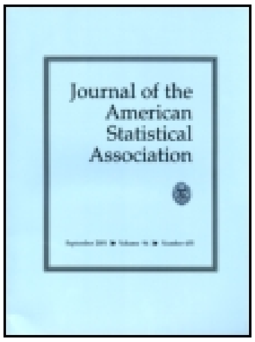

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Tractable Bayesian variable selection: beyond normality

David Rossell & Francisco J. Rubio

To cite this article: David Rossell & Francisco J. Rubio (2017): Tractable Bayesian variable selection: beyond normality, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1371025

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1371025

Tractable Bayesian variable selection: beyond normality

David Rossell[1], Francisco J. Rubio[2]

Author's Footnote

[1] Universitat Pompeu Fabra, Department of Business and Economics, Barcelona (Spain)

[2] London School of Hygiene & Tropical Medicine, London (United Kingdom)

Abstract

Bayesian variable selection often assumes normality, but the effects of model misspecification are not sufficiently understood. There are sound reasons behind this assumption, particularly for large p: ease of interpretation, analytical and computational convenience. More flexible frameworks exist, including semi- or non-parametric models, often at the cost of some tractability. We propose a simple extension that allows for skewness and thicker-than-normal tails but preserves tractability. It leads to easy interpretation and a log-concave likelihood that facilitates optimization and integration. We characterize asymptotically parameter estimation and Bayes factor rates, under certain model misspecification. Under suitable conditions misspecified Bayes factors induce sparsity at the same rates than under the correct model. However, the rates to detect signal change by an exponential factor, often reducing sensitivity. These deficiencies can be ameliorated by inferring the error distribution, a simple strategy that can improve inference substantially. Our work focuses on the likelihood and can be combined with any likelihood penalty or prior, but here we focus on non-local priors to induce extra sparsity and ameliorate finite-sample effects caused by misspecification. We show the importance of considering the likelihood rather than solely the prior, for Bayesian variable selection. The methodology is in R package 'mombf'.

Keywords: Variable selection, two-piece errors, Bayes factors, model misspecification, robust regression.

1. INTRODUCTION

The rise of high-dimensional problems has generated a renewed interest in simple models. Beyond the obvious issue that modest sample sizes limit the number of parameters that can be learned accurately, simple models remain a central choice due to their analytical and computational tractability, ease of interpretation, and the fact that they often work well in practice. There is, however, a pressing need to seek extensions which, while retaining the aforementioned advantages, incorporate additional flexibility and can be studied without unrealistically assuming that the posed model is correct. Ideally such extensions should detect when the added flexibility is not needed so that one can fall back onto simpler models. We focus on canonical variable selection in linear regression from a Bayesian standpoint, although some results may also be useful for penalized likelihood methods. Given that the number of models to consider is exponential in the number of variables, it is highly convenient to adopt error models that lead to fast within-model calculations, *e.g.* closed forms or fast approximations for the integrated likelihood. Our work is based on two-piece distributions, an easily interpretable family that has a long history and which we fully characterize in the linear model case (synthesizing and extending current results) under model misspecification. Our main contributions are showing that two-piece errors (specifically when applied to the Normal and Laplace families) lead to tractable inference, proposing simple computational algorithms, and characterizing variable selection under model misspecification, including when this likelihood is combined with non-local priors (NLPs, Johnson and Rossell (2010)). We show that in the presence of asymmetries or heavy tails the Normal model incurs a significant loss of power, and propose a formal strategy to detect such departures from normality. When these departures are negligible our model collapses onto Normal errors, for which closed-form expressions are often available.

To fix ideas, we consider the linear regression model

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

where $y = (y_1, \ldots, y_n)^T$ is the observed outcome for n individuals, X is an $n \times p$ matrix with potential predictors, $\theta = (\theta_1, \ldots, \theta_p)^T \in \mathbb{R}^p$ are regression coefficients and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are independent and identically distributed (id) errors (see Section 5.2 for a discussion on non-id errors).

The goal is to determine the non-zero coefficients in θ under an arbitrary data-generating distribution for the ϵ_i's, building a framework that remains convenient for large p. Let $\gamma_j = \mathrm{I}(\theta_j \neq 0)$ for $j = 1, \ldots, p$ be variable inclusion indicators and $p_\gamma = \sum_{j=1}^{p} \gamma_j$ the number of active variables. To consider that residuals may be asymmetric and/or have thicker-than-normal tails $\gamma_{p+1} = 1$ denotes the presence of asymmetry ($\gamma_{p+1} = 0$ otherwise) and $\gamma_{p+2} = 1$ that of thick tails ($\gamma_{p+2} = 0$ for Normal tails). Thus $\gamma = (\gamma_1, \ldots, \gamma_{p+2})$ denotes the assumed model. X_γ and θ_γ are the corresponding submatrix of X and subvector of θ, respectively. We denote the i^{th} row in X and X_γ by $x_i^T \in \mathbb{R}^p$ and $x_{\gamma i}^T \in \mathbb{R}^{p_\gamma}$.

There are a number of proposals to relax the normality assumption. Within the frequentist literature Wang et al. (2007) proposed median regression with LASSO penalties (LASSO-LAD) and Wang and Li (2009) with rank-based SCAD penalties. Arslan (2012) extended LASSO median regression by weighting observations and Fan et al. (2014) considered adaptive LASSO quantile regression. These approaches are formally connected to assuming either Laplace or asymmetric Laplace errors. There are also model-free M-estimation methods, *e.g.* combining Huber's loss with an adaptive LASSO penalty (Lambert-Lacroix 2011), sparse trimmed-means LASSO (Alfons et al. 2013), and non-negative garrote extensions to induce robustness to outliers (Gijbels and Vrinssen 2015). Theoretical characterizations also exist, *e.g.* Mendelson (2014) proved the consistency and asymptotic normality of high-dimensional M-estimators and Loh (2017) extended the results to generalized M-estimators with non-convex loss functions. Within the Bayesian framework, Gottardo and Raftery (2007) and Wang et al. (2016) consider variable selection after transforming y_i and/or x_i, the former allowing for t errors and the latter inducing NLPs on θ via the transformation's Jacobian. While certainly interesting, the transformed conditional mean $E(y_i \mid x_i)$ is no longer linear in x_i and parameter interpretation and prior elicitation is less straightforward. Our main interest is in linear predictors with simple error distributions. Along these lines, Yu et al. (2013) proposed Gibbs sampling for model choice in Bayesian quantile regression using a latent scale augmentation, and Yan and Kottas (2015) extended Azzalini's skew Normal to Laplace errors within Bayesian quantile regression, which leads to easily-implementable MCMC, and induced sparsity via LASSO penalties. Related to our work Rubio and Genton (2016) and Rubio and Yu (2017) employ skew-symmetric and two-piece errors in linear regression, respectively, albeit the set of covariates is fixed and they focus on prediction and censored responses. Yet another possible avenue is to pose highly flexible errors, *e.g.* Chung and Dunson (2009) set a non-parametric model to simultaneously learn the effect of x_i on the mean and on the shape of the residual distribution. Kundu and Dunson (2014) proposed variable selection with non-parametric symmetric residuals, for which notably Chae et al. (2016) proved high-dimensional model selection consistency and concentration rates under model misspecification. Most Bayesian work uses Markov Chain Monte Carlo (MCMC) for parameter estimation and computation of marginal likelihoods and does not collapse onto the Normal model when warranted by the data, hampering its computational scalability as p or n grow, further the theoretical study is typically M-closed.

In contrast, we show that simpler parametric error models equipped with efficient analytical approximations to the integrated likelihood achieve selection consistency under model misspecification, and embed these models within a framework that when appropriate collapses onto normality. We also show that model misspecification can markedly decrease the sensitivity to detect truly active variables, *e.g.* under asymmetry or heavy tails. Our results complement the examples in Grünwald and van Ommen (2014), where the presence of inliers favoured the addition of spurious variables (see also Figure 1 in Kundu and Dunson (2014)). We show that asymptotically misspecified Bayes factors to discard spurious models essentially multiply the correct Bayes factor by a constant term, but when detecting true signals this term is exponential in n. That is, asymptotically model misspecification has more serious effects on sensitivity than on false positives. For finite n, false positives can be an important issue. We use the example in Grünwald and van Ommen (2014) to illustrate how such finite n effects can be reduced by penalizing small coefficients via NLPs (Section 6.2).

Before presenting our approach we clarify our main contributions relative to earlier work in two-piece distributions. Rubio and Steel (2014) showed that Jeffreys priors and their associated posteriors for location-scale two-piece models are improper, and that the (improper) independence Jeffreys prior leads to a proper posterior. Rubio and Yu (2017) extended the study to linear regression, again under improper priors. Unfortunately, improper priors cannot be used for Bayesian model selection as they lead to the well-known Jeffreys-Lindley-Bartlett paradox. There is also literature (e.g. Arellano-Valle et al. (2005)) on MLE consistency and asymptotic normality in the case with no covariates. Checks of the large sample theory technical conditions are however hard to come by, which are non-standard due to the non-existence of certain derivatives. Our two-piece likelihood properties, specifically log-concavity and asymptotic analysis under model misspecification are, to our knowledge, new. As well as our results on Bayes factors, indeed the main theme of our paper: model selection. The M-estimation technical machinery for the theorems is also of interest as an avenue for asymptotic analysis of Bayesian model selection under misspecification. Finally optimization and integration algorithms built on interior-point methods are newly developed here to scale with n and p. A particular case of our framework provides a new approach to Bayesian quantile regression. We also propose a novel strategy to infer the error model from the data.

The manuscript is structured as follows. Section 2 reviews two-piece distributions and establishes the concavity of the log-likelihood in the asymmetric Normal and Laplace cases. Section 3 proposes a prior formulation based on NLPs that enforces sparsity and discards degrees of asymmetry that are irrelevant in practice. Section 4 tackles maximum likelihood and posterior mode estimation, specifically giving asymptotic distributions and optimization algorithms that capitalize on likelihood tractability. Section 5 outlines a framework to select both variables and the residual distribution, proposes fast approximations to the integrated likelihood and characterizes asymptotically the associated Bayes factors. Section 6 shows results on simulated and experimental data, and Section 7 offers concluding remarks. The supplementary material contains all proofs and further results. R code to reproduce our results is also provided as a supplement to this article.

2. LOG-LIKELIHOOD

We recall the definition of a two-piece distribution for model (1) and predictors X_γ.

Definition 1. *A random variable $y_i \in \mathbb{R}$ following a two-piece distribution with location $x_{\gamma i}^T\theta_\gamma$, scale $\sqrt{\vartheta} \in \mathbb{R}^+$ and asymmetry α has density function $s(y_i; x_{\gamma i}^T\theta_\gamma, \vartheta, \alpha) =$*

$$\frac{2}{\sqrt{\vartheta}[a(\alpha)+b(\alpha)]}\left[f\left(\frac{y_i - x_{\gamma i}^T\theta_\gamma}{\sqrt{\vartheta}a(\alpha)}\right) I(y_i < x_{\gamma i}^T\theta_\gamma) + f\left(\frac{y_i - x_{\gamma i}^T\theta_\gamma}{\sqrt{\vartheta}b(\alpha)}\right) I(y_i \geq x_{\gamma i}^T\theta_\gamma)\right], \tag{2}$$

where $f(\cdot)$ is a symmetric unimodal density with mode at 0 *and support on* $\mathbb{R}$*, and* $a(\alpha), b(\alpha) \in \mathbb{R}^+$.

Two-piece distributions induce asymmetry by (continuously) merging two symmetric densities that have the same mode $x_{\gamma i}^T\theta_\gamma$ but different scale parameters $\sqrt{\vartheta}a(\alpha)$, $\sqrt{\vartheta}b(\alpha)$ on each side of the mode. Some popular parameterizations are the inverse scale factors $\{a(\alpha), b(\alpha)\} = \{\alpha, 1/\alpha\}$ for $\alpha \in \mathbb{R}^+$ (Fernández and Steel 1998) or the epsilon-skew parameterization $\{a(\alpha), b(\alpha)\} = \{1-\alpha, 1+\alpha\}$ for $\alpha \in [-1, 1]$ (Mudholkar and Hutson 2000). We adopt the latter as it leads to orthogonality in the expected log-likelihood hessian between α and ϑ, also it allows easy interpretation as the total variation distance between $s(y_i; x_{\gamma i}^T\theta_\gamma, \vartheta, \alpha)$ and its symmetric counterpart $s(y_i; x_{\gamma i}^T\theta_\gamma, \vartheta, 0)$ is $|\alpha|/2$ (Dette et al. 2016). Further, a classical skewness coefficient proposed by Arnold-Groeneveld defined as $\mathrm{AG} = 1 - 2F(x_{\gamma i}^T\theta_\gamma) \in [-1, 1]$ for a univariate random variable with mode at $x_{\gamma i}^T\theta_\gamma$ and cumulative distribution function $F()$, is equal to $\mathrm{AG} = -\alpha$ (Rubio and Steel 2014).

Two-piece distributions are appealing for regression given that the mode of $s()$ is $x_{\gamma i}^T\theta_\gamma$, its mean (when defined) depends on $x_{\gamma i}$ only through $x_{\gamma i}\theta_\gamma$ and its variance is proportional to ϑ (see below for specific expressions), facilitating interpretation and prior elicitation. Despite these properties and them being a classical strategy with a fascinating history, proposed at least as early as 1897 and rediscovered multiple times (Wallis 2014), their popularity has been limited due to practical concerns, *e.g.* log-likelihood maximization may be hampered by discontinuous gradients or hessians. For this reason we focus on two-piece Normal and Laplace errors, for which we prove log-concavity and thus analytical and computational tractability, giving a practical mechanism to capture asymmetry and heavier-than-normal tails. Specifically, the two-piece Normal is obtained by letting $f(z) = N(z; 0, 1)$ in (2) be the standard Normal density, and gives $E(y_i \mid x_{\gamma i}) = x_{\gamma i}^T\theta_\gamma - \alpha\sqrt{8\vartheta/\pi}$, $\mathrm{Var}(y_i \mid x_{\gamma i}) = \vartheta[(3 - 8/\pi)\alpha^2 + 1]$ and a median that is also linear in $x_{\gamma i}$ (Mudholkar and Hutson 2000). The corresponding likelihood has the simple expression $\log L_1(\theta_\gamma, \vartheta, \alpha) =$

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\vartheta) - \frac{1}{2\vartheta}\left(\sum_{i \in A(\theta_\gamma)} \frac{(y_i - x_{\gamma i}^T\theta_\gamma)^2}{(1+\alpha)^2} + \sum_{i \notin A(\theta_\gamma)} \frac{(y_i - x_{\gamma i}^T\theta_\gamma)^2}{(1-\alpha)^2}\right) =$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\vartheta) - \frac{1}{2\vartheta}(y - X_\gamma\theta_\gamma)^T W^2 (y - X_\gamma\theta_\gamma). \qquad (3)$$

where $A(\theta_\gamma) = \left\{i : y_i < x_{\gamma i}^T\theta_\gamma\right\}$ are the observations with negative residuals, $W = \mathrm{diag}(w)$, $w_i =$

$|1+\alpha|^{-1}$ if $i \in A(\theta_\gamma)$ and $w_i = |1-\alpha|^{-1}$ if $i \notin A(\theta_\gamma)$. For later convenience we denote by $\overline{w}$ the signed weight vector with $\overline{w}_i = w_i$ if $i \in A(\theta_\gamma)$ and $\overline{w}_i = -w_i$ if $i \notin A(\theta_\gamma)$, by $w^k = (w_1^k, \ldots, w_n^k)$ the element-wise k^{th} power of a vector, $\overline{w}^k = (\text{sign}(\overline{w}_1)|\overline{w}_1|^k, \ldots, \text{sign}(\overline{w}_n)|\overline{w}_n|^k)^T$ and $\overline{W}^k = \text{diag}(\overline{w}^k)$. Note that (3) is linked to asymmetric least square regression and is the Normal likelihood for $\alpha = 0$.

The two-piece Laplace is obtained by setting $f(z) = 0.5\exp(-|z|)$ in (2). This distribution is more commonly referred to as asymmetric Laplace, we denote it $y_i \sim \text{AL}(x_{\gamma i}^T\theta_\gamma, \vartheta, \alpha)$ and note that $E(y_i \mid x_{\gamma i}, \theta_\gamma, \vartheta, \alpha) = x_{\gamma i}^T\theta_\gamma - 2\alpha\sqrt{\vartheta}$ and $\text{Var}(y_i \mid x_{\gamma i}) = 2\vartheta(1+\alpha^2)$ (Arellano-Valle et al. 2005). For coherency from here onwards, we also refer to the two-piece Normal as asymmetric Normal and denote $y_i \sim \text{AN}(x_{\gamma i}^T\theta_\gamma, \vartheta, \alpha)$. The asymmetric Laplace log-likelihood is $\log L_2(\theta_\gamma, \vartheta, \alpha) =$

$$-n\log(2) - \frac{n}{2}\log(\vartheta) - \frac{1}{\sqrt{\vartheta}}\left(\sum_{i \in A(\theta_\gamma)} \frac{|y_i - x_{\gamma i}^T\theta_\gamma|}{1+\alpha} + \sum_{i \notin A(\theta_\gamma)} \frac{|y_i - x_{\gamma i}^T\theta_\gamma|}{1-\alpha}\right). \tag{4}$$

The symmetric Laplace case is obtained for $\alpha = 0$, in which case optimization of (4) with respect to θ_γ is equivalent to median regression, whereas for fixed $\alpha \neq 0$ it leads to quantile regression. Hence a particular case of our framework is obtained when conditioning upon asymmetric Laplace errors with a fixed α, this leads to Bayesian quantile regression for the quantile $\tau = (1+\alpha)/2$. Fixing α can be interesting in certain applications, is implemented in our software and illustrated in the DLD data (Section 6.5). However by default we recommend treating α as a parameter to be learnt from the data. This reduces sensitivity to model misspecification: conditioning upon non-optimal α increases the KL-divergence between the assumed model class and the data-generating truth, which may decrease power to detect truly active variables (Proposition 5 and follow-up discussion). Further, we propose a framework to infer the error distribution, clearly there one wishes to use the best-fitting α. Finally, each α conditioned upon may lead to different selected variables, this can be interesting but in applications one often is more interested in global variable selection.

Our first results regarding the tractability of (3)-(4) are given in Propositions 1-2 (Proposition 1(i) was already shown by Mudholkar and Hutson (2000)).

Proposition 1. *The asymmetric Normal log-likelihood in (3) satisfies:*

(i) Its gradient is continuous and is given by

$$g_1(\theta_\gamma,\vartheta,\alpha) = \begin{pmatrix} \frac{1}{\vartheta}X_\gamma^T W^2(y - X_\gamma\theta_\gamma) \\ -\frac{n}{2\vartheta} + \frac{1}{2\vartheta^2}(y - X_\gamma\theta_\gamma)^T W^2(y - X_\gamma\theta_\gamma) \\ \frac{1}{\vartheta}(y - X_\gamma\theta_\gamma)^T\overline{W}^3(y - X_\gamma\theta_\gamma) \end{pmatrix}.$$

(ii) Its Hessian with respect to θ_γ is continuous everywhere except on the zero Lebesgue measure set $\{\theta_\gamma \in \mathbb{R}^p : x_{\gamma i}^T\theta_\gamma = y_i$ for some $i = 1,\ldots,n\}$, and is $H_1(\theta_\gamma,\vartheta,\alpha) = \vartheta^{-1}\times$

$$\begin{pmatrix} -X_\gamma^T W^2 X_\gamma & \frac{1}{\vartheta}X_\gamma^T W^2(X_\gamma\theta_\gamma - y) & -2X_\gamma^T\overline{W}^3(y - X_\gamma\theta_\gamma) \\ \frac{1}{\vartheta}(X_\gamma\theta_\gamma - y)^T W^2 X_\gamma & \frac{n}{2\vartheta} - \frac{(y-X_\gamma\theta_\gamma)^T\overline{W}^2(y-X_\gamma\theta_\gamma)}{\vartheta^2} & -\frac{1}{\vartheta}(y - X_\gamma\theta_\gamma)^T\overline{W}^3(y - X_\gamma\theta_\gamma) \\ -2(y - X_\gamma\theta_\gamma)^T\overline{W}^3 X_\gamma & -\frac{1}{\vartheta}(y - X_\gamma\theta_\gamma)^T\overline{W}^3(y - X_\gamma\theta_\gamma) & -3(y - X_\gamma\theta_\gamma)^T W^4(y - X_\gamma\theta_\gamma) \end{pmatrix},$$

(iii) If rank$(X_\gamma) = p_\gamma$, then $H_1(\theta_\gamma,\vartheta,\alpha)$ is strictly negative definite with respect to (θ_γ,α) and (3) has a unique maximum $(\widehat{\theta_\gamma},\widehat{\vartheta},\widehat{\alpha})$. Alternatively, if rank$(X_\gamma) < p_\gamma$, then $H_1(\theta_\gamma,\vartheta,\alpha)$ is negative semidefinite.

The implication is that, analogously to Normal errors, when X_γ has full rank (3) is continuous and concave almost everywhere in (θ_γ,α). This fact, combined with $\log L_1$ having a continuous gradient, guarantees overall concavity and hence a unique maximum (see the proof for a formal argument). Further, inspection of (1) reveals that $\log L_1$ is locally quadratic as a function of θ_γ within regions of constant $A(\theta_\gamma)$ and that its maximizer with respect to (θ_γ,α) does not depend on ϑ, two observations that facilitate optimization.

Proposition 2 shows that, although $\log L_2$ is piecewise-linear in θ_γ and thus has a singular hessian, one can prove concavity and uniqueness of a maximum in terms of (θ_γ,α) as in Proposition 1, extending the well-known result of concavity with respect to only θ_γ (Koenker 2005). In Sections 4-5 we describe how this result facilitates computation, in particular leading to simple optimization and analytical approximations to integrated likelihoods, and asymptotic characterizations.

Proposition 2. *The asymmetric Laplace log-likelihood in (4) satisfies:*

(i) It is continuously differentiable with gradient

$$g_2(\theta_\gamma, \vartheta, \alpha) = \vartheta^{-\frac{1}{2}} \times \begin{pmatrix} -X_\gamma^T \overline{w} \\ -\frac{n}{2\vartheta^{\frac{1}{2}}} + \frac{1}{2\vartheta} w^T |y - X_\gamma \theta_\gamma| \\ |y - X_\gamma \theta_\gamma|^T \overline{w}^2 \end{pmatrix},$$

except on the zero Lebesgue measure set $\{\theta_\gamma \in \mathbb{R}^p : x_{\gamma i}^T \theta_\gamma = y_i$ *for some* $i = 1, \ldots, n\}$*, where the gradient is undefined.*

(ii) Its Hessian with respect to θ_γ *is continuous everywhere except on the zero Lebesgue measure set* $\{\theta_\gamma \in \mathbb{R}^p : x_{\gamma i}^T \theta_\gamma = y_i$ *for some* $i = 1, \ldots, n\}$*, and is* $H_2(\theta_\gamma, \vartheta, \alpha) = \vartheta^{-1/2} \times$

$$\begin{pmatrix} 0 & \frac{1}{2\vartheta} X_\gamma^T \overline{w} & X_\gamma^T w^2 \\ \frac{1}{2\vartheta} \overline{w}^T X_\gamma & \frac{n}{2\vartheta^{\frac{3}{4}}} - \frac{3}{4\vartheta^2} w^T |y - X_\gamma \theta_\gamma| & -\frac{1}{2\vartheta} |y - X_\gamma \theta_\gamma|^T \overline{w}^2 \\ (X_\gamma^T w^2)^T & -\frac{1}{2\vartheta} |y - X_\gamma \theta_\gamma|^T \overline{w}^2 & -2|y - X_\gamma \theta_\gamma|^T \overline{w}^3 \end{pmatrix}.$$

(iii) If rank$(X_\gamma) = p_\gamma$*, then (4) is strictly concave in* (θ_γ, α) *and has a unique maximum* $(\widehat{\theta_\gamma}, \widehat{\vartheta}, \widehat{\alpha})$*. Alternatively, if rank*$(X_\gamma) < p_\gamma$*, then it is non-strictly concave in* (θ_γ, α)*.*

Parameter estimates maximizing (3)-(4) can be interpreted as the best-fitting linear model under weighted least-squares or weighted least absolute deviations, respectively. Different weights are assigned to observations on each side of the estimated $x_i^T \theta$. The weights are determined by α, which captures residual asymmetry and converges to a unique KL-optimal value (Section 4.1). Selected variables can be interpreted in a similar fashion, essentially as defining the smallest model amongst those minimizing each criterion (Section 5.2). That is, variable selection can be understood in terms of optimal variable configurations under well-known criteria.

3. PRIOR FORMULATION

We complete the Bayesian model via priors on the model indicators γ and the model-specific parameters (θ_γ, α). For $p(\gamma)$ by default we adopt the standard Beta-Binomial(a_γ, b_γ) prior (Scott and Berger 2010) where $a_\gamma, b_\gamma > 0$ are known constants (by default $a_\gamma = b_\gamma = 1$), although our implementation also incorporates uniform and Binomial priors. The four posed residual distributions (Normal, asymmetric Normal, Laplace and asymmetric Laplace) are assigned equal prior probability independently from the variable inclusions. Therefore

$$p(\gamma) = \frac{1}{4}\frac{B(a_\gamma + \sum_{j=1}^p \gamma_j, b_\gamma + p - \sum_{j=1}^p \gamma_j)}{B(a_\gamma, b_\gamma)}, \tag{5}$$

where $B()$ is the Beta function. Any model with $p_\gamma > n$ is assigned $p(\gamma) = 0$, as it would result in data interpolation.

Regarding $p(\theta_\gamma \mid \gamma)$, given that the mode, mean and median of y_i are linear in $x_{\gamma i}^T\theta_\gamma$ the usual prior specification strategies under Normal errors remain sensible. The possibilities are too numerous to list here, see *e.g.* Bayarri et al. (2012) or Mallick and Nengjun (2013) and references therein. We focus on the class of NLPs introduced by Johnson and Rossell (2010), as these lead to stronger sparsity than conventional (local) priors and (under suitable conditions) consistency of posterior model probabilities in high-dimensional Normal regression where $p = o(n)$ (Johnson and Rossell 2012) or $\log p = o(n)$ (Shin et al. 2015). However our theory also applies to local priors. The basic intuition is that, under model γ, all elements in θ_γ are assumed to be non-zero. Thus, $p(\theta_\gamma \mid \gamma)$ should vanish as any element in θ_γ approaches 0. We focus on two specific choices (Johnson and Rossell 2012; Rossell et al. 2013)

$$p_M(\theta_\gamma \mid \vartheta, \gamma) = \prod_{\gamma_j=1} \frac{\theta_j^2}{kg_\theta\vartheta} N(\theta_j; 0, g_\theta k\vartheta), \tag{6}$$

$$p_E(\theta_\gamma \mid \vartheta, \gamma) = \prod_{\gamma_j=1} \exp\left\{\sqrt{2} - \frac{g_\theta k\vartheta}{\theta_j^2}\right\} N(\theta_j; 0, g_\theta k\vartheta), \tag{7}$$

called product MOM and eMOM priors (respectively), where g_θ is a known prior dispersion. For Normal or asymmetric Normal errors $k = 1$, and for the Laplace or asymmetric Laplace $k = 2$ as then $\text{Var}(\epsilon_i)$ is proportional to 2ϑ. Along the same lines for the scale parameter we set a standard inverse gamma $p(\vartheta \mid \gamma) = \text{IG}(\vartheta; a_\vartheta/2, kb_\vartheta/2)$ (in our examples $a_\vartheta = b_\vartheta = 0.01$). MOM vanishes at a quadratic speed around the origin and accelerates polynomial Bayes factor sparsity rates, whereas eMOM vanishes exponentially and leads to quasi-exponential rates (Johnson and

Rossell 2010; Rossell and Telesca 2017), a result we extend here for our new class of models and under model misspecification (Section 5). In our examples, we follow the default recommendation in Johnson and Rossell (2010) and set $g_\theta = 0.348, 0.119$ for MOM and eMOM (respectively), under the rationale that they assign 0.01 prior probability to $|\theta_i/\sqrt{\vartheta}| < 0.2$, *i.e.* effect sizes often deemed practically irrelevant. Naturally, whenever prior information is available we recommend using it to set g_θ. The supplementary material describes a third prior class called iMOM that provides a thick-tailed counterpart to the eMOM. Although the iMOM is implemented in our software, we do not consider it further here given that its performance was very similar to the eMOM but it has the unappealing property of leading to non-convex optimization (akin to other thick-tailed priors, *e.g.* Cauchy), and when considering $p(\alpha)$ (see below) it leads to a density that diverges on the boundary ($\alpha = -1$ or $\alpha = 1$).

To set $p(\alpha \mid \gamma_{p+1} = 1)$ ($\alpha = 0$ under $\gamma_{p+1} = 0$) we reparameterize $\tilde{\alpha} = \text{atanh}(\alpha) \in \mathbb{R}$ as in Rubio and Steel (2014). These authors proposed $0.5(1+\alpha) \sim \text{Beta}(2,2)$, which places the prior mode at $\alpha = 0$ and thus defines a local prior. Our goal here is to detect situations where the degree of asymmetry is practically relevant and to otherwise allow the posterior to collapse on the symmetric model. To achieve this, we consider $p_M(\tilde{\alpha} \mid \gamma_{p+1} = 1) = \tilde{\alpha}^2 \phi(\tilde{\alpha}/\sqrt{g_\alpha})/\sqrt{g_\alpha}$, and $p_E(\tilde{\alpha} \mid \gamma_{p+1} = 1) = e^{\sqrt{2} - g_\alpha/\tilde{\alpha}^2} N(\tilde{\alpha}; 0, g_\alpha)$, where $g_\alpha \in \mathbb{R}^+$ is a fixed prior dispersion parameter. To set g_α, by default we consider that Arnold-Groeneveld asymmetry coefficients $|\alpha| < 0.2$ are often practically irrelevant. Thus, we set g_α such that $P(|\alpha| \geq 0.2) = 0.99$. Also, note that $\alpha = 2$ gives a total variation distance of $|\alpha|/2 = 0.1$, *i.e.* the largest difference $|P(\epsilon_i \in A \mid \alpha = 0) - P(\epsilon_i \in A \mid \alpha)|$ for any set A is 0.1, which we typically view as irrelevant. Since $\text{atanh}(0.2) = 0.203$, a direct calculation gives that $P(|\tilde{\alpha}| \geq 0.203) = 0.99$ when $g_\alpha = 0.357, 0.122$ under MOM and eMOM. To assess sensitivity in our examples, we also considered g_α such that $P(|\alpha| \geq 0.1) = 0.99$ (total variation distance=0.05), giving $g_\alpha = 0.087, 0.030$. Figure 1 depicts $p(\alpha)$ under these settings. Our results showed that variable selection is typically robust to choices of g_α within this range.

[Figure 1 about here.]

4. PARAMETER ESTIMATION

We obtain some results for parameter estimation under a given γ that are also useful to establish variable selection rates (see Section 5 for results on Bayesian model averaging). Section 4.1 gives the limiting distribution of $(\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma) = \arg\max_{\theta_\gamma,\vartheta,\alpha} \log L_k(\theta_\gamma, \vartheta, \alpha)$ as $n \to \infty$ for asymmetric Normal ($k = 1$) and Laplace ($k = 2$) when data are generated from (1) but the error model may be misspecified. Briefly, as is typically the case, we obtain parameter estimation consistency and asymptotic normality, albeit there is a loss of efficiency and an underestimation of uncertainty. Section 4.2 presents novel optimization algorithms for maximum likelihood and posterior mode estimation designed to improve the computational scalability of current related methods.

4.1 Asymptotic distributions

We lay out technical conditions for our asymptotic results to hold.

A1. The parameter space $\Gamma \subset \mathbb{R}^p \times \mathbb{R}_+ \times (-1, 1)$ is compact and convex.

A2. Data are truly generated as $y_i = x_i^T \theta^* + \epsilon_i$ for some $\theta^* \in \mathbb{R}^p$, fixed $p_{\gamma^*} = \sum_{j=1}^p \mathrm{I}(\theta_j^* \neq 0)$ and ϵ_i are *i.i.d.* and independent of x_i. Let the data-generating $y_i|x_i \overset{i.i.d.}{\sim} S_0(\cdot|x_i)$ with density $s_0(y_i \mid x_i) > 0$ for all y_i.

A3. For all γ there is some n_0 such that $X_\gamma^T X_\gamma$ is strictly positive definite almost surely for all $n > n_0$.

A4. Denote by $x_i \overset{i.i.d.}{\sim} \Psi(\cdot)$ the generating process of the covariates (which can be either stochastic or deterministic).

$$\int |y_1|^j dS_0(y_1|x_1) d\Psi(x_1) < \infty,$$
$$\int ||x_1||^j d\Psi(x_1) < \infty,$$

where $j = 1, 2,$ or 4, and we specify the order j of interest in each of the results below, and $||\cdot||$ denotes the Euclidean distance $||z|| = (\sum z_i^2)^{\frac{1}{2}}$.

A5. For $\eta \in \Gamma$

$$\begin{aligned}
\int \frac{\partial}{\partial \eta_j}\left[\int m_\eta(y_1, x_1) dS_0(y_1|x_1)\right] d\Psi(x_1) &= \frac{\partial}{\partial \eta_j}\int\int m_\eta(y_1, x_1) dS_0(y_1|x_1) d\Psi(x_1), \\
\int \frac{\partial^2}{\partial \eta_i \eta_j}\left[\int m_\eta(y_1, x_1) dS_0(y_1|x_1)\right] d\Psi(x_1) &= \frac{\partial^2}{\partial \eta_i \eta_j}\int\int m_\eta(y_1, x_1) dS_0(y_1|x_1) d\Psi(x_1).
\end{aligned}$$

These conditions are in line with those in classical robust regression, *e.g.* see Huber (1973) or Koenker and Bassett (1982). Condition A1 is made out of technical convenience, naturally one may take an arbitrarily large Γ. Condition A2 states that data truly arise from a linear model, where the key assumption is that the residuals are independent. Extensions to non-id errors are discussed in Section 5.2. Condition A3 holds whenever the rows of X are regarded as a deterministic sequence satisfying the condition, or for instance when x_i are independent and identically distributed from an underlying distribution of fixed dimension with positive-definite $\text{Cov}(x_1)$, as then $X^T X$ converges almost surely to a positive-definite matrix by the strong law of large numbers. We focus on fixed p, extensions to p growing with n are possible along the lines in Mendelson (2014), but its detailed treatment is beyond the scope of this paper. Condition A4 requires existence of moments up to a certain order. Condition A5 requires being able to exchange integration and differentiation, and is needed only to prove asymptotic normality.

Our results summarize and extend classical studies focusing on θ_γ in least squares, median and quantile regression to consider the whole parameter vector $(\theta_\gamma, \vartheta, \alpha)$. Briefly, Eicker (1964) and Srivastava (1971) showed that the least squares estimator ($k = 1, \alpha = 0$) satisfies $\sqrt{n}V^T(\widehat{\theta}_\gamma - \theta_0) \xrightarrow{D} N(0, \text{Var}(\epsilon_1)I)$, where θ_0 minimizes Kullback-Leibler divergence to the data-generating truth and $VV^T = X_\gamma^T X_\gamma/n$, assuming that $\text{Var}(\epsilon_1) < \infty$ and minimum conditions on $X_\gamma^T X_\gamma$. To our knowledge, the asymmetric Normal has been much less studied, *e.g.* Kimber (1985), Mudholkar and Hutson (2000) and Arellano-Valle et al. (2005) considered the case with no covariates and no checks of the conditions required by large sample theory are shown, which are non-trivial given that $H_1(\theta_\gamma, \vartheta, \alpha)$ is discontinuous. Regarding Laplace errors ($k = 2, \alpha = 0$), Pollard (1991) and Knight (1999) showed $2f_0\sqrt{n}V^T(\widehat{\theta}_\gamma - \theta_0) \xrightarrow{D} N(0, I)$, where $f_0 = p(\epsilon_0)$ and ϵ_0 is the median of $s_0(\epsilon_i)$, under mild conditions on $X_\gamma^T X_\gamma$ and $f_0 > 0$. Koenker (1994) generalized the result to the asymmetric Laplace, obtaining $2f_0\sqrt{n/(1-\alpha^2)}V^T(\widehat{\theta}_\gamma - \theta_0) \xrightarrow{D} N(0, I)$, where $f_0 = p(\epsilon)$ evaluated at the τ^{th}

quantile $\epsilon = S_0^{-1}(\tau)$, where in our parameterization $\tau = (1+\alpha)/2$. Proposition 3 establishes the consistency of the maximum likelihood estimator $\widehat{\eta}_\gamma = (\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma)$ to the Kullback-Leibler optimal parameter values, whereas Proposition 4 gives asymptotic normality.

Proposition 3. *Assume Conditions A1–A4 with $p < n$, where $j = 2$ in A4 when $k = 1$ and $j = 1$ when $k = 2$. Then, the function $M_k(\theta_\gamma, \vartheta, \alpha) = \mathbb{E}[\log L_k(y_1 | x_1^T \theta_\gamma, \vartheta, \alpha)]$ has a unique maximizer $(\theta_\gamma^*, \vartheta_\gamma^*, \alpha_\gamma^*) = \text{argmax}_\Gamma M_k(\theta_\gamma, \vartheta, \alpha)$. Moreover, the maximum likelihood estimator $(\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma) \xrightarrow{P} (\theta_\gamma^*, \vartheta_\gamma^*, \alpha_\gamma^*)$ as $n \to \infty$.*

Proposition 4. *Assume Conditions A1–A5, with $j = 4$ in A4 when $k = 1$ and $j = 2$ when $k = 2$. Denote $\eta = (\theta_\gamma, \vartheta, \alpha)$, $m_\eta(y_1, x_1) = \log s_k(y_1 | x_1^T \theta_\gamma, \vartheta, \alpha)$, $Pm_\eta = \mathbb{E}\left[m_\eta(y_1, x_1)\right]$, and $\eta_\gamma^* = (\theta_\gamma^*, \vartheta_\gamma^*, \alpha_\gamma^*) = \text{argmax}_\Gamma Pm_\eta$. Then, the sequence $\sqrt{n}(\widehat{\eta}_\gamma - \eta_\gamma^*)$ is asymptotically Normal with mean 0 and covariance matrix $V_{\eta_\gamma^*}^{-1} \mathbb{E}[\dot{m}_{\eta_\gamma^*} \dot{m}_{\eta_\gamma^*}^T] V_{\eta_\gamma^*}^{-1}$, where $\dot{m}_{\eta_\gamma^*}$ is the gradient of $m_\eta(\cdot)$, with respect to η, evaluated at η_γ^* and $V_{\eta_\gamma^*}$ is the second derivative matrix of Pm_η evaluated at η_γ^*.*

The sandwich covariance $V_{\eta_\gamma^*}^{-1} \mathbb{E}[\dot{m}_{\eta_\gamma^*} \dot{m}_{\eta_\gamma^*}^T] V_{\eta_\gamma^*}^{-1}$ is typically an inflated version of that obtained when the true model is assumed ($V_{\eta_\gamma^*}^{-1}$), implying the well-known consequence of model misspecification that parameter estimation suffers a loss of efficiency and uncertainty is underestimated. To gain insight, Corollary 1 gives specific asymptotic variances under various model misspecification cases. For instance, when truly $\epsilon_i \sim N(0, \vartheta)$ wrongly assuming Laplace errors increases the variance by a factor $\pi/2$, and a similar phenomenon is observed when ignoring the presence of residual asymmetry. We defer discussion of the implications for variable selection to Section 5 and the examples in Section 6.

Corollary 1. *The asymptotic distribution of $\widehat{\theta}_\gamma$ obtained by maximizing either the Normal, ANormal, Laplace or ALaplace likelihood is $V(\widehat{\theta}_\gamma - \theta_\gamma^*) \xrightarrow{D} N(0, vI)$, for some $v > 0$. The asymptotic variances v, when ϵ_i truly arise* i.i.d. *under four specific distributions, are given below.*

	Maximized log-likelihood			
True model	*Normal*	*ANormal*	*Laplace*	*ALaplace*
$N(0,\vartheta)$	ϑ	ϑ	$\frac{\pi}{2}\vartheta$	$\frac{\pi}{2}\vartheta$
$AN(0,\vartheta,\alpha)$	$\vartheta(1+0.454\alpha^2)$	$\vartheta(1-\alpha^2)$ $(\star)$	$\frac{\pi}{2}\vartheta k_\alpha$	$\dfrac{\pi}{2}\vartheta(1-\alpha_\gamma^{*2})$
$L(0,\vartheta)$	2ϑ	2ϑ	ϑ	ϑ
$AL(0,\vartheta,\alpha)$	$2\vartheta(1+\alpha^2)$	$2\vartheta w_{\alpha,\alpha_\gamma^*}$ $(\star)$	$\vartheta(1+\lvert\alpha\rvert)^2$	$\vartheta(1-\alpha^2)$

where $k_\alpha = \exp\left\{\left[\Phi^{-1}\left(\frac{1}{2(1+|\alpha|)}\right)\right]^2\right\} \geq 1,\ w_{\alpha,\alpha_\gamma^*} = \dfrac{(1+\alpha)^2 - 2\alpha\left(1+\alpha_\gamma^*\right)}{\left(1-\alpha^2\right)^2} \in [0,1],$ *and* α_γ^* *is as in Proposition 4. Cases marked* $(\star)$ *were derived assuming that covariates have zero mean.*

4.2 Optimization

We outline simple, efficient algorithms to obtain $(\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma) = \arg\max_{\theta_\gamma,\vartheta,\alpha} \log L_k(\theta_\gamma, \vartheta, \alpha)$, where $k \in \{1,2\}$ are the asymmetric Normal and Laplace log-likelihoods (3)-(4). We also consider the corresponding posterior modes $(\tilde{\theta}_\gamma, \tilde{\vartheta}_\gamma, \tilde{\alpha}_\gamma) = \arg\max_{\theta_\gamma,\vartheta,\alpha} \log L_k(\theta_\gamma, \vartheta, \alpha) + \log p(\theta_\gamma, \vartheta, \alpha \mid \gamma)$, where $p(\theta_\gamma, \vartheta, \alpha \mid \gamma)$ is the prior density (Section 3). The algorithms are useful to obtain parameter estimates or Laplace approximations to the integrated likelihood. Mudholkar and Hutson (2000) and Arellano-Valle et al. (2005) gave an algorithm to obtain $\widehat{\theta}_\gamma$ for $\log L_1$ in the case with no covariates ($p_\gamma = 1$). To tackle point discontinuities in the derivatives their algorithm requires solving n separate optimization problems, which does not scale up with increasing n, or alternatively using method of moments estimators. Maximum likelihood estimation of θ_γ under the asymmetric Laplace and fixed α is connected to quantile regression (see below). Regarding Bayesian frameworks, most rely on MCMC for parameter estimation but this is too costly when we wish to consider a potentially large number of models. Instead, we propose a generic framework for jointly obtaining $(\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma)$ or $(\tilde{\theta}_\gamma, \tilde{\vartheta}_\gamma, \tilde{\alpha}_\gamma)$ applicable to both the asymmetric Normal and Laplace. The key result we exploit is concavity of the log-likelihood given by Propositions 1-2, which allows iteratively optimizing first θ_γ and then (ϑ, α). Optimization with respect to (ϑ, α) has closed form, whereas setting θ_γ can be seen as weighted least squares for the asymmetric Normal and as quantile regression for the asymmetric Laplace. The latter task of maximizing $\log L_2$ with respect to θ_γ is a classical problem that can be framed as linear programming, for which simplex and interior-point methods are available. However, these are not applicable to the posterior mode as the target is no longer piecewise linear and even efficient implementations have computational complexity greater than cubic in p and supra-linear in n (Koenker 2005).

We outline two simple algorithms that have lower complexity and can be readily adapted to obtain the posterior mode. Briefly, in Algorithm 1, Step 2 follows from setting first derivatives to zero and directly extends Mudholkar and Hutson (2000) (Proposition 4.4) and Arellano-Valle et al. (2005) (Section 4.2). Step 3 is essentially a Levenberg-Marquardt algorithm (Levenberg 1944; Marquardt 1963) exploiting gradient continuity. g_θ and H_θ denote the gradient and hessian with respect to θ_γ as in Propositions 1-2, where for $\log L_2()$ we use the asymptotic hessian $X^T X/(\vartheta(1-\alpha^2))$. Its updates are in between those of a Newton-Raphson and gradient descent algorithms and can be interpreted as restricting the Newton-Raphson step to a trust region where the quadratic approximation is accurate (Sorensen 1982). For large regularization parameter λ the update δ converges to the gradient algorithm, which by continuity is guaranteed to increase the target, whereas for small λ it converges to the Newton-Raphson algorithm, achieving quadratic convergence as $\theta_\gamma^{(t)}$ approaches the optimum.

Algorithm 1. Optimization via Levenberg-Marquardt

1. *Initialize* $\widehat{\theta}_\gamma^{(0)} = (X^T X)^{-1} X^T y$, $\lambda = 0$. *Set* $t = 1$

2. *Let* $s_1 = \sum_{i \in A(\widehat{\theta}_\gamma^{(t-1)})} |y_i - x_{\gamma i}^T \widehat{\theta}_\gamma^{(t-1)}|^{3-k}$, $s_2 = \sum_{i \notin A(\widehat{\theta}_\gamma^{(t-1)})} |y_i - x_{\gamma i}^T \widehat{\theta}_\gamma^{(t-1)}|^{3-k}$. *Update*

$$\widehat{\alpha}^{(t)} = \frac{s_1^{\frac{k}{2+k}} - s_2^{\frac{k}{2+k}}}{s_1^{\frac{k}{2+k}} + s_2^{\frac{k}{2+k}}}; \widehat{\vartheta}^{(t)} = \frac{1}{4n^k} \left(s_1^{\frac{k}{2+k}} + s_2^{\frac{k}{2+k}} \right)^{2+k}.$$

3. *Propose* $m = \theta_\gamma^{(t-1)} + \delta$, *where*

$$\delta = -\left(H_\theta + \lambda diag(H_\theta) \right)^{-1} g_\theta,$$

and g_θ, H_θ *are the subsets of* $g_k(\widehat{\theta}_\gamma^{(t-1)}, \widehat{\vartheta}^{(t)}, \widehat{\alpha}^{(t)})$ *and* $H_k(\widehat{\theta}_\gamma^{(t-1)}, \widehat{\vartheta}^{(t)}, \widehat{\alpha}^{(t)})$ *corresponding to* θ_γ. *If* $\log L_k(m, \vartheta^{(t)}, \alpha^{(t)}) > \log L_k(\theta_\gamma^{(t-1)}, \vartheta^{(t)}, \alpha^{(t)})$ *set* $\theta_\gamma^{(t)} = m$ *and* $\lambda = \lambda/2$, *else update* $\lambda = 1 + \lambda$ *and repeat Step 3.*

Given a good initial guess $\widehat{\theta}_\gamma^{(0)}$, the fact that $\log L_k$ are locally well approximated by a quadratic function in θ_γ ($\log L_1$ is exactly locally quadratic) results in Algorithm 1 usually converging after a few iterations. As usual, with second-order optimization each iteration requires a matrix inversion that is costly when p is large. As an alternative, Algorithm 2 uses coordinate descent to optimize each $\theta_{\gamma j}$ sequentially, which only requires univariate updates, where updating the set $A(\theta_\gamma)$ for each $\theta_{\gamma j}$ implies that Step 3 has cost $O(np)$. In contrast, Algorithm 1 determines $A(\theta_\gamma)$ once per iteration and performs matrix inversion, with total cost $O(n + p^3)$ per iteration. Hence, although Algorithm 1 usually requires fewer iterations than Algorithm 2, for large p the latter is typically preferrable. A related study of computational cost is offered in Breheny and Huang (2011) in the context of penalized likelihood optimization, who found that coordinate descent is often preferrable to multivariate updates. These results show that, contrary to historical beliefs, two-piece distributions lead to convenient optimization. R package mombf (Rossell et al. 2016) incorporates both algorithms but our examples are based on Algorithm 2, the results were essentially identical to those of Algorithm 1 but the running time was substantially shorter.

We adapted both algorithms to find the posterior mode by simply redefining g_k and H_k to be the gradient and Hessian of $\log L_k(\theta_\gamma, \vartheta, \alpha) + \log p(\theta_\gamma, \vartheta, \alpha \mid \gamma)$. The corresponding expressions are in Supplementary Section 3.2. We remark that due to the penalty around the origin NLPs such as $p_M()$ and $p_E()$ in (6)-(7) are not log-concave, however this is not an issue as they are symmetric and log-concave in each quadrant (fixed $\text{sign}(\theta_\gamma, \alpha)$). Thus $\log p(\theta_\gamma, \vartheta, \alpha \mid y, \gamma)$ is concave in each quadrant, its unique global mode lies in the same quadrant as the maximum likelihood estimator and we may initialize the algorithm at $(\tilde{\theta}_\gamma^{(0)}, \tilde{\vartheta}^{(0)}, \tilde{\alpha}^{(0)}) = (\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma)$. Convergence is typically achieved after a few iterations.

Algorithm 2. Optimization via coordinate descent

1. *Set an arbitrary $c > 1$ and initialize $\theta_\gamma^{(0)}$, $\lambda = 0$ as in Algorithm 1.*

2. *Update $(\widehat{\vartheta}^{(t)}, \widehat{\alpha}^{(t)})$ as in Algorithm 1.*

3. *For $j = 1, \ldots, p_\gamma$, let $m = \theta_{\gamma j}^{(t-1)} - \frac{g_j}{h_{jj}(1+\lambda)}$, where g_j is the j^{th} element in $g_1(\theta_\gamma)$ and h_{jj} the (j,j) element in $H_1(\theta_\gamma)$ at $\theta_\gamma = (\theta_{\gamma 1}^{(t)}, \ldots, \theta_{\gamma j-1}^{(t)}, \theta_{\gamma j}^{(t-1)}, \ldots, \theta_{\gamma p_\gamma}^{(t-1)})$. If L_k evaluated at $\theta_{\gamma j}^{(t)} = m$ increases, set $\theta_{\gamma j}^{(t)} = m$, $\lambda = \lambda/c$, else iteratively update $\lambda = c + \lambda$ and m until L_k increases.*

5. MODEL SELECTION

Under a standard Bayesian framework $p(\gamma \mid y) = p(y \mid \gamma)p(\gamma)/p(y)$, with integrated likelihood

$$\begin{aligned}
p(y \mid \gamma) &= \int L_1(\theta_\gamma, \vartheta, 0)p(\theta_\gamma, \vartheta)d\theta_\gamma d\vartheta, \text{ if } \gamma_{p_\gamma+1} = 0, \gamma_{p_\gamma+2} = 0,\\
p(y \mid \gamma) &= \int L_1(\theta_\gamma, \vartheta, \alpha)p(\theta_\gamma, \vartheta, \alpha)d\theta_\gamma d\vartheta d\alpha, \text{ if } \gamma_{p_\gamma+1} = 1, \gamma_{p_\gamma+2} = 0,\\
p(y \mid \gamma) &= \int L_2(\theta_\gamma, \vartheta, 0)p(\theta_\gamma, \vartheta)d\theta_\gamma d\vartheta, \text{ if } \gamma_{p_\gamma+1} = 0, \gamma_{p_\gamma+2} = 1,\\
p(y \mid \gamma) &= \int L_2(\theta_\gamma, \vartheta, \alpha)p(\theta_\gamma, \vartheta, \alpha)d\theta_\gamma d\vartheta d\alpha, \text{ if } \gamma_{p_\gamma+1} = 1, \gamma_{p_\gamma+2} = 1. \qquad (8)
\end{aligned}$$

Section 5.1 discusses how to compute $p(y \mid \gamma)$ and Section 5.2 the asymptotic properties of the associated Bayes factors and Bayesian model averaging, along with a discussion on model misspecification and to what extent these results can be generalized to non-identically distributed errors (e.g. under heteroscedasticity or hetero-asymmetry). Section 5.3 outlines a stochastic model search algorithm that can be used when p is too large for exhaustive enumeration of the 2^{p+2} models.

5.1 Integrated likelihood

Computing (8) in the case $\gamma_{p+1} = \gamma_{p+2} = 0$ corresponds to Normal linear regression, for which existing methods are typically available, *e.g.* Johnson and Rossell (2012) gave closed-form expressions for the MOM and Laplace approximations for the eMOM. The three remaining cases require numerical evaluation, for which we propose Laplace and Monte Carlo approximations. The former are appealing due to log-likelihood concavity and asymptotic normality (Section 4). Indeed, in our examples they delivered very similar inference and were orders of magnitude faster than Monte Carlo. Hence, by default we recommend Laplace approximations over Monte Carlo, except in small p situations where the latter is still practical. To ensure that the parameter support is on the real numbers Laplace approximations are based on the reparameterization $\eta = (\theta_\gamma, \log(\vartheta), \text{atanh}(\alpha))$ and given by

$$\widehat{p}(y \mid \gamma) = \exp\{\log L_k(\tilde{\eta}) + \log p(\tilde{\eta})\} \frac{(2\pi)^{\sum_{j=1}^{p+2} \gamma_j/2}}{|H_k(\tilde{\eta})|^{1/2}}, \qquad (9)$$

where $k = 1, 2$ for $\gamma_{p+2} = 0, 1$ respectively, $\tilde{\eta}$ and $H_k(\tilde{\eta})$ are the posterior mode and hessian of $\log L_k(\eta) + \log p(\eta)$. The specific expressions are given in Supplementary Section 3. Expression (9) simply requires the posterior mode (Algorithms 1-2) and evaluating the hessian. The latter is straightforward for $k = 1$, but for $k = 2$ it is singular in θ_γ, requiring some care. The reasoning behind (9) is to approximate the log-integrand in (8) by a smooth function that has strictly positive definite hessian in θ_γ, which is facilitated in our setting by $\log L_2$ concavity and asymptotic normality. We found that a simple yet effective strategy is to replace H_2 by the asymptotic expected hessian $\overline{H}_2$ obtained under independent asymmetric Laplace errors.

Although we did not find the following concern to be a practical issue in our examples, we remark that in principle $\overline{H}_2$ may underestimate the underlying uncertainty in θ_γ and thus inflate $|\overline{H}_2|$, *e.g.* under truly non-Laplacian independent and identically distributed errors one needs to add a multiplicative constant (Section 4.1), whereas independent but heteroscedastic errors require a matrix-reweighting adjustment (Kocherginsky et al. 2005). Typical strategies to improve the estimated curvature rely either on direct estimation under the assumption of independent and identically distributed errors, or indirect estimation via inversion of score tests, although these only provide univariate confidence intervals and their cost does not scale well with p, or sampling-based methods such as bootstrap or Monte Carlo. As a practical alternative here we consider that the goal is really to approximate the actual curvature of $\log L_2$, which can be easily done with a few point evaluations of $\log L_2$ in a neighbourhood of $\tilde{\eta}_\gamma$. Briefly, we consider the adjustment $D\overline{H}_2 D$, where D is a diagonal matrix such that its element d_{ii} gives the best approximation of $\log L_2$ as a quadratic function of θ_i in the least squares sense. $D\overline{H}_2 D$ matches the actual curvature in $\log L_2$ and is thus less dependent on asymptotic theory than other strategies, and has the advantage that D can be computed quickly. See Supplementary Section 3.3 for further details and Supplementary Figure 1S for an example. Given that the unadjusted $\overline{H}_2$ performed well in our examples and the associated results were practically indistinguishable to those based on Monte Carlo, unless otherwise stated our results are based on $\overline{H}_2$.

As our Monte Carlo alternative, we implemented an importance sampling estimator based on multivariate T draws and covariance matching the asymptotic posterior covariance. Specifically, let $\eta^{(b)} \sim T_3(\tilde{\eta}, \tilde{H}_k^{-1}/3)$ for $b = 1, \ldots, B$ where B is a large integer, then

$$\widehat{p}_I(y \mid \gamma) = B^{-1} \sum_{b=1}^{B} L_k(\eta^{(b)}) p(\eta^{(b)}) / T_3(\eta^{(b)}; \tilde{\eta}, \tilde{H_k}^{-1}/3). \tag{10}$$

We remark that NLPs are multimodal in (θ_γ, α), thus some care is needed when using Laplace approximations. To give an honest characterization of the properties of our preferred computational method, in Section 5 we obtain asymptotic rates for Bayes factors based on $\hat{p}(y \mid \gamma)$ in (9). Rossell and Telesca (2017) studied the discrepancies between $p(y \mid \gamma)$ and $\hat{p}(y \mid \gamma)$ for MOM, iMOM and eMOM priors and Normal errors. Briefly, given that secondary modes vanish asymptotically for truly active covariates but not for spurious covariates, $\hat{p}(y \mid \gamma)$ imposes a stronger penalty on spurious variables than $p(y \mid \gamma)$, however for such models $p(y \mid \gamma)$ decreases fast enough that both approximations typically lead to very similar inference.

5.2 Bayes factor rates

Let $\gamma^* = (\mathrm{I}(\theta_1^* \neq 0), \ldots, \mathrm{I}(\theta_p^* \neq 0), \mathrm{I}(\alpha^* \neq 0), \mathrm{I}(k^* = 2))$ be the optimal model, that is $(\theta^*, \vartheta^*, \alpha^*, k^*) = \arg\max_{\Gamma,k} M_k(\theta, \vartheta, \alpha)$ maximize the expected log-likelihood across $k = 1, 2$, and the expectation is with respect to the data-generating density in Condition A1. We indicate by $\gamma^* \subset \gamma$ that γ^* is a submodel of γ, *i.e.* $\gamma_j^* \leq \gamma_j$ for $j = 1, \ldots, p+1$, and by $\gamma^* \not\subset \gamma$ that $\gamma_j^* > \gamma_j$ for some j. If the data were truly generated from the assumed error distribution, it is well-known that the Bayes factor in favour of γ decreases exponentially with n when $\gamma^* \not\subset \gamma$ (γ is missing important variables). Conversely when γ adds spurious variables to γ^* the Bayes factor is only $O_p(n^{-(p_\gamma - p_{\gamma^*})/2})$ under local priors, an imbalance that is ameliorated under NLPs, which achieve faster polynomial or quasi-exponential rates depending on their chosen parametric form (Johnson and Rossell 2010; Johnson and Rossell 2012). Proposition 5 gives an extension under model misspecification, the first result of this kind for NLPs. We remark that the rates apply directly to the Laplace approximations (9). As studied by Rossell and Telesca (2017) (Supplementary Section 5, Supplementary Figure 8), when γ contains spurious parameters the non-local posterior $p(\theta_\gamma, \alpha_\gamma \mid \gamma, y)$ can have non-vanishing multimodality, in which case Laplace approximations $\hat{p}(y \mid \gamma)$ underestimate $p(y \mid \gamma)$ even as $n \to \infty$. In our experience this is not a major concern (e.g. Table 3S compares Laplace with importance sampling estimates), but we find it preferrable to characterize inference under our recommended computational framework, i.e. for $\hat{p}(y \mid \gamma)$. A critical condition for Proposition 5 is that the prior density be strictly positive at the optimum, $p(\theta^*_{\gamma^*}, \alpha^*_{\gamma^*} \mid \gamma^*) > 0$, which is trivially satisfied by pMOM and peMOM priors. It also holds for local priors, which for simplicity we define as $p(\theta_\gamma, \vartheta, \alpha \mid \gamma) > 0$ for all $(\theta_\gamma, \vartheta, \alpha) \in \Gamma_\gamma$ and we assume to be continuous.

Proposition 5. *Suppose that Conditions A1-A3 hold, fixed* p_γ, p_{γ^*} *and* $n \to \infty$. *If* $\gamma^* \not\subset \gamma$ *then* $\frac{1}{n} \log(\widehat{p}(y \mid \gamma)/\widehat{p}(y \mid \gamma^*)) \xrightarrow{P} -a_1$ *for local, pMOM and peMOM priors and some constant* $a_1 > 0$.

Conversely, if $\gamma^* \subset \gamma$ *then* $\widehat{p}(y \mid \gamma)/\widehat{p}(y \mid \gamma^*) = O_p(b_n)$ *where* $b_n = n^{-(p_\gamma - p_{\gamma^*})/2}$ *for local priors,* $b_n = n^{-3(p_\gamma - p_{\gamma^*})/2}$ *for the pMOM prior, and* $b_n = e^{-c\sqrt{n}}$ *for the peMOM prior where* $c > 0$.

Corollary 2. *Let* $E(\theta_i \mid y) = \sum_\gamma E(\theta_i \mid y, \gamma) p(\gamma \mid y)$ *be Bayesian model averaging estimates,* $r^+ = max_{p_\gamma = p_{\gamma^*}+1} p(\gamma)/p(\gamma^*)$, $r^- = max_{p_\gamma \leq p_{\gamma^*}} p(\gamma)/p(\gamma^*)$, *where* $p(\gamma)$ *is non-increasing in* p_γ *and* $\log r^- = O(n)$. *Under the conditions in Proposition 5, if* $\theta^*_i = 0$ *then* $E(\theta_i \mid y) = r^+ O_p(n^{-2})$ *under the pMOM prior and* $r^+ O_p(e^{-c\sqrt{n}})$ *under the peMOM prior. If* $\theta^*_i \neq 0$ *then* $E(\theta_i \mid y) = \theta^*_i + O_p(n^{-1/2})$ *under the pMOM and peMOM priors.*

Proposition 5 implies model selection consistency with Bayes factor rates that have the same functional form as when the correct model is assumed. We emphasize that this does not imply that there is no cost due to assuming an incorrect model: the coefficient a_1 in the exponential or those in the polynomial rates are affected. The constant a_1 determines how quickly one can detect truly active variables (asymptotically) and is given by the KL divergence between the assumed model class and the data-generating truth. That is, under the true model a_1 takes a different value than under a misspecified model and hence the ratio of the correct versus misspecified Bayes factors to detect signals is essentially exponential in n. In contrast, when $\gamma^* \subset \gamma$ this ratio converges to a constant, hence the effects of model misspecification on false positives vanishes asymptotically. We remark that, for finite n, misspecification can have a marked effect on false positives, see Section 6.2 for examples. Corollary 2 is the trivial implication that Bayes factors also drive parameter estimation shrinkage in a Bayesian model averaging setting (Rossell and Telesca 2017). When $\theta^*_i = 0$, the shrinkage is $1/n^2$ or $e^{-\sqrt{n}}$ for pMOM and peMOM respectively, in contrast to $1/n$ for local priors and $1/\sqrt{n}$ for the unregularized MLE, times a term given by model prior probabilities.

We remark that Conditions A1-A3 for Proposition 5 assume independent and identically distributed (id) errors. It is possible to relax these conditions, particularly that of id errors. Loosely speaking, the three main ingredients in the proof are that $(\hat{\theta}_\gamma, \hat{\alpha}_\gamma, \hat{\vartheta}_\gamma) \xrightarrow{P} (\theta^*_\gamma, \alpha^*_\gamma, \vartheta^*_\gamma)$ (MLE consistency), that asymptotically $P\left(n^{-p_\gamma} |H_k(\tilde{\eta}_\gamma)| \in [c_1, c_2]\right) \longrightarrow 1$ for some constants $c_1 > 0, c_2 > 0$, and that the likelihood ratio statistic between γ^* and a supra-model γ is bounded in probability. The MLE and likelihood ratio conditions hold quite generally for non-id errors, in particular the latter is satisfied whenever its limiting distribution is say a chi-square or mixture of chi-squares. Regarding H_k, under independent but non-id errors the ALaplace model has $H_2^{-1} = s(X_\gamma^T F_\gamma X_\gamma)^{-1}(X_\gamma^T X_\gamma)(X_\gamma^T F_\gamma X_\gamma)^{-1}$, where $s > 0$ is a constant depending on α and F_γ an $n \times n$ diagonal matrix accounting for each observation's variance (Kocherginsky et al. 2005). The Laplace model is a particular case of this result. Under Normal errors the MLE has the non-asymptotic covariance $H_1^{-1} = (X_\gamma^T X_\gamma)^{-1} X_\gamma^T \mathrm{Cov}(\epsilon) X_\gamma (X_\gamma^T X_\gamma)^{-1}$, and similarly for the asymmetric least squares criterion implied by the two-piece Normal. Provided that $\max_{i=1,\ldots,n} \mathrm{Var}(\epsilon_i)$ is bounded or grows at a slower-than-polynomial rate with n and the eigenvalues of $n(X_\gamma^T X_\gamma)^{-1}$ lie between two positive constants, then $P\left(n^{-p_\gamma} |H_k(\tilde{\eta}_\gamma)| \in [c_1, c_2]\right) \longrightarrow 1$ for some $c_1 > 0, c_2 > 0$. Relaxing the independence assumption requires more care, *e.g.* under very strong dependence $|H_k|$ could grow at a slower rate than n^{p_γ}. We remark that these observations are simply meant to provide intuition, obtaining precise conditions for Proposition 5 under non-iid settings is an interesting question for future research.

From the discussion above model misspecification affects sensitivity via the constant a_1. In our experience, typically there is a loss of power. Fully characterizing this issue theoretically is complicated as a_1 depends on the unknown data-generating truth, but it is possible to provide some intuition. Consider an arbitrary variable configuration $(\gamma_1, \ldots, \gamma_p) \not\subset (\gamma_1^*, \ldots, \gamma_p^*)$ that is missing some truly active variables. Suppose that, as in Condition A1, truly $\epsilon_i \sim s_0(\epsilon_i) = s(\epsilon_i \mid \xi_0)$ for some error density family $s(\epsilon_i \mid \xi)$, $\xi \in \Xi$, and fixed $\xi_0 \in \Xi$. Denote by $L_0(\theta_\gamma, \xi)$ the likelihood under the correct $\epsilon_i \sim s(\epsilon_i \mid \xi)$ and $p_0(y \mid \gamma) = \int L_0(y \mid \theta_\gamma, \xi) p(\theta_\gamma, \xi) d\theta_\gamma d\xi$ the associated integrated likelihood under some prior $p(\theta_\gamma, \xi) > 0$. The interest is in comparing the correct Bayes factor $p_0(y \mid \gamma^*)/p_0(y \mid \gamma)$ to the misspecified $\hat{p}(y \mid \gamma^*)/\hat{p}(y \mid \gamma)$. Under fairly general conditions

$$\log(p_0(y \mid \gamma^*)/p_0(y \mid \gamma)) \approx n\mathrm{D}_0(p_0(y \mid \theta^*_\gamma, \xi^*_\gamma, \gamma)),$$

plus lower order terms analogous to those in Proposition 5 when $\gamma \not\subset \gamma^*$, where $\mathrm{D}_0(p_0(y \mid \theta_\gamma^*, \xi_\gamma^*, \gamma))$ is the Kullback-Leibler divergence between the data-generating $p_0(y \mid \theta_{\gamma^*}^*, \xi_0, \gamma^*)$ and the KL-optimal $p_0(y \mid \theta_\gamma^*, \xi_\gamma^*, \gamma)$ under γ. Trivial algebra gives

$$\log\left(\frac{p_0(y \mid \gamma^*)/p_0(y \mid \gamma)}{\hat{p}(y \mid \gamma^*)/\hat{p}(y \mid \gamma)}\right) \approx n\left(\mathrm{D}_0(p_0(y \mid \theta_\gamma^*, \xi_\gamma^*, \gamma)) + \mathrm{D}_0(p(y \mid \eta_{\gamma^*}^*, \gamma^*)) - \mathrm{D}_0(p(y \mid \eta_\gamma^*, \gamma))\right). \quad (11)$$

The sign of the right hand side in (11) determines whether the misspecified Bayes factor has lower or greater asymptotic power than the correct Bayes factor. A precise study of (11) deserves separate treatment, but the expression can be loosely interpreted as a type of triangle inequality. If the divergence due to simultaneously using the wrong error distribution and γ instead of γ^*, $\mathrm{D}_0(p(y \mid \eta_\gamma^*, \gamma))$, is smaller than the sum of the divergences due to only using the wrong error distribution plus that of only using γ instead of γ^*. Then, misspecifiying the error distribution results in slower (but still exponential) Bayes factor rates to detect truly active variables. To our knowledge there is no guarantee that (11) is positive in general for any assumed model and data-generating truth, however in all our examples misspecified Bayes factors exhibited such a loss of power, suggesting that this is often the case.

5.3 Model exploration

Algorithm 3 describes a novel Gibbs sampling that can be used when p_γ is too large for exhaustive enumeration of all $2^{p_\gamma+2}$ models. Although conceptually simple, Algorithm 3 extends a method that delivered good results for high-dimensional variable selection under Normal errors (Johnson and Rossell 2012), and is designed to spend most iterations in the Normal model whenever it is a good enough approximation. That is, as illustrated in our examples the computational effort adapts automatically to the nature of the data, so that the cost associated to abandoning the Normal model is only incurred when this is required to improve inference. Our implementation also allows the user to fix $(\gamma_{p+1}, \gamma_{p+2})$, so that one can condition on Normal, asymmetric Normal, Laplace or asymmetric Laplace errors whenever this is desired.

The number of iterations T should ideally be large enough for the chain to converge, see for instance Johnson (2013) for a discussion of formal convergence diagnostics based on coupling methods. In practice, it usually suffices to monitor some posterior quantities of interest. For instance, in the setting of variable selection with NLPs Rossell and Telesca (2017) found useful to set T large enough so that sampling-based estimates of $p(\gamma_j = 1 \mid y)$ are close enough to estimates based on renormalizing posterior probabilities across the models visited so far.

Algorithm 3. Gibbs model space search.

1. *Let* $\gamma_{p+1}^{(0)} = \gamma_{p+2}^{(0)} = 0$ *and set* $\gamma_1^{(0)}, \ldots, \gamma_p^{(0)}$ *using the greedy forward-backward initialization algorithm in Johnson and Rossell (2012). Set* $t = 1$.

2. *For* $j = 1, \ldots, p$, *update* $\gamma_j^{(t)} = 1$ *with probability*

$$\frac{p(\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, 1, \gamma_{j+1}^{(t-1)}, \ldots, \gamma_p^{(t-1)} \mid y)}{\sum_{\gamma_j=0}^{1} p(\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j, \gamma_{j+1}^{(t-1)}, \ldots, \gamma_p^{(t-1)} \mid y)}.$$

3. *Update* $(\gamma_{p+1}^{(t)}, \gamma_{p+2}^{(t)}) = (l, m)$ *with probability*

$$\frac{p(\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}, l, m \mid y)}{\sum_{\gamma_{p+1}=0}^{1} \sum_{\gamma_{p+2}=0}^{1} p(\gamma_1^{(t)}, \ldots, \gamma_p^{(t)}, \gamma_{p+1}, \gamma_{p+2} \mid y)}.$$

If $t \leq T$, *set* $t = t + 1$ *and go back to Step 2, otherwise stop.*

6. RESULTS

We studied via simulations the practical implications of model misspecification on variable selection, both on small and large p (Sections 6.1-6.3), as well as the ability of our framework to detect asymmetries ($\gamma_{p+1} = 1$) and heavier-than-normal tails ($\gamma_{p+2} = 1$). The heteroscedastic errors simulation in Section 6.2 and the DLD example in Section 6.5 also illustrates how to perform quantile regression for multiple fixed quantile levels as a particular case of our framework.

Computations were carried out using function modelSelection in R package mombf 1.9.2 (Rossell et al. 2016), using default prior settings (Section 3) and Laplace approximations to $p(y \mid \gamma)$ unless otherwise stated. Although our goal is to build a Bayesian framework to cope with simple departures from normality, for comparison we included some penalized likelihood methods with available R implementation: standard LASSO penalties on least squares regression (LASSO-LS, Tibshirani (1996)), LASSO penalties on least absolute deviation (LASSO-LAD, Wang and Li (2009)), SCAD penalties on least squares (Fan and Li 2001), and LASSO penalties on quantile regression (LASSO-QR, Wu and Liu (2009)). For LASSO-LS, LASSO-LAD, LASSO-QR and SCAD we set the penalization parameter with 10-fold cross-validation using functions mylars, rq.lasso.fit and ncvreg in R packages parcor 0.2.6, rqPen 1.5.1 and ncvreg 3.4.0 (respectively) with default parameters. LASSO-LAD corresponds to setting the 0.5 quantile in rq.lasso.fit, whereas for LASSO-QR we set the optimal quantile $(1+\alpha)/2$ where α is the data-generating truth. That is, we performed a conservative comparison where results for LASSO-QR may be slightly optimistic. All R code is provided in supplementary files.

6.1 Low-dimensional simulation

[Figure 2 about here.]

We started by simulating 200 data sets from a linear model with Normal residuals, each with $n = 100$, $p = 6$, $\theta = (0, 0.5, 1, 1.5, 0, 0)$ ($\theta_1 = 0$ corresponds to the intercept), $\vartheta = 2$. Covariate values were generated from a multivariate Normal centered at 0, with unit variances and all pairwise correlations $\rho_{ij} = 0.5$. We compared the results under assumed Normal, asymmetric Normal, Laplace and asymmetric Laplace errors, and also when inferring the residual distribution with our framework (Section 5). Throughout, we used MOM priors with default $g_\theta = 0.348$, $g_\alpha = 0.357$ and uniform model probabilities $p(\gamma) \propto 1$. Given that p is small we enumerated and computed $p(\gamma \mid y)$ for all models. Figure 2 (top left) shows the marginal posterior probabilities $p(\gamma_j = 1 \mid y)$. These were almost identical under assumed Normal and asymmetric Normal errors. Both models were preferrable to Laplace or asymmetric Laplace errors, mainly in giving higher $p(\gamma_j = 1 \mid y)$ for truly active variables.

We repeated the simulation study, this time generating $\epsilon_i \sim \text{AN}(0, 2, -0.5)$, $\epsilon_i \sim L(0, 2)$ and finally $\epsilon_i \sim \text{AL}(0, 2, -0.5)$. Here we observed more marked differences than under $\epsilon_i \sim N(0, 2)$, specifically failing to account for thick tails caused a substantial drop in $p(\gamma_j = 1 \mid y)$ for truly active predictors. As an example, when truly $\epsilon_i \sim \text{AL}(0, 4, -0.5)$ the mean $p(\theta_3 \neq 0 \mid y)$ increased from 0.63 under assumed Normal errors to 0.89 under asymmetric Laplace errors. These results suggest that wrongly assuming Normal errors may has more pronounced consequences on inference than using more robust error distributions. Interestingly, including asymmetry in the model had no noticeable adverse effects on inference even when residuals were truly symmetric, and improved power when residuals were truly asymmetric. Hence the reasoning for adopting symmetric models seems mostly computational.

Our framework based on inferring $(\gamma_{p+1}, \gamma_{p+2})$ showed a highly competitive behaviour, usually fairly close to assuming the true distribution (Figure 2). The mean posterior probability assigned to the true error distribution was always > 0.8 (Supplementary Table 3S), indicating that the desired departures from normality were effectively detected.

We repeated all the analyses above first using Monte Carlo estimates of $p(y \mid \gamma)$ based on $B = 10,000$ importance samples, and then again using our alternative default $g_\alpha = 0.087$. Supplementary Table 3S shows that inference on the error distribution remained remarkably stable, albeit as expected reducing $g_\alpha = 0.357$ to 0.087 increases slightly $p(\alpha \neq 0 \mid y)$ in all settings. Supplementary Figures 2S-3S show $p(\gamma_j = 1 \mid y)$. These are virtually indistinguishable from those in Figure 2, indicating that the results are robust to these implementation details.

Finally, we assessed the behaviour of the least-squares initialization in Algorithms 1-2 under different data-generating mechanisms, specifically in terms of CPU times. Table 1S gives mean times across $10,000$ independent simulations with $p = 6$ and increasing data-generating truths $\alpha^* = 0, -0.25, -0.5, -0.75$, both for two-piece Normal and two-piece Laplace errors. These are for the whole model-fitting process, including exhaustive model enumeration and computation of posterior model probabilities. The time increases were of roughly 25% from $\alpha^* = 0$ to $\alpha^* = -0.75$. This is as expected, under asymmetry least-squares is a poorer initial $\hat{\boldsymbol{\theta}}^{(0)}$. The increase is however mild, indicating that a larger fraction of the computation cost arises from other operations (e.g. matrix inversion after the mode has been found). These results support that our $\hat{\boldsymbol{\theta}}^{(0)}$ is not particularly problematic. One could certainly consider alternative $\hat{\boldsymbol{\theta}}^{(0)}$, say median regression or trimmed least squares, but these are typically costlier that least-squares hence the overall gains are likely to be moderate at best.

6.2 Non-identically distributed errors

[Figure 3 about here.]

We investigate the effect of deviations from the identically distributed errors assumption. We repeated the simulations in Section 6.1 under heteroscedastic and hetero-asymmetric errors, and reproduced a pathological example reported by Grünwald and van Ommen (2014). Under heteroscedasticity, we set $\tilde{\epsilon}_i = e^{x_i^T \theta} \epsilon_i / c$ where c was set such that $\text{Var}(\tilde{\epsilon}_i) = \text{Var}(\epsilon_i)$, so that the signal-to-noise was comparable to our earlier simulations. This example mimics that used by Koenker (2005) (Figure 1.6) to illustrate the potential interest of conditioning upon multiple quantile levels, except that ours has a stronger (exponential) association between mean and variance. We first apply our framework without conditioning on α. Figure 3 shows $P(\gamma_j = 1 \mid y)$ for $p = 6$. The main feature is that the Laplace and asymmetric Laplace models clearly outperform the Normal model both in sensitivity and specificity. For instance, when truly $\theta_2^* = 0.5$ the mean $P(\gamma_2 = 1 \mid y)$ increased from 0.33 to 0.78 under assumed Normal and Laplace residuals respectively. The mean for truly inactive $\theta_5^* = \theta_6^* = 0$ decreased from 0.063 to 0.021. Interestingly, inferring the error model chose Laplace errors even when these were truly Normal and showed a highly competitive performance (Supplementary Table 7S). Intuitively, heteroscedasticity gives an overabundance of residuals at the origin and at the tails relative to a homoscedastic Normal. Such errors are better captured by a Laplace model.

Next, following Koenker (2005) we assessed the performance of quantile regression at fixed quantile levels $q = 0.05, 0.25, 0.75, 0.95$. The usual motivation for conditioning upon multiple quantiles is to consider that each quantile could potentially depend on a different subset of predictors. This corresponds to conditioning upon asymmetric Laplace errors and fixed $\alpha = 2q - 1$ (Section 2). The marginal posterior inclusion probabilities in Table 8S show that $q = 0.5$ (the KL-optimal value) led to substantially higher sensitivity than say $q = 0.05$ or $q = 0.95$. We remark that under our heteroscedastic data-generating truth the q^{th} conditional quantile is $\mathbf{x}_i^T \boldsymbol{\theta} + z_q \sqrt{e^{\mathbf{x}_i^T \boldsymbol{\theta}} / c}$ where z_q is the q^{th} standard Normal quantile. The results illustrate that, in this and similar situations where all quantiles depend on the same subset of variables, inferring α can lead to better variable selection than conditioning upon poor choices of α. Naturally, under more complex scenarios where quantiles do depend on different variable subsets, conditioning upon multiple α can provide a richer description of the dependence of y_i on $\mathbf{x}_i$.

Our second simulation scenario considered the presence of non-constant asymmetry. Specifically, we generated $\tanh(\alpha_i) \sim N(\text{atanh}(\bar{\alpha}), 1/4^2)$ where the median asymmetry is $\bar{\alpha} = 0, -0.5$ as before.

Under this setting when $\bar{\alpha} = 0$ then $\alpha_i \in (-0.45, 0.45)$ with 0.95 probability and when $\bar{\alpha} = -0.5$ it is $(-0.78, -0.06)$, i.e. there is substantial variation in asymmetry. Supplementary Figure 8S displays $P(\gamma_j = 1 \mid y)$ for $p = 6$. These results are qualitatively similar to those in Figure 2 where α_i was held fixed. We remark that although in these examples non-constant asymmetry was not a concern, its impact could be more serious in other settings, *e.g.* under strong dependencies between the asymmetry and the mean. See Section 7 for some further discussion.

Finally, we mimic the example in Grünwald and van Ommen (2014), Section 5.1.2. The authors set $(y_i, x_{i1}, \ldots, x_{ip}) = (0, 0, \ldots, 0)$ with probability 0.5 and $y_i = x_i^T \theta^* + \epsilon_i$ with probability 0.5, where $x_{ij} \sim N(0, 1)$, $\theta^* = (0.1, 0.1, 0.1, 0.1, 0.1, 0, \ldots, 0)$ and $\epsilon_i \sim N(0, \vartheta)$. This extreme case of non-id errors is interesting in that the degeneracy at the origin results in inliers, rather than the more commonly considered outliers in y_i or leverage points in x_i. We selected variables under assumed Normal errors for $p = n = 50$, for this (n, p) the authors reported a particularly large inflation of false positives (as $n \to \infty$ these disappeared). Specifically we set $\vartheta^* = 2$, Zellner's $p(\theta_\gamma \mid \gamma) = N(\theta_\gamma; 0, n(X_\gamma^T X_\gamma)^{-1})$ and the Beta-Binomial(1,1) prior for $p(\gamma)$. The posterior mode selected a striking 21.3 out of the 45 spurious variables (mean across 100 independent simulations), confirming their findings (Supplementary Table 9S). Under a pMOM prior the mean false positives decreased to 12.1 when conditioning on Normal errors and further to 10.5 when inferring the error model. Interestingly under the peMOM prior and Normal errors the mean false positives were only 2.9. All methods showed similar sensitivity, selecting roughly 3 out of the 5 active variables. This example illustrates that, while serious model misspecification can have marked effects for finite n, these can be partially mitigated by adopting priors that penalize small coefficients and flexible error models. In this particular example the exponential peMOM penalties were more effective than the pMOM penalties in lowering false positives.

6.3 High dimensional simulation

[Figure 4 about here.]

We repeated the simulation study in Section 6.1 with $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$ by adding 95 spurious predictors for a total of $p = 100$ covariates, and subsequently 400 more spurious predictors for a total $p = 500$. Given that the model space is too large for a full enumeration, we run the

Gibbs algorithm in Section 5.3 with $T = 10,000$ iterations. To initialize the chain we used the greedy Gibbs algorithm from Johnson and Rossell (2012), which starts at $\gamma = (0, \dots, 0)$ and keeps adding or removing individual covariates until a local mode is found. We set $p(\gamma)$ to the default Beta-Binomial(1,1) and left all other settings as in Section 6.1.

We conducted one first set of simulations under $\vartheta = 1$. Figure 4 shows the proportion of simulations in which the posterior mode $\widehat{\gamma} = \arg\max_{\gamma} p(\gamma \mid y)$ was equal to the simulation truth $\gamma_0 = (0, 1, 1, 1, 0, \dots, 0)$. The main finding was that assuming the wrong error distribution had a marked detrimental impact on Bayesian variable selection, particularly in the presence of asymmetries or thicker-than-normal tails. Supplementary Table 5S gives the exact figures, as well as the number of false and true positives. All Bayesian formulations compared favourably to LASSO-LS, LASSO-LAD, LASSO-QR and LASSO-SCAD, mainly due to the latter incurring an inflated number of false positives. This is in agreement with earlier findings (Johnson and Rossell 2012; Rossell and Telesca 2017) when comparing NLPs to penalized likelihoods, and likely partially related to the fact that cross-validation focuses on predictive ability and thus tends to favour the inclusion of a few spurious covariates. Interestingly, in our study LASSO-LAD showed little advantages over LASSO-LS, even under truly Laplace errors. LASSO-QR did improve slightly upon LASSO-SCAD when truly $\alpha^* \neq 0$ both in sensitivity and specificity. Analogously to the $p = 6$ case in Figure 2, when $p = 101, 501$ the marginal inclusion probabilities for truly active variables suffered a drop when ignoring the presence of asymmetry or heavy tails (Supplementary Figures 4S-5S). Our framework to infer the error distribution delivered highly competitive inference.

Supplementary Table 2S indicates CPU times for $p = 100$. The Normal model exhibited lower times under truly Normal or Laplace errors, likely due to the availability of closed-form expressions for $p(\gamma \mid y)$. The presence of asymmetry encouraged the inclusion of an intercept term under the Normal model, the associated increase in model dimension cancelled the computational savings. Times for our inferred residuals framework were highly competitive under all scenarios.

To emulate a situation with lower signal-to-noise ratio we repeated the simulation study under $\vartheta = 2$. The results are shown in Supplementary Table 6S and Supplementary Figures 6S-7S. Briefly, the performance of all methods suffered in this more challenging setting due to a drop in the power to detect truly active predictors, however their relative performances were largely analogous to those for $\vartheta = 1$.

6.4 TGFB data

[Figure 5 about here.]

We illustrate our methodology with the human microarray gene expression data in colon cancer patients from Calon et al. (2012). Briefly, following upon Rossell and Telesca (2017), we aim to detect which amongst p =10,172 candidate genes have an effect on the expression levels of TGFB, a gene known to play an important role in colon cancer progression. These data contain moderately correlated covariates with absolute Pearson correlations ranging in (0,0.956) and 99% of them being in the interval (0,0.375). Both response and predictors were standardized to zero mean and unit variance. The dataset and further information are provided in Rossell and Telesca (2017).

[Table 1 about here.]

We start by considering inference under the Normal model, i.e. conditional on $\gamma_{p+1} = \gamma_{p+2} = 0$. We ran 1,000 Gibbs iterations (*i.e.* $10^3 \times 10,172$ model updates), which was deemed sufficient for practical convergence (see supplementary material in Rossell and Telesca (2017)). Table 1 shows the highest posterior probability models. The top model included the 6 genes ARL4C, AOC3, URB2, FAM89B, PCGF2, CCDC102B and had an estimated $p(\gamma \mid y) = 0.299$. Alternatively, selecting variables with marginal $p(\gamma_j = 1 \mid y) > 0.5$ (Barbieri and Berger 2004) returned 5 out of these 6 genes ($p(\gamma_j = 1 \mid y) = 0.482$ for FAM89B). Briefly, according to genecards.org FAM89B is a TGFB regulator, ARL4C and PCGF2 have been related to various cancer types and AOC3 is used to alleviate cancer symptoms, reinforcing the plausibility that these genes may be indeed related to TGFB. URB2 and CCDC102B have no known relation to cancer, although the latter is connected to ARL4D in the STRING interaction networks.

We next considered the possibility that the Normal model might not be adequate for these data. As an exploratory check, a quantile-quantile plot based on the residuals under the top model revealed no strong departure from normality (Figure 5). Although this is somewhat reassuring one cannot discard a lack of normality under a different set of predictors, as the top model was selected under assumed normality. To conduct a more formal analysis we run Algorithm 3 ($T = 1,000$

iterations) now including $\gamma_{p+1}, \gamma_{p+2}$. The posterior probabilities for Normal, asymmetric Normal, Laplace and asymmetric Laplace errors were 0.998, 0.0002, 0.0018 and 1.3×10^{-27}, respectively. The six top models and their posterior probabilities closely matched those under the assumed Normal model (Table 1), and the correlation between marginal inclusion probabilities under Normal and inferred residuals was 0.96. These results support that our framework to infer $(\gamma_{p+1}, \gamma_{p+2})$ in Algorithm 3 is able to detect when errors are approximately Normal.

6.5 DLD data

We consider another genomics study by Yuan et al. (2016). In contrast to Section 6.4, here RNA-sequencing was used to measure gene expression, a newer and more precise technology than microarrays. The study included 100 colorectal, 36 prostate, and 6 pancreatic cancer and 50 healthy control patients, for a total of $n = 192$ patients. Briefly, the authors used a measure of expression called RPM. RPM considers the number of reads mapped to a given gene relative to the gene length and may exhibit heavy tails or asymmetries, even after log or other transformations. We focus on the 58 messenger RNA genes identified in the exRNA species diversity analysis provided by the authors in Supplementary Table S1. To illustrate our methodology, we consider predicting the expression of gene DLD based on the remaining 57 genes and the 3 binary variables indicating the patient type (colorectal, prostate, pancreatic). According to genecards.org, the protein encoded by DLD can perform mechanistically distinct functions, it can regulate the energy metabolism and has been found to be associated with dehydrogenase and leukocyte adhesion defficiencies.

We first applied our methodology conditioning on Normal errors ($\gamma_{p+1} = \gamma_{p+2} = 0$). We used 10,000 Gibbs iterations. The highest posterior probability model had $p(\gamma \mid y) = 0.58$ and contained 5 genes (C6orf226, ECH1, CSF2RA, FBXL19, RRP1B), however, its residuals showed a clear departure from normality (Figure 5, right). We run again our Gibbs algorithm, this time inferring γ_{p+1} and γ_{p+2}. The analysis returned an overwhelming $p(\gamma_{p+1} = 1, \gamma_{p+2} = 0 \mid y) = 0.999$ in favour of Laplace residuals. The top model had posterior probability 0.36 and contained the same 5 predictors plus an extra gene MTMR1. MTMR1 encodes a protein related to the myotubularin family containing consensus sequences for protein tyrase phosphatases, whereas the response gene DLD has a post-translational modification based on tyrosine phosphorylation, thus giving a plausible biological mechanism connecting MTMR1 and DLD. Supplementary Table 10S lists the six largest marginal variable inclusion probabilities under Normal and inferred error distribution.

So far, we treated α as a parameter to be learnt from the data. We now condition upon asymmetric Laplace errors and fixed $\alpha = -0.5, 0, 0.5$. This leads to quantile regression for the $(1+\alpha)/2 = 0.25, 0.5, 0.75$ percentiles (Section 2). Supplementary Table 11S displays the top 5 models for each α. Briefly, five genes (C6orf226, CSF2RA, ECH1, RRP1B and FBXL19) featured in the top model for all α's, the first four with marginal inclusion probability > 0.99. FBXL19 had higher probability under $\alpha = 0$ than $\alpha = -0.5, 0.5$ (0.783 vs. 0.516 and 0.467 respectively). MTMR1 featured in the top model only for $\alpha = 0$ (marginal probability 0.619). Given the biological plausibility that MTMR1 is related to DLD, these results suggest that setting $\alpha = 0$ (the value inferred from the data) may have led to higher power to detect MTMR1 than conditioning on Normal or asymmetric Laplace residuals with $\alpha = -0.5$ or $\alpha = 0.5$. This is in agreement with Proposition 5 and our simulations in Sections 6.1-6.3.

7. DISCUSSION

Most efforts in Bayesian variable selection focus either on the Normal model or on flexible alternatives that require MCMC. Our framework represents a middle-ground to add flexibility in a parsimonious manner that remains analytically and computationally tractable, facilitating applications where either p is large or n is too moderate to fit more complex models accurately. Our results show that model misspecification is a non-ignorable issue with important consequences for model selection. Bayes factor rates typically retain the same functional dependence on n (*e.g.* polynomial or exponential) as when the model is correctly specified, however the coefficients governing these rates do change. Specifically, the ratio of the correct *vs.* misspecified Bayes factors to detect truly active variables grows exponentially with n when a triangle-type inequality holds, signaling the potential for an important drop in sensitivity. Our empirical studies support this finding: failing to account for simple forms of asymmetry or heavy tails reduced the proportion of correct model selections by several folds. Misspecification also has an effect on false positives. Although here the ratio of correct vs. misspecified Bayes factors is essentially a constant, the effect can be noticeable for finite n. Hence it is important to consider flexible likelihoods and, when possible, also adopt false positive correction mechanisms for finite n. As a possible venue for the latter, we illustrated in an example how non-local priors helped discard small spurious parameters arising from misspecification. A more detailed study would be interesting future work.

Other future avenues include extensions to allow for polynomial error tails, dependent errors, heteroscedasticity or covariate-dependent asymmetry. We remark that fully non-parametric strategies already exist, *e.g.* Chung and Dunson (2009). The challenge is to build models that provide an intermediate level of flexibility while giving tractable variable selection. For instance, allowing the variance or asymmetry to depend on x_i is an interesting task for which there is no unique agreed-upon solution. One possibility is to let $\vartheta_i = \exp\left(x_i^T \beta\right)$, where $|\beta| \leq |\theta|$, akin to what Daye et al. (2012) for Normal errors. The authors found that the log-likelihood for β for fixed θ is log-concave, and so is that for θ under fixed β, enabling fast optimization. It would be interesting to develop similar strategies for the asymmetry and non-Normal errors. An issue here would be dealing with the increased problem dimensionality due to selecting variables also for β. Another interesting venue stemming from our work is posing non-parametric models that can collapse onto simple parametric forms when the extra flexibility is not needed. Again the idea is to strike a balance between the tractability offered by simple models and the ultimate goal of providing accurate inference. Other extensions are developing more advanced optimization or model search strategies, our goal here was to illustrate that even relatively simple methods can be competitive. Such computational issues are particularly meaningful in increasingly challenging settings, *e.g.* large graphical or spatio-temporal models. Overall, we hope to have provided a basic framework that others can build on to tackle these exciting applications.

ACKNOWLEDGMENTS

DR was partially funded by the NIH grant R01 CA158113-01 and the Ramón y Cajal Fellowship RYC-2015-18544.

REFERENCES

Alfons, A., Croux, C., and Gelper, S. (2013), “Sparse least trimmed squares regression for analyzing high-dimensional large data sets,” *The Annals of Applied Statistics*, 7(1), 226–248.

Arellano-Valle, R., Gómez, H., and Quintana, F. (2005), "Statistical inference for a general class of asymmetric distributions," *Journal of Statistical Planning and Inference*, 128(2), 427–443.

Arslan, O. (2012), "Weighted LAD-LASSO method for robust parameter estimation and variable selection in regression," *Computational Statistics and Data Analysis*, 56(6), 1952–1965.

Barbieri, M., and Berger, J. (2004), "Optimal predictive model selection," *The Annals of Statistics*, 32(3), 870–897.

Bayarri, M., Berger, J., Forte, A., and Garcia-Donato, G. (2012), "Criteria for Bayesian Model Choice with Application to Variable Selection," *The Annals of statistics*, 40(3), 1550–1577.

Breheny, P., and Huang, J. (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *Annals of Applied Statistics*, 5(1), 232–253.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D., Iglesias, M., Céspedes, M., Sevillano, M., Nadal, C., Jung, P., Zhang, X.-F., Byrom, D., Riera, A., Rossell, D., Mangues, R., Massague, J., Sancho, E., and Batlle, E. (2012), "Dependency of colorectal cancer on a TGF-beta-driven programme in stromal cells for metastasis initiation," *Cancer Cell*, 22(5), 571–584.

Chae, M., Lin, L., and Dunson, D. (2016), "Bayesian sparse linear regression with unknown symmetric error," *arXiv*, 1608.02143, 1–34.

Chung, Y., and Dunson, D. (2009), "Nonparametric Bayes conditional distribution modeling with variable selection," *Journal of the American Statistical Association*, 104(488), 1646–1660.

Daye, Z., Chen, J., and Li, H. (2012), "High-Dimensional Heteroscedastic Regression with an Application to eQTL Data Analysis," *Biometrics*, 68(1), 316–326.

Dette, H., Ley, C., and Rubio, F. (2016), "Natural (non-) informative priors for skew-symmetric distributions," *arXiv preprint arXiv:1605.02880*, .

Eicker, F. (1964), "Asymptotic normality and consistency of the least squares estimator for families of linear regressions," *Annals of Mathematical Statistics*, 34(2), 447–456.

Fan, J., Fan, Y., and Barut, E. (2014), "Adaptive robust variable selection," *The Annals of Applied Statistics*, 42(1), 324–351.

Fan, J., and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96(456), 1348–1360.

Fernández, C., and Steel, M. (1998), "On Bayesian modeling of fat tails and skewness," *Journal of the American Statistical Association*, 93(441), 359–371.

Gijbels, I., and Vrinssen, I. (2015), "Robust nonnegative garrote variable selection in linear regression," *Computational Statistics and Data Analysis*, 85, 1–22.

Gottardo, R., and Raftery, A. (2007), "Bayesian robust transformation and variable selection: a unified approach," *The Canadian Journal of Statistics*, 37(3), 361–380.

Grünwald, P., and van Ommen, T. (2014), "Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it," *arXiv*, 1412.3730, 1–70.

Huber, P. (1973), "Robust regression: asymptotics, conjectures and Monte Carlo," *The Annals of Statistics*, 1(5), 799–821.

Johnson, V. (2013), "On numerical aspects of Bayesian model selection in high and ultrahigh-dimensional settings," *Bayesian Analysis*, 8(4), 741.

Johnson, V., and Rossell, D. (2010), "Prior Densities for Default Bayesian Hypothesis Tests," *Journal of the Royal Statistical Society B*, 72(2), 143–170.

Johnson, V., and Rossell, D. (2012), "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, 24(498), 649–660.

Kimber, A. C. (1985), "Methods for the two-piece normal distribution," *Communications in Statistics - Theory and Methods*, 14(1), 235–245.

Knight, K. (1999), "Asymptotics for L1-estimators of regression parameters under heteroscedasticity," *The Canadian Journal of Statistics*, 27(3), 497–507.

Kocherginsky, M., He, X., and Mu, Y. (2005), "Practical confidence intervals for regression quantiles," *Journal of Computational and Graphical Statistics*, 14(1), 41–55.

Koenker, R. (1994), Confidence Intervals for Regression Quantiles,, in *Proceedings of the 5th Prague Symposium on Asymptotic Statistics*, Springer-Verlag, pp. 349–359.

Koenker, R. (2005), *Quantile regression*, Cambridge: Cambridge University Press.

Koenker, R., and Bassett, G. (1982), "Tests of linear hypotheses and L1 estimation," *Econometrica*, 50(6), 1577–1584.

Kundu, S., and Dunson, D. (2014), "Bayes variable selection in semiparametric linear models," *Journal of the American Statistical Association*, 109(505), 437–447.

Lambert-Lacroix, S. (2011), "Robust regression through the Huber's criterion and adaptive lasso penalty," *Electronic Journal of Statistics*, 5, 1015–1053.

Levenberg, K. (1944), "A Method for the Solution of Certain Non-Linear Problems in Least Squares," *Quarterly of Applied Mathematics*, 2(2), 164–168.

Loh, P.-L. (2017), "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators," *The Annals of Statistics*, 45(2), 866–896.

Mallick, H., and Nengjun, Y. (2013), "Bayesian methods for high dimensional linear models," *Journal of Biometrics & Biostatistics*, 1, 005.

Marquardt, D. (1963), "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM Journal on Applied Mathematics*, 11(2), 431–441.

Mendelson, S. (2014), "Learning without concentration for general loss functions," *ArXiv*, 1410.3192.

Mudholkar, G., and Hutson, A. (2000), "The epsilon-skew-normal distribution for analyzing near-normal data," *Journal of Statistical Planning and Inference*, 83(2), 291–309.

Pollard, D. (1991), "Asymptotics for least absolute deviation regression estimators," *Econometric Theory*, 7(2), 186–199.

Rossell, D., Cook, J., Telesca, D., and Roebuck, P. (2016), *mombf: Moment and Inverse Moment Bayes Factors*. R package version 1.8.1.
URL: *https://CRAN.R-project.org/package=mombf*

Rossell, D., and Telesca, D. (2017), "Non-local priors for high-dimensional estimation," *Journal of the American Statistical Association*, 112, 254–265.

Rossell, D., Telesca, D., and Johnson, V. (2013), High-dimensional Bayesian classifiers using non-local priors,, in *Statistical Models for Data Analysis XV*, Springer, pp. 305–314.

Rubio, F., and Genton, M. (2016), "Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis," *Statistics in Medicine*, 35(4), 2441–2454.

Rubio, F., and Steel, M. (2014), "Inference in Two-Piece Location-Scale Models with Jeffreys Priors (with discussion)," *Bayesian Analysis*, 9(1), 1–22.

Rubio, F., and Yu, K. (2017), "Flexible objective Bayesian linear regression with applications in survival analysis," *Journal of Applied Statistics*, 44.

Scott, J., and Berger, J. (2010), "Bayes and empirical Bayes multiplicity adjustment in the variable selection problem," *The Annals of Statistics*, 38(5), 2587–2619.

Shin, M., Bhattacharya, A., and Johnson, V. (2015), "Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings," *Texas A&M University (technical report)*, pp. 1–33.

Sorensen, D. (1982), "Newton's Method with a Model Trust Region Modification," *SIAM Journal of Numerical Analysys*, 19(2), 409–426.

Srivastava, M. (1971), "On fixed width confidence bounds for regression parameters," *The Annals of Mathematical Statistics*, 42(4), 1403–1411.

Tibshirani, R. (1996), "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58(1), 267–288.

Wallis, K. (2014), "The two-piece normal, binormal, or double Gaussian distribution: its origin and rediscoveries," *Statistical Science*, 29(1), 106–112.

Wang, H., Li, G., and Jiang, G. (2007), "Robust Regression Shrinkage and Consistent Variable Selection through the LAD-LASSO," *Journal of Business & Economic Statistics*, 25(3), 347–355.

Wang, L., and Li, R. (2009), "Weighted Wilcoxon-Type Smoothly Clipped Absolute Deviation Method," *Biometrics*, 65(2), 564–571.

Wang, L., Tang, Y., Sinha, D., Pati, D., and Lipsitz, S. (2016), "Bayesian Variable Selection for Skewed Heteroscedastic Response," *arXiv preprint arXiv:1602.09100*, .

Wu, L., and Liu, Y. (2009), "Variable selection in quantile regression," *Statistica Sinica*, 19(2), 801–809.

Yan, Y., and Kottas, A. (2015), A new family of error distributions for Bayesian quantile regression,, Technical report, University of California Santa Cruz.

Yu, K., Chen, C., Reed, C., and Dunson, D. (2013), "Bayesian variable selection in quantile regression," *Statistics and its Interface*, 6(2), 261–274.

Yuan, T., Huang, X., Woodcock, M., Du, M., Dittmar, R., Wang, Y., Tsai, S., Kohli, M., Boardman, L., Patel, T., and Wang, L. (2016), "Plasma extracellular RNA profiles in healthy and cancer patients," *Scientific Reports*, 6, 1–11.

List of Figures

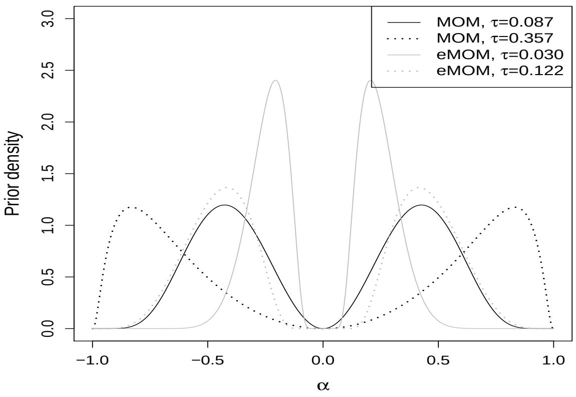

Figure 1: Default priors for α.

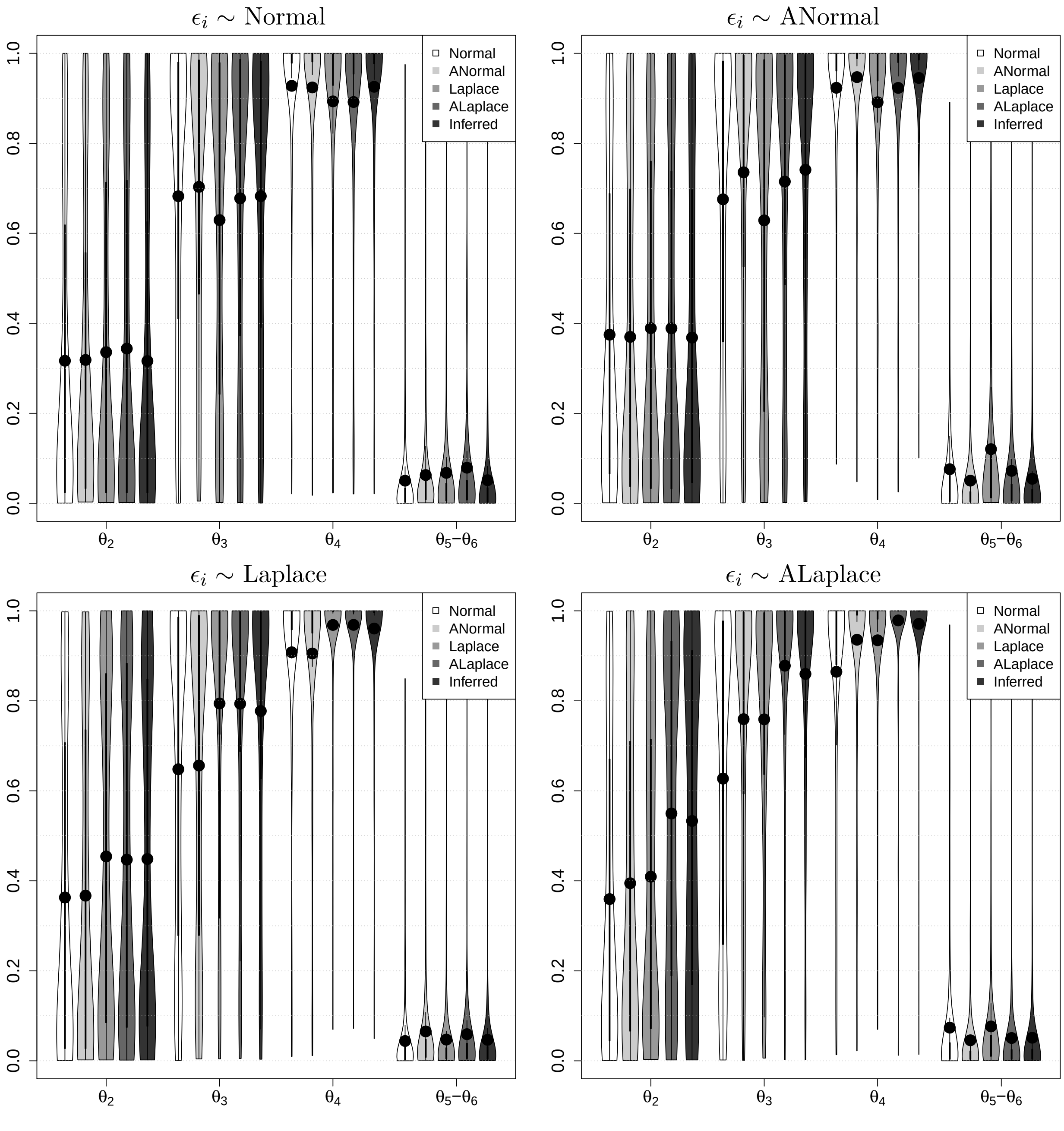

Figure 2: $P(\theta_i \neq 0 \mid y)$ for simulation with constant $\vartheta = 2$, $\alpha = 0, 0.5$. $P(\theta_i \neq 0 \mid y)$ for $p = 6$, $\theta = (0, 0.5, 1, 1.5, 0, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

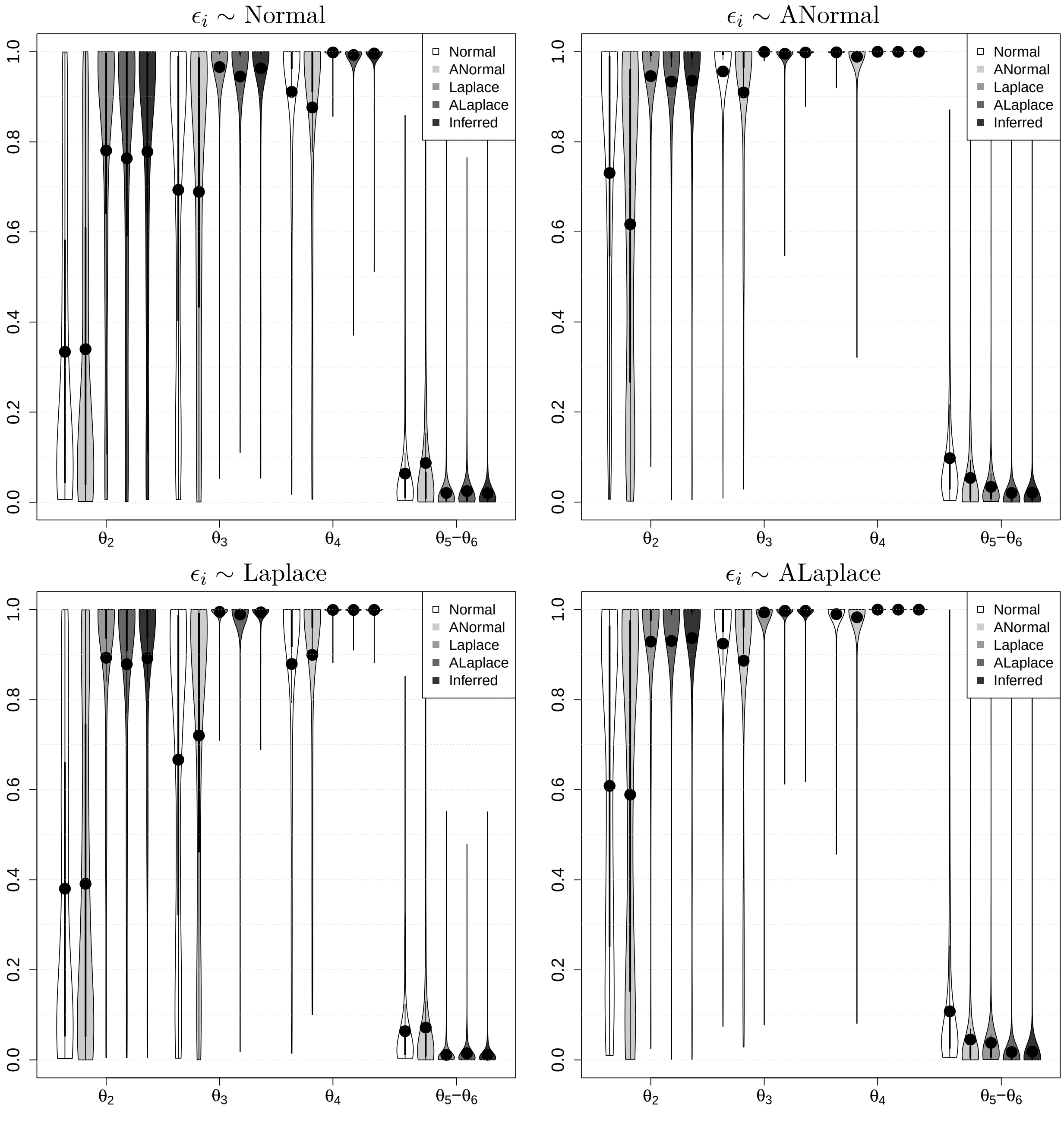

Figure 3: $P(\theta_i \neq 0 \mid y)$ for simulation with $\vartheta_i \propto e^{x_i^T\theta}$, constant $\alpha = 0, -0.5$. $p = 6$, $\theta = (0, 0.5, 1, 1.5, 0, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

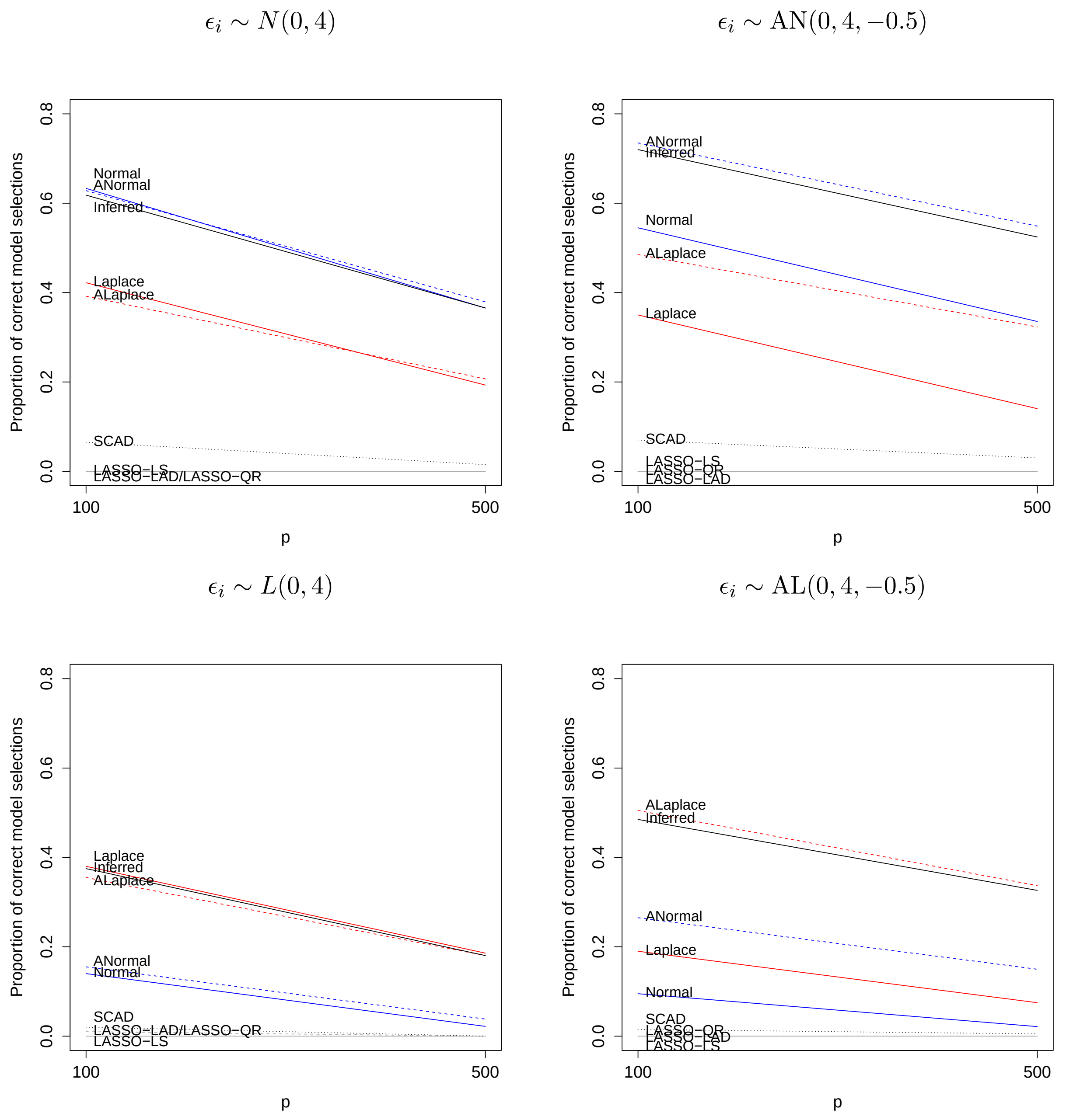

Figure 4: Proportion of correct model selections $p(\widehat{\gamma} = \gamma_0 \mid y)$. $\vartheta = 1$, $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$.

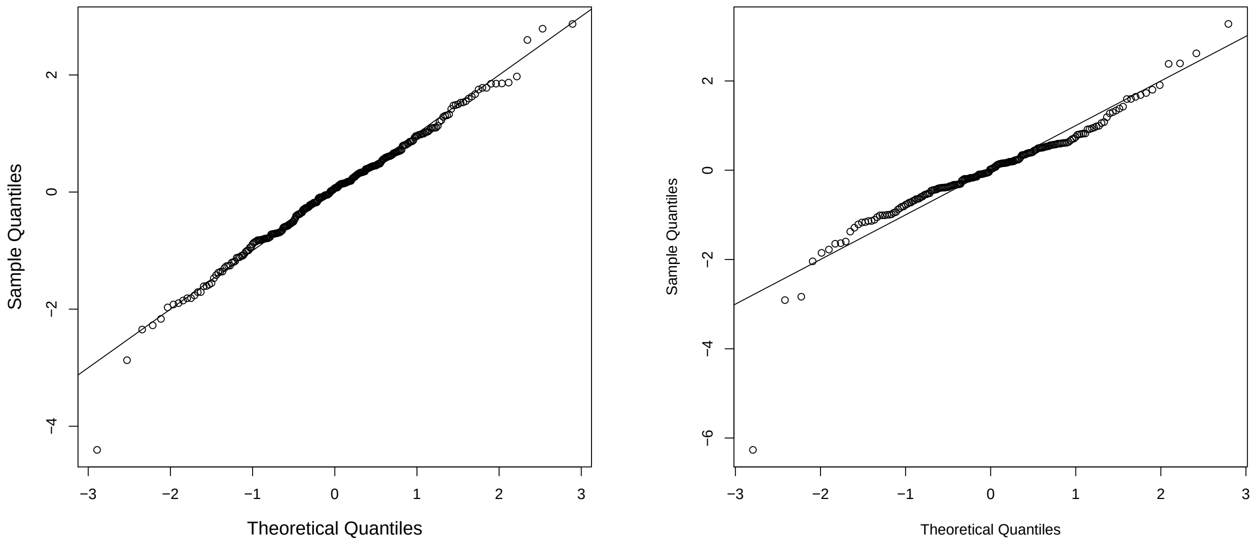

Figure 5: QQ Normal plot for TGFB (left) DLD (right) data.

List of Tables

Gene symbol	$p(\gamma \mid y)$	
	Normal	Inferred
ARL4C,AOC3,URB2,FAM89B,PCGF2,CCDC102B	0.299	0.304
ARL4C,CNRIP1,AOC3,PCGF2	0.165	0.167
ARL4C,CNRIP1,PCGF2	0.161	0.163
ARL4C,CNRIP1,AOC3,PCGF2,RPS6KB2	0.045	0.046
ARL4C,AOC3,PCGF2,CCDC102B	0.028	0.028
ARL4C,AOC3,FAM89B,PCGF2,CCDC102B	0.025	0.025

Table 1: TGFB data. Highest probability models under Normal and inferred error distribution.