.

ii. Let $\{\boldsymbol{\theta}^{(m)}\}_{m \geq 1}$ be any sequence such that $\lim_{m \to \infty} d(\boldsymbol{\theta}^{(m)}) = \infty$. Then $\lim_{m \to \infty} \pi(\boldsymbol{\theta}^{(m)}) / \pi^L(\boldsymbol{\theta}^{(m)}) = c$, where $c > 0$ is either a positive constant or ∞. In particular, if $\int \frac{\pi(\lambda)}{h(\lambda)} d\lambda < \infty$ then $c < \infty$.

Property (i) is important as Bayes factor rates depend on the penalty, which we see is given by the smallest $d(\theta_1), \ldots, d(\theta_p)$. Property (ii) shows that $\pi(\boldsymbol{\theta})$ inherits its tail behavior from $\pi^L(\boldsymbol{\theta})$. Corollary 9 is an extension to multiple truncations.

Corollary 9—Let $\pi(\boldsymbol{\theta})$ be the marginal NLP for $\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \frac{\pi^L(\boldsymbol{\theta})}{h(\boldsymbol{\lambda})} \prod_{i=1}^{p} I(d_i(\theta_i) > \lambda_i) \pi_i(\lambda_i)$, where $h(\boldsymbol{\lambda}) = P_u(d_1(\theta_1) > \lambda_1, \ldots, d_p(\theta_p) > \lambda_p)$ under $\pi^L(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\lambda})$ is absolutely continuous.

i. Let $\{\boldsymbol{\theta}^{(m)}\}_{m\geq 1}$ such that $\lim_{m\to\infty} d_i(\theta_i^{(m)}) = 0$ for i = 1, …, p. Then for some $\lambda_i^{(m)} \in (0, d(\theta_i))$, $\lim_{m\to\infty} \pi(\boldsymbol{\theta}^{(m)})/(\pi^L(\boldsymbol{\theta}^{(m)})\pi(\boldsymbol{\lambda}^{(m)})\prod_{i=1}^{p} d_i(\theta_i^{(m)})) = 1$.

ii. Let $\{\boldsymbol{\theta}^{(m)}\}_{m\geq 1}$ such that $\lim_{m\to\infty} d_i(\theta_i^{(m)}) = \infty$ for $i = 1, \ldots p$. Then $\lim_{m\to\infty} \pi(\boldsymbol{\theta}^{(m)})/\pi^L(\boldsymbol{\theta}^{(m)}) = c > 0$ where $c \in \mathbb{R}^+ \cup \{\infty\}$. In particular, if $E(h(\boldsymbol{\lambda})^{-1}) < \infty$ under $\pi(\boldsymbol{\lambda})$, then $c < \infty$.

That is, multiple independent truncation variables give a multiplicative penalty $\prod_{i=1}^{p} d_i(\theta_i)$ and tails are at least as thick as those of $\pi^L(\boldsymbol{\theta})$. Once a functional form for $\pi(\boldsymbol{\theta})$ is chosen, we need to set its parameters. Although the asymptotic rates (Section 2) hold for any fixed parameters, their value can be relevant in finite samples. Given that posterior inference depends solely on the marginal prior $\pi(\boldsymbol{\theta})$, whenever possible we recommend eliciting $\pi(\boldsymbol{\theta})$ directly. For instance, Johnson and Rossell (2010) defined practical significance in linear regression as signal-to-noise ratios $|\theta_i|/\sqrt{\phi} > 0.2$, and gave default τ assigning $P(|\theta_i|/\sqrt{\phi} > 0.2) = 0.99$. Rossell et al. (2013) found analogous τ for probit regression, and also considered learning τ either via a hyper-prior or minimizing posterior predictive loss (Gelfand and Ghosh, 1998). Consonni and La Rocca (2010) devised objective Bayes strategies. Yet another possibility is to match the unit information prior e.g. setting $V(\theta_i/\sqrt{\phi}) = 1$ which can be regarded as minimally informative (in fact prior *e.g.* $V(\theta_i/\sqrt{\phi}) = 1.074$ for the MOM default $\tau = 0.358$). When $\pi(\boldsymbol{\theta})$ is not in closed-form prior elicitation depends both on τ and $\pi(\boldsymbol{\lambda})$, but prior draws can be used to estimate $P(|\theta_i|/\sqrt{\phi} > t)$ for any t. An analytical alternative is to set $\pi(\boldsymbol{\lambda})$ so that $E(\lambda) = d(\theta_i, \phi)$when $\theta_i/\sqrt{\phi} = t$, *i.e.* $E(\lambda)$ matches a practical relevance threshold. For instance, for $t = 0.2$ and $\pi(\lambda) \sim \text{IG}(a, b)$ under the MOM prior we would set $E(\lambda) = b/(a-1) = 0.2^2/\tau$, and under the eMOM prior $b/(a-1) = e^{\sqrt{2}-\tau/0.2^2}$. Both expressions illustrate the dependence between τ and $\pi(\lambda)$. Here we use default τ (Section 5), but as discussed other strategies are possible.

4. Posterior sampling

We use the latent truncation characterization to derive posterior sampling algorithms. Section 4.1 provides two Gibbs algorithms to sample from arbitrary posteriors, and Section 4.2 adapts them to linear models. Sampling is conditional on a given M_k, hence we drop M_k to keep notation simple.

4.1. General algorithm

First consider a NLP defined by a single latent truncation, *i.e.* $\pi(\boldsymbol{\theta}|\lambda) = \pi^L(\boldsymbol{\theta})\text{I}(d(\boldsymbol{\theta}) > \lambda)/h(\lambda)$, where $h(\lambda) = P_u(d(\boldsymbol{\theta}) > \lambda)$ and $\pi(\lambda)$ a prior on $\lambda \in \mathbb{R}^+$. The joint posterior is

$$\pi(\boldsymbol{\theta}, \lambda \mid \mathbf{y}_n) \propto f(\mathbf{y}_n \mid \boldsymbol{\theta}) \frac{\pi^L(\boldsymbol{\theta}) \mathrm{I}(d(\boldsymbol{\theta}) > \lambda)}{h(\lambda)} \pi(\lambda). \quad (5)$$

Sampling from $\pi(\boldsymbol{\theta} \mid \mathbf{y}_n)$ directly is challenging as it is highly multi-modal, but straightforward algebra gives the following k^{th} Gibbs iteration to sample from $\pi(\boldsymbol{\theta}, \lambda \mid \mathbf{y}_n)$.

Algorithm 1. Gibbs sampling with a single truncation

1. Draw $\lambda^{(k)} \sim \pi(\lambda \mid \mathbf{y}_n, \boldsymbol{\theta}^{(k-1)}) \propto \mathrm{I}(d(\boldsymbol{\theta}) > \lambda)\pi(\lambda)/h(\lambda)$. When $\pi(\lambda) \propto h(\lambda)$ as in Proposition 6, $\lambda^{(k)} \sim \mathrm{Unif}(0, d(\boldsymbol{\theta}^{(k-1)}))$.
2. Draw $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y}_n, \lambda^{(k)}) \propto \pi^L(\boldsymbol{\theta} \mid \mathbf{y}_n)\mathrm{I}(d(\boldsymbol{\theta}) > \lambda^{(k)})$.

That is, $\lambda^{(k)}$ is sampled from a univariate distribution that reduces to a uniform when setting $\pi(\lambda) \propto h(\lambda)$, and $\boldsymbol{\theta}^{(k)}$ from a truncated version of $\pi^L()$, which may be a LP that allows posterior sampling. As a difficulty, the truncation region $\{\boldsymbol{\theta}: d(\boldsymbol{\theta}) > \lambda^{(k)}\}$ is non-linear and non-convex so that jointly sampling $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ may be challenging. One may apply a Gibbs step to each element in $\theta_1, \ldots, \theta_p$ sequentially, which only requires univariate truncated draws from $\pi^L(\cdot)$, but the mixing of the chain may suffer. The multiple truncation representation in Corollary 7 provides a convenient alternative. Consider $\pi(\boldsymbol{\theta} \mid \lambda) = \pi^L(\boldsymbol{\theta}) \prod_{i=1}^{p} \mathrm{I}(d_i(\theta_i) > \lambda_i)\pi(\boldsymbol{\lambda})/h(\boldsymbol{\lambda})$, where $h(\boldsymbol{\lambda}) = P_U(d_1(\theta_1) > \lambda_1, \ldots d_p(\theta_p) > \lambda_p)$.

The following steps define the k Gibbs iteration:

Algorithm 2. Gibbs sampling with multiple truncations

1. Draw $\boldsymbol{\lambda}^{(k)} \sim \pi(\boldsymbol{\lambda} \mid \mathbf{y}_n, \boldsymbol{\theta}^{(k-1)}) = \prod_{i=1}^{p} \mathrm{Unif}(\lambda_i; 0, d_i(\theta_i)) \frac{\pi(\boldsymbol{\lambda})}{h(\boldsymbol{\lambda})}$. If $\pi(\boldsymbol{\lambda}) \propto h(\boldsymbol{\lambda})$ as in Corollary 7, $\lambda_i^{(k)} \sim \mathrm{Unif}(0, d_i(\theta_i))$.
2. Draw $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta} \mid \mathbf{y}_n, \boldsymbol{\lambda}^{(k)}) \propto \pi^L(\boldsymbol{\theta} \mid \mathbf{y}_n) \prod_{i=1}^{p} \mathrm{I}(d_i(\theta_i) > \lambda_i^{(k)})$.

Now the truncation region in Step 2 is defined by hyper-rectangles, which facilitates sampling. As in Algorithm 1, by setting the prior conveniently Step 1 avoids evaluating $\pi(\boldsymbol{\lambda})$ and $h(\boldsymbol{\lambda})$.

4.2. Linear models

We adapt Algorithm 2 to a linear regression $\mathbf{y}_n \sim N(X\boldsymbol{\theta}, \phi I)$ with the three priors in (1)-(3). We set the prior $\phi \sim \mathrm{IG}(a_\phi/2, b_\phi/2)$. For all three priors, Step 2 in Algorithm 2 samples from a multivariate Normal with rectangular truncation around $\mathbf{0}$, for which we developed an efficient algorithm. Kotecha and Djuric (1999) and Rodriguez-Yam et al. (2004) proposed Gibbs after orthogonalization strategies that result in low serial correlation, which Wilhelm and Manjunath (2010) implemented in the R package `tmvtnorm` for restrictions $l \leq \theta_i \leq u$. Here we require sampling under $d_l(\theta_i) \geq l$, a non-convex region. Our adapted algorithm is in Supplementary Section 3 and implemented in R package `mombf`. An important property is that the algorithm produces independent samples when the posterior probability of the truncation region becomes negligible. Since NLPs only assign high posterior probability to a

model when the posterior for non-zero coefficients is well shifted from the origin, the truncation region is indeed often negligible. We outline the algorithm separately for each prior.

4.2.1. pMOM prior—Straightforward algebra gives the full conditional posteriors

$$\pi(\boldsymbol{\theta}|\phi,\mathbf{y}_n) \propto \left(\prod_{i=1}^{p}\theta_i^2\right) N(\boldsymbol{\theta};\mathbf{m},\phi S^{-1}) \qquad (6)$$
$$\pi(\phi|\boldsymbol{\theta},\mathbf{y}_n) = \mathrm{IG}\left(\frac{a_\phi+n+3p}{2}, \frac{b_\phi+s_R^2+\boldsymbol{\theta}'\boldsymbol{\theta}/\tau}{2}\right),$$

where $S = X'X + \boldsymbol{\tau}^{-1}I$, $\mathbf{m} = S^{-1}X'\mathbf{y}_n$ and $s_R^2 = (\mathbf{y}_n - X\boldsymbol{\theta})'(\mathbf{y}_n - X\boldsymbol{\theta})$ is the sum of squared residuals. Corollary 7 represents the pMOM prior in (1) as

$$\pi(\boldsymbol{\theta}|\phi,\boldsymbol{\lambda}) = N(\boldsymbol{\theta};\mathbf{0},\tau\phi I)\prod_{i=1}^{p}\mathrm{I}\left(\frac{\theta_i^2}{\tau\phi} > \lambda_i\right)\frac{1}{h(\lambda_i)} \qquad (7)$$

marginalized with respect to $\pi(\lambda_i) = h(\lambda_i) = P\left(\frac{\theta_i^2}{\tau\phi} > \lambda_i|\phi\right)$, where $h(\cdot)$ is the survival of a chi-square with 1 degree of freedom. Algorithm 2 and simple algebra give the k^{th} Gibbs iteration

1. $\phi^{(k)} \sim \mathrm{IG}\left(\frac{a_\phi+n+3p}{2}, \frac{b_\phi+s_R^2+(\boldsymbol{\theta}^{(k-1)})'\boldsymbol{\theta}^{(k-1)}/\tau}{2}\right)$
2. $\boldsymbol{\lambda}^{(k)} \sim \pi(\boldsymbol{\lambda}|\boldsymbol{\theta}^{(k-1)},\phi^{(k)},\mathbf{y}_n) = \prod_{i=1}^{p}\mathrm{I}\left(\frac{(\theta_i^{(k-1)})^2}{\tau\phi^{(k)}} > \lambda_i\right)$
3. $\boldsymbol{\theta}^{(k)} \sim \pi(\boldsymbol{\theta}|\boldsymbol{\lambda}^{(k)},\phi^{(k)},\mathbf{y}_n) = N(\boldsymbol{\theta};\mathbf{m},\phi^{(k)}S^{-1})\prod_{i=1}^{p} I\left(\frac{\theta_i^2}{\tau\phi^{(k)}} > \lambda_i\right)$.

Step 1 samples unconditionally on $\boldsymbol{\lambda}$, so that no efficiency is lost for introducing these latent variables. Step 3 requires truncated multivariate Normal draws.

4.2.2. piMOM prior—We assume $\dim(\Theta) < n$. The full conditional posteriors are

$$\pi(\boldsymbol{\theta}|\phi,\mathbf{y}_n) \propto \left(\prod_{i=1}^{p}\frac{\sqrt{\tau\phi}}{\theta_i^2}e^{-\frac{\tau\phi}{\theta_i^2}}\right)N(\boldsymbol{\theta};\mathbf{m},\phi S^{-1}) \quad (8)$$

$$\pi(\phi|\boldsymbol{\theta},\mathbf{y}_n) = e^{-\tau\phi\sum_{i=1}^{p}\theta_i^{-2}}\,\mathrm{IG}\left(\phi;\frac{a_\phi+n-p}{2},\frac{b_\phi+s_R^2}{2}\right),$$

where $S = X'X$, $\mathbf{m} = S^{-1}X'\mathbf{y}_n$ and $s_R^2 = (\mathbf{y}_n - X\boldsymbol{\theta})'(\mathbf{y}_n - X\boldsymbol{\theta})$. Now, the piMOM prior is $\boldsymbol{\pi}_I(\boldsymbol{\theta}|\boldsymbol{\phi}) =$

$$N(\boldsymbol{\theta};\mathbf{0};\tau_N\phi\mathrm{I})\prod_{i=1}^{p}\frac{\frac{\sqrt{\tau\phi}}{\sqrt{\pi\theta_i^2}}e^{-\frac{\phi\tau}{\theta_i^2}}}{N(\theta_i;0,\tau_N\phi)} = N(\boldsymbol{\theta};\mathbf{0};\tau_N\phi\mathrm{I})\prod_{i=1}^{p}d_i(\theta_i,\phi). \quad (9)$$

In principle any τ_N may be used, but $\tau_N \geq 2\tau$ guarantees $d(\theta_i, \phi)$ to be monotone increasing in θ_i^2, so that its inverse exists (Supplementary Section 4). By default we set $\tau_N = 2\tau$. Corollary 7 gives

$$\pi(\boldsymbol{\theta}|\phi,\boldsymbol{\lambda}) = N(\boldsymbol{\theta};\mathbf{0},\tau_N\phi\mathrm{I})\prod_{i=1}^{p}\mathrm{I}(d(\theta_i,\phi)>\lambda_i)\frac{1}{h(\lambda_i)} \quad (10)$$

and $\pi(\boldsymbol{\lambda}) = \prod_{i=1}^{p}h(\lambda_i)$, where $h(\lambda_i) = P(d(\theta_i, \phi) > \lambda_i)$ which we need not evaluate. Algorithm 2 gives the following MH within Gibbs procedure.

1. MH step
   a. Propose $\phi^* \sim \mathrm{IG}\left(\phi;\frac{a_\phi+n-p}{2},\frac{b_\phi+s_R^2}{2}\right)$
   b. Set $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^*$ with probability $\min\left\{1, e^{\left(\phi^{(k-1)}-\phi^*\right)\tau\sum_{i=1}^{p}\theta_i^{-2}}\right\}$, else $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^{(k-1)}$.
2. $\boldsymbol{\lambda}^{(k)} \sim \prod_{i=1}^{p}\mathrm{Unif}\left(\lambda_i;0,d\left(\theta_i^{(k-1)},\phi^{(k)}\right)\right)$
3. $\boldsymbol{\theta}^{(k)} \sim N\left(\boldsymbol{\theta};\mathbf{m},\phi^{(k)}S^{-1}\right)\prod_{i=1}^{p}\mathrm{I}\left(d\left(\theta_i,\phi^{(k)}\right)>\lambda_i^{(k)}\right)$.

Step 3 requires the inverse $d^{-1}(\cdot)$, which can be evaluated efficiently combining an asymptotic approximation with a linear interpolation search (Supplementary Section 4). As a

token, 10,000 draws for $p = 2$ variables required 0.58 seconds on a 2.8 GHz processor running OS X 10.6.8.

4.2.3. peMOM prior—The full conditional posteriors are

$$\pi(\boldsymbol{\theta}|\phi,\mathbf{y}_n) \propto \left(\prod_{i=1}^{p} e^{-\frac{\tau\phi}{\theta_i^2}}\right) N\left(\boldsymbol{\theta};\mathbf{m},\phi S^{-1}\right); \pi(\phi|\boldsymbol{\theta},\mathbf{y}_n) \propto e^{-\sum_{i=1}^{p}\frac{\tau\phi}{\theta_i^2}} \mathrm{IG}\left(\phi;\frac{a^*}{2},\frac{b^*}{2}\right), \quad (11)$$

where $S = X'X + \boldsymbol{\tau}^{-1}I$, $\mathbf{m} = S^{-1}X'\mathbf{y}_n$, $a^* = a_\phi + n + p$, $b^* = b_\phi + s_R^2 + \boldsymbol{\theta}'\boldsymbol{\theta}/\tau$. Corollary 7 gives

$$\pi(\boldsymbol{\theta}|\phi,\boldsymbol{\lambda}) = N(\boldsymbol{\theta};\mathbf{0},\tau\phi I)\prod_{i=1}^{p}\mathrm{I}\left(e^{\sqrt{2}-\frac{\tau\phi}{\theta_i^2}} > \lambda_i\right)\frac{1}{h(\lambda_i)} \quad (12)$$

and $\pi(\lambda_i) = h(\lambda_i) = P\left(e^{\sqrt{2}-\frac{\tau\phi}{\theta_i^2}} > \lambda_i|\phi\right)$. Again $h(\lambda_i)$ has no simple form but is not required by Algorithm 2, which gives the k^{th} Gibbs iteration

1. $\phi^{(k)} \sim e^{-\sum_{i=1}^{p}\frac{\tau\phi}{\theta_i^2}}\mathrm{IG}\left(\phi;\frac{a^*}{2},\frac{b^*}{2}\right)$
   a. Propose $\phi^* \sim \mathrm{IG}\left(\phi;\frac{a^*}{2},\frac{b^*}{2}\right)$
   b. Set $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^*$ with probability $\min\left\{1, e^{(\phi^{(k-1)}-\phi^*)\tau\sum_{i=1}^{p}(\theta_i^{(k-1)})^{-2}}\right\}$, else $\boldsymbol{\phi}^{(k)} = \boldsymbol{\phi}^{(k-1)}$.
2. $\boldsymbol{\lambda}^{(k)} \prod_{i=1}^{p}\mathrm{Unif}\left(\lambda_i;0,e^{\sqrt{2}-\tau\phi/(\theta_i^{(k-1)})^2}\right)$
3. $\boldsymbol{\theta}^{(k)} \sim N(\boldsymbol{\theta};\mathbf{m},\phi^{(k)}S^{-1})\prod_{i=1}^{p}\mathrm{I}\left(\theta_i^2 > |\frac{\phi\tau}{\log(\lambda_i^{(k)})-\sqrt{2}}|\right)$

5. Examples

We assess our posterior sampling algorithms and the use of NLPs for high-dimensional estimation. Section 5.1 shows a simple yet illustrative multi-modal example. Section 5.2 studies $p \geq n$ cases and compares the BMA estimators induced by NLPs with benchmark

priors (BP, Fernández et al. (2001)), hyper-g priors (HG, Liang et al. (2008)), SCAD (Fan and Li, 2001), LASSO (Tibshirani, 1996) and Adaptive LASSO (ALASSO, Zhou (2006)). For NLPs and BP we used R package mombf 1.6.0 with default prior dispersions $\tau = 0.358$, 0.133, 0.119 for pMOM, piMOM and peMOM (respectively), which assign 0.01 prior probability to $|\theta_i/\sqrt{\phi}| < 0.2$ (Johnson and Rossell, 2010), and $\phi \sim \text{IG}(0.01/2, 0.01/2)$. The model search and posterior sampling algorithms are described in Supplementary Section 5. Briefly, we performed 5,000 Gibbs iterations to sample from $P(M_k|\mathbf{y}_n)$ and subsequently sampled $\boldsymbol{\theta}_k$ given M_k, $\mathbf{y}_n$ as outlined in Section 4.2. For HG we used R package BMS 0.3.3 with default alpha=3 and 10^5 MCMC iterations in Section 5.2, for the larger example in Section 5.3 we used package BAS with 3×10^6 iterations as it provided higher accuracy at lower running times. For LASSO, ALASSO and SCAD we set the penalization parameter with 10-fold cross-validation using functions mylars and ncvreg in R packages parcor 0.2.6 and ncvreg 3.2.0 (respectively) with default parameters. The R code is in the supplementary material. For all Bayesian methods we set a Beta-Binomial(1,1) prior on the model space. This is an interesting sparsity-inducing prior, *e.g.* for M_k with $p_k = p_t + 1$ it assigns $P(M_k)/P(M_t) = 1/(p - p_t)$. From Proposition 3 if $p > n$ this penalty more than doubles the shrinkage of $E(\theta_i|\mathbf{y}_n)$ under LPs, *i.e.* they should perform closer to NLPs. Also note that BP sets $\boldsymbol{\theta}_k|\phi_k, M_k \sim N(\mathbf{0}; g\phi X'_{k,n}X_{k,n})$ with $g = \max\{n, p^2\}$, which in our $p \geq n$ simulations induces extra sparsity and thus shrinkage. We assess the relative merits of each method without any covariate pre-screening procedures.

5.1. Posterior samples for a given model

We simulated $n = 1,000$ realizations from $y_i \sim N(\theta_1 x_{1i} + \theta_2 x_{2i}, 1)$, where (x_{1i}, x_{2i}) are drawn from a bivariate Normal with $E(x_{1i}) = E(x_{2i}) = 0$, $V(x_{1i}) = V(x_{2i}) = 2$, $\text{Cov}(x_{1i}, x_{2i}) = 1$. We first consider $\theta_1 = 0.5$, $\theta_2 = 1$, and compute posterior probabilities for the four possible models. We assign equal a priori probabilities and obtain exact $m_k(\mathbf{y}_n)$ using `pmomMarginalU`, `pimomMarginalU` and `pemomMarginalU` in `mombf` (the former has closed-form, for the latter two we used 10^6 importance samples). The posterior probability assigned to the full model under all three priors is 1 (up to rounding) (Supplementary Table 1). Figure 2 (left) shows 900 Gibbs draws (100 burn-in) obtained under the full model. The posterior mass is well-shifted away from 0 and resembles an elliptical shape for the three priors. Supplementary Table 2 gives the first-order auto-correlations, which are very small. This example reflects the advantages of the orthogonalization strategy, which is particularly efficient as the latent truncation becomes negligible.

We now set $\theta_1 = 0$, $\theta_2 = 1$ and keep $n = 1000$ and (x_{1i}, x_{2i}) as before. We simulated several data sets and in most cases did not observe a noticeable posterior multi-modality. We portray a specific simulation that did exhibit multi-modality, as this poses a greater challenge from a sampling perspective. Table 1 shows that the data-generating model has highest posterior probability. Although the full model was clearly dismissed in light of the data, as an exercise we drew from its posterior. Figure 2 (right) shows 900 Gibbs draws after a 100 burn-in, and Supplementary Table 2 shows a low auto-correlation. The samples adequately captured the multiple modes.

5.2. High-dimensional estimation

5.2.1. Growing p, fixed n and θ—We perform a simulation study with $n = 100$ and growing $p = 100, 500, 1000$. We set $\theta_i = 0$ for $i = 1, \ldots, p - 5$, the remaining 5 coefficients to (0.6, 1.2, 1.8, 2.4, 3) and residual variances $\phi = 1, 4, 8$. Covariates were sampled from $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{ii} = 1$ and all correlations set to $\rho = 0$ or $\rho = 0.25$. We remark that ρ are population correlations, the maximum sample correlations when $\rho = 0$ were 0.37, 0.44, 0.47 for $p = 100, 500, 1000$ (respectively), and 0.54, 0.60, 0.62 when $\rho = 0.25$. We simulated 1,000 data sets under each setup.

Figure 3 shows sum of squared errors (SSE) averaged across simulations for $\phi = 1, 4, 8$, $\rho = 0, 0.25$. pMOM and piMOM perform similarly and present a lower SSE as p grows than other methods in all scenarios. To obtain more insight on how the lower SSE is achieved, Supplementary Figures 2–3 show SSE separately for $\theta_i = 0$ (left) and $\theta_i \neq 0$ (right). The largest differences between methods were observed for $\theta_i = 0$, the performance of pMOM and piMOM coming closer for smaller signal-to-noise ratios $|\theta_i|/\sqrt{\phi_i}$. For $\theta_i \neq 0$ differences in SSE are smaller, iMOM slightly outperforming MOM. For all methods as $|\theta_i|/\sqrt{\phi_i}$ decrease the SSE worsens relative to the oracle least squares (Supplementary Figures 2–3, right panels, black horizontal segments).

5.2.2. Growing p, θ = O(n$^{-1/4}$)—We extend the simulations by considering $p = 100, 500, 1000$ and $\rho = 0, 0.25$ as before in a setting with vanishing $\boldsymbol{\theta} = O(n^{-1/4})$. Specifically, we set $n = 100, 250, 500$ for $p = 100, 500, 1000$ (respectively), $\theta_i = 0$ for $i = 1, \ldots, p - 5$ as before and the remaining 5 coefficients to $n^{-1/4}(0.6, 1.2, 1.8, 2.4, 3)$ and $\phi = 1$. The goal is to investigate if NLP shrinkage rate comes at a cost of reduced precision when the coefficients are truly small. Note that $n^{-1/4}$ is only slightly larger than the $n^{-1/2}$ error of the MLE, and hence represents fairly small coefficients.

Figure 4 shows the total SSE and Supplementary Figure 4 that for zero (left) and non-zero (right) coefficients. MOM and iMOM present the lowest overall SSE in most situations but HG and ALASSO achieve similar performance, certainly closer than the earlier sparser scenario with fixed $\boldsymbol{\theta}$, $n = 100$ and growing p.

Because NLPs assign high prior density to a certain range of $|\theta_i|/\sqrt{\phi}$ values, we conducted a further study when $\boldsymbol{\theta}$ contains an ample range of non-zero coefficients (*i.e.* both large and small). To this end, we set $n = 100, 250, 500$ for $p = 100, 500, 1000$ with $\phi = 1$ as before, $\theta_i = 0$ for $i = 1, \ldots, p - 11$, vanishing $(\theta_{p-10}, \ldots, \theta_{p-6}) = n^{-1/4}$ (0.6, 1.2, 1.8, 2.4, 3) and fixed $(\theta_{p-5}, \ldots, \theta_p) = (0.6, 1.2, 1.8, 2.4, 3)$. Figure 5 shows the overall MSE and Supplementary Figure 5 that for $\theta_i = 0$ and $\theta_i \neq 0$ separately. The lowest overall MSE is achieved by iMOM and MOM, followed by HG and BP, whereas ALASSO is less competitive than in the earlier simulations where all $\theta_i = O(n^{-1/4})$. Overall, these results support that NLPs remain competitive even with small signals and that their performance relative to competing methods is best in sparse situations, agreeing with our theoretical findings.

5.3. Gene expression data

We assess predictive performance in high-dimensional gene expression data. (Calon et al., 2012) used mice experiments to identify 172 genes potentially related to the gene TGFB, and showed that these were related to colon cancer progression in an independent data set with $n = 262$ human patients. TGFB plays a crucial role in colon cancer and it is important to understand its relation to other genes. Our goal is to predict TGFB in the human data, first using only the $p = 172$ genes and then adding 10,000 extra genes that we selected randomly from the 18,178 genes with distinct Entrez identifier contained in the experiment. Their absolute Pearson correlations with the 172 genes ranged from 0 to 0.892 with 95% of them being in (0.003,0.309). Both response and predictors were standardized to zero mean and unit variance (data and R code in Supplementary Material). We assessed predictive performance via the leave-one-out cross-validated R^2 coefficient between predictions and observations. For Bayesian methods we report the posterior expected number of variables in the model (*i.e.* the mean number of predictors used by BMA), and for SCAD and LASSO the number of selected variables.

Table 1 shows the results. For $p = 172$ all methods achieve similar R^2, that for LASSO being slightly higher, although pMOM, piMOM and BP used substantially less predictors. These results appear reasonable in a moderately dimensional setting where genes are expected to be related to TGFB. However, when using $p = 10,172$ predictors important differences between methods are observed. The BMA estimates based on pMOM and piMOM remain parsimonious (6.5 and 10.3 predictors, respectively) and the cross-validated R^2 increases roughly to 0.62. The BP prior dispersion parameter $g = 172^2$ induces strong parsimony, though relative to NLPs the non-selectiveness of this penalty causes some loss of prediction power ($R^2 = 0.586$). For the remaining methods the number of predictors increased sharply and R^2 did not improve relative to the $p = 172$ case. Predictors with large marginal inclusion probabilities in pMOM/piMOM included genes related to various cancer types (ESM1, GAS1, HIC1, CILP, ARL4C, PCGF2), TGFB regulators (FAM89B) or AOC3 which is used to alleviate certain cancer symptoms. These findings suggest that NLPs effectively detected a parsimonious subset of predictors in this high-dimensional example. We also note that computation times were highly competitive. BP and NLPs are programmed in mombf in an identical manner (piMOM has no closed-form expressions, hence the higher time) whereas HG is implemented in BAS with a slightly more advanced MCMC model search algorithm (*e.g.* pre-ranking variables and considering swaps). NLPs focus $P(M_k \mid \mathbf{y}_n)$ on smaller models, which alleviates the cost required by matrix inversions (non-linear in the model size). NLPs also concentrate $P(M_k \mid \mathbf{y}_n)$ on a smaller subset of models, which tend to be revisited and hence the marginal likelihood need not be recomputed. Regarding the efficiency of our posterior sampler for $(\boldsymbol{\theta}, \phi)$, we ran 10 independent chains with 1,000 iterations each and obtained mean serial correlations of 0.32 (pMOM) and 0.26 (piMOM) across all non-zero coefficients. The mean correlation between $\hat{E}(\boldsymbol{\theta} \mid \mathbf{y}_n)$ across all chain pairs was > 0.99 (pMOM and piMOM). Supplementary Section 5 contains further convergence assessments.

6. Discussion

We showed how combining BMA with NLPs gives a coherent joint framework encouraging model selection parsimony and selective shrinkage for spurious coefficients. Beyond theory, the latent truncation construction motivates NLPs from first principles, adds flexibility in prior choice and enables effective posterior sampling even under strong multi-modalities. We obtained strong results when $p \gg n$ in simulations and gene expression data, with parsimonious models achieving accurate cross-validated predictions and good computation times. Note that these did not require procedures to pre-screen covariates, which can cause a loss of detection power. Interestingly, NLPs achieved low estimation error even in settings with vanishing coefficients: their slightly higher SSE for active coefficients was compensated by a lower SSE for inactive coefficients. That is, NLPs can be advantageous even with sparse vanishing $\boldsymbol{\theta}$, although of course they may be less competitive in non-sparse situations. An important point is that inducing sparsity via $P(M_k)$ (*e.g.* Beta-Binomial) or vague $\pi(\boldsymbol{\theta}_k \mid M_k)$ (*e.g.* the BP) also performed reasonably well, although relative to the NLP data-adaptive sparsity there can be a loss of detection power.

Our results show that it is not only possible to use the same prior for estimation and selection, but may indeed be desirable. We remark that we used default informative priors, which are relatively popular for testing, but perhaps less readily adopted for estimation. Developing objective Bayes strategies to set the prior parameters is an interesting venue for future research, as well as determining shrinkage rates in more general $p \gg n$ cases, and adapting the latent truncation construction beyond linear regression, *e.g.* generalized linear, graphical or mixture models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Both authors were partially funded by the NIH grant R01 CA158113-01. We thank Merlise Clyde for providing the BAS package.

References

Bhattacharya A, Pati D, Pillai NS, Dunson DB. Bayesian shrinkage Technical report. arXiv preprint arXiv:1212.6088. 2012

Calon A, Espinet E, Palomo-Ponce S, Tauriello DVF, Iglesias M, Céspedes MV, Sevillano M, Nadal C, Jung P, Zhang XHF, Byrom D, Riera A, Rossell D, Mangues R, Massague J, Sancho E, Batlle E. Dependency of colorectal cancer on a tgf-beta-driven programme in stromal cells for metastasis initiation. Cancer Cell. 2012; 22(5):571–584. [PubMed: 23153532]

Castillo I, Van der Vaart AW. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. The Annals of Statistics. 2012; 40(4):2069–2101.

Castillo I, Schmidt-Hieber J, van der Vaart AW. Bayesian linear regression with sparse priors. Technical report, arXiv preprint arXiv:1403.0735. 2014

Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95(3):759–771.

Consonni, G., La Rocca, L. On moment priors for Bayesian model choice with applications to directed acyclic graphs. In: Bernardo, JM.Bayarri, MJ.Berger, JO.Dawid, AP.Heckerman, D.Smith, AFM.,

West, M., editors. Bayesian Statistics 9 - Proceedings of the ninth Valencia international meeting. Oxford University Press; 2010. p. 119-144.

Dawid, AP. The trouble with Bayes factors Technical report. University College London; 1999.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statistica Sinica. 2010; 20:101–140. [PubMed: 21572976]

Fernández C, Ley E, Steel MFJ. Benchmark priors for Bayesian model averaging. Journal of Econometrics. 2001; 100:381–427.

Gelfand AE, Ghosh SK. Model choice: A minimum posterior predictive loss approach. Biometrika. 1998; 85:1–11.

Johnson VE, Rossell D. Prior densities for default Bayesian hypothesis tests. Journal of the Royal Statistical Society B. 2010; 72:143–170.

Johnson VE, Rossell D. Bayesian model selection in high-dimensional settings. Journal of the American Statistical Association. 2012; 24(498):649–660.

Kotecha, JH., Djuric, PM. Proceedings, 1999 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE Computer Society; 1999. Gibbs sampling approach for generation of truncated multivariate gaussian random variables; p. 1757-1760.

Lai TL, Robbins H, Wei CZ. Strong consistency of least squares in multiple regression. Journal of multivariate analysis. 1979; 9:343–361.

Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g-priors for Bayesian variable selection. Journal of the American Statistical Association. 2008; 103:410–423.

Liang F, Song Q, Yu K. Bayesian modeling for high-dimensional generalized linear models. Journal of the American Statistical Association. 2013; 108(502):589–606.

Martin R, Walker SG. Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. Technical report. 2013 arXiv preprint arXiv:1304.7366.

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

Redner R. Note on the consistency of the maximum likelihood estimator for nonidentifiable distributions. Annals of Statistics. 1981; 9(1):225–228.

Rodriguez-Yam, G., Davis, RA., Scharf, LL. PhD thesis. Department of Statistics, Colorado State University; 2004. Efficient Gibbs sampling of truncated multivariate normal with application to constrained linear regression.

Rossell, D., Telesca, D., Johnson, VE. Statistical Models for Data Analysis XV. Springer; 2013. High-dimensional Bayesian classifiers using non-local priors; p. 305-314.

Rousseau, J. Approximating interval hypothesis: p-values and Bayes factors. In: Bernardo, JM.Bayarri, MJ.Berger, JO., Dawid, AP., editors. Bayesian Statistics. Vol. 8. Oxford University Press; 2007. p. 417-452.

Shin, M., Bhattacharya, A., Johnson, VE. Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings; arXiv. 2015. p. 1-33.http://arxiv.org/abs/1507.07106

Springer MD, Thompson WE. The distribution of products of beta, gamma and gaussian random variables. SIAM Journal of Applied Mathematics. 1970; 18(4):721–737.

Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, B. 1996; 58:267–288.

Verdinelli, I., Wasserman, L. Bayes factors, nuisance parameters and imprecise tests. In: Bernardo, JM.Berger, JO.Dawid, AP., Smith, AFM., editors. Bayesian Statistics. Vol. 5. Oxford University Press; 1996. p. 765-771.

Walker AM. On the asymptotic behaviour of posterior distributions. Jornal of the Royal Statistical Society B. 1969; 31(1):80–88.

Wilhelm S, Manjunath BG. tmvtnorm: a package for the truncated multivariate normal distribution. The R Journal. 2010; 2:25–29.

Zhou H. The adaptive LASSO and its oracle properties. Journal of the American Statistical Association. 2006; 101(476):1418–1429.

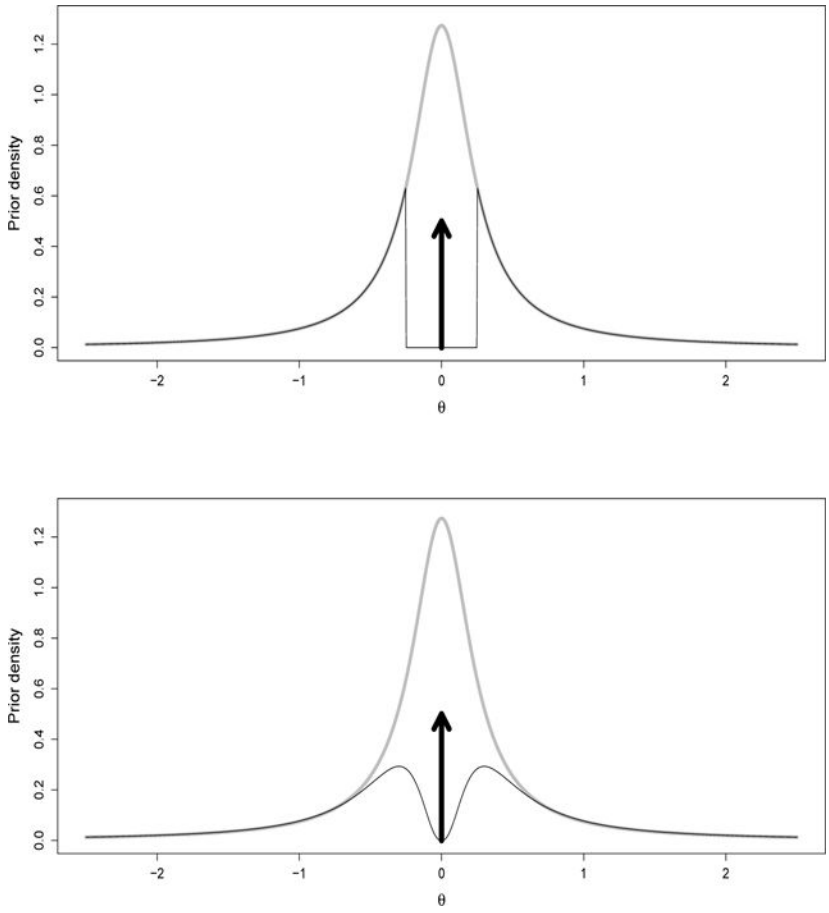

Figure 1.

Marginal priors for $\theta \in \mathbb{R}$ (estimation prior Cauchy(0, 0.0625) shown in grey). Top: mixture of point mass at 0 and Cauchy(0, 0.0625) truncated at $\lambda = 0.25$; Bottom: same as top with $\lambda \sim IG(3, 10)$

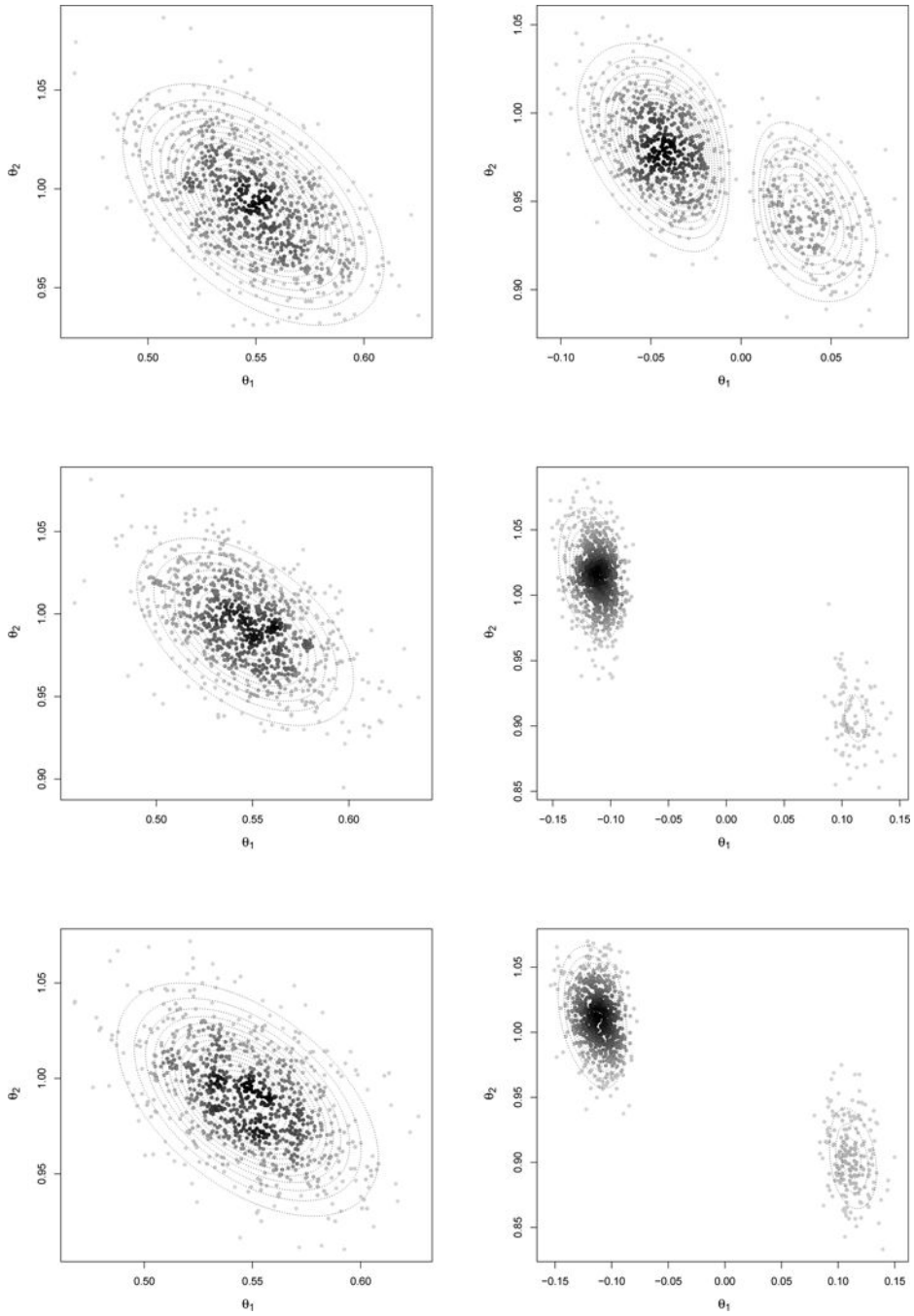

Figure 2.

900 Gibbs draws when $\boldsymbol{\theta} = (0.5, 1)'$ (left) and $\boldsymbol{\theta} = (0, 1)'$ (right) and posterior density contours. Top: MOM ($\tau = 0.358$); Middle: iMOM ($\tau = 0.133$); Bottom: eMOM ($\tau = 0.119$)

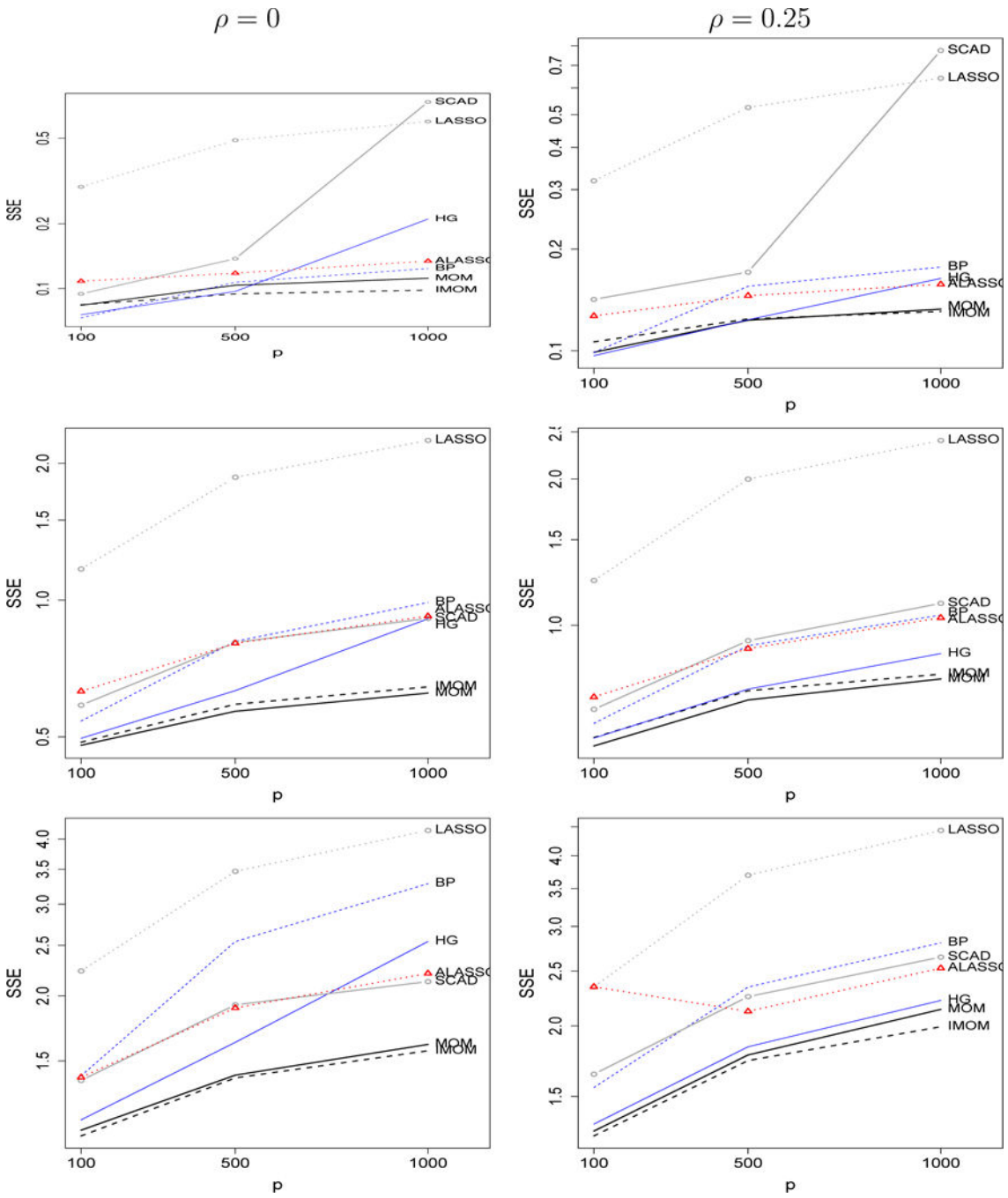

Figure 3.

Mean SSE when $\phi = 1, 4, 8$ (top, middle, bottom), $\rho = 0, 0.25$ (left, right). Simulation settings: $n = 100$, $p = 100, 500, 1000$ and 5 non-zero coefficients 0.6, 1.2, 1.8, 2.4, 3.0.

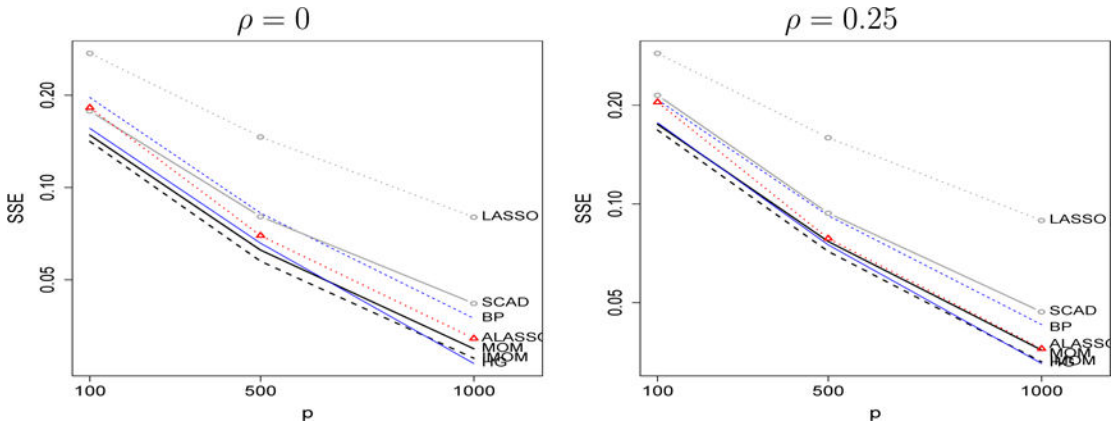

Figure 4.

Mean SSE when non-zero $\theta = n^{-1/4}$ (0.6, 1.2, 1.8, 2.4, 3.0), $\rho = 0$, 0.25 (left, right), $\phi = 1$. Simulation settings: ($n = 100$, $p = 100$), ($n = 250$, $p = 500$), ($n = 500$, $p = 1000$)

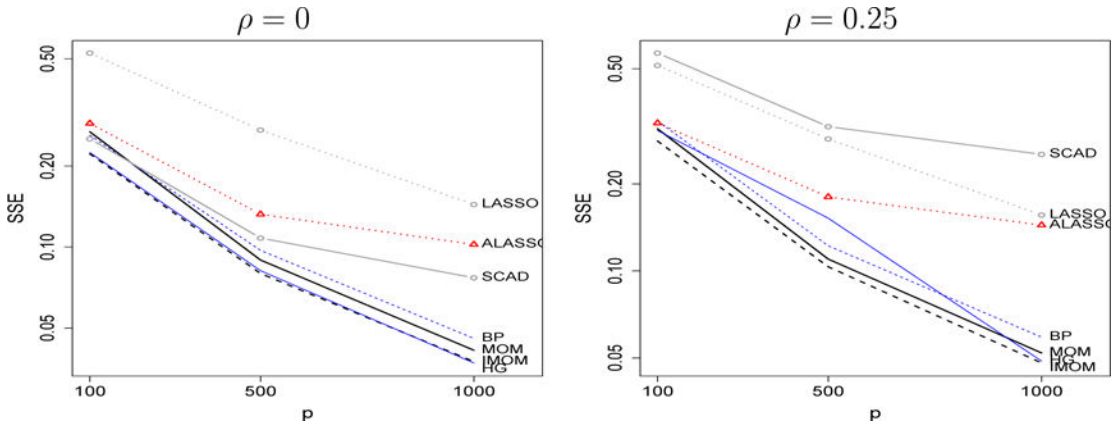

Figure 5.

Mean SSE when non-zero $(\theta_{p-10}, \ldots, \theta_{p-6}) = n^{-1/4}(0.6, 1.2, 1.8, 2.4, 3)$, $(\theta_{p-5}, \ldots, \theta_p) = (0.6, 1.2, 1.8.2.4, 3)$ and $\rho = 0, 0.25$ (left, right), $\phi = 1$. Simulation settings: $(n = 100, p = 100)$, $(n = 250, p = 500)$, $(n = 500, p = 1000)$

Table 1

Expression data with $p = 172$ or 10, 172 genes. $\bar{p}$: mean (MOM, iMOM, BP, HG) or selected number of predictors (SCAD, LASSO, ALASSO). R^2 coefficient is between $(y_i, \hat{y}_i)$ (leave-one-out cross-validation). CPU time on Linux OpenSUSE 13.1, 64 bits, 2.6GHz processor, 31.4Gb RAM for 1,000 Gibbs iterations (MOM,iMOM,BP) or 3×10^6 model updates (HG)

	$p = 172$		$p = 10, 172$		
	$\bar{p}$	R^2	$\bar{p}$	R^2	CPU time
MOM	4.3	0.566	6.5	0.617	1m 52s
iMOM	5.3	0.560	10.3	0.620	59m
BP	4.2	0.562	3.0	0.586	1m 23s
HG	11.3	0.562	26.4	0.522	11m 49s
SCAD	29	0.565	81	0.535	16.7s
LASSO	42	0.586	159	0.570	23.7s
ALASSO	24	0.569	10	0.536	2m 49s

J Am Stat Assoc. Author manuscript; available in PMC 2018 June 05.