Supplementary material. Tractable Bayesian variable selection: beyond normality

1. IMOM PRIOR

The product iMOM prior density on θ_γ (Johnson and Rossell 2012) is given by

$$p_I(\theta_\gamma \mid \vartheta, \gamma) = \prod_{\gamma_j=1} \frac{(g_\theta \vartheta)^{\frac{1}{2}}}{\sqrt{\pi}\theta_j^2} \exp\left\{-\frac{g_\theta \vartheta}{\theta_j^2}\right\}, \tag{1}$$

where by default $g_\theta = 0.133$ assigns $p(|\theta/\vartheta^{1/2}| > 0.2) = 0.99$. Regarding the asymmetry parameter $\tilde{\alpha} = \text{atanh}(\alpha)$, the prior is $p_I(\tilde{\alpha} \mid \gamma_{p+1} = 1) = \tilde{\alpha}^{-2}\sqrt{g_\alpha/\pi}e^{-g_\alpha/\tilde{\alpha}^2}$, and the default prior dispersions are $g_\alpha = 0.033$ to obtain $P(|\tilde{\alpha}| \geq 0.1) = 0.99$ and $g_\alpha = 0.136$ for $P(|\alpha| \geq 0.2) = 0.99$.

2. PROOFS

For simplicity, we drop γ from the notation in the proof of Propositions 1-4 and Corollay 1, given that all arguments are conditional on a given model γ.

2.1 Proof of Proposition 1

We start by stating a useful lemma stating that positive definite hessian plus continuous gradient guarantees concavity.

Lemma 1. *Let $f(\theta)$ be a function with continuous gradient $g(\theta)$, for all θ, and negative definite hessian $H(\theta)$ almost everywhere with respect to the Lebesgue measure. Then, $f(\theta)$ is strictly concave. If $H(\theta)$ is negative semidefinite, then $f(\theta)$ is concave.*

Proof. Let θ_1 and θ_2 be two arbitrary values and denote $\theta_w = (1-w)\theta_1 + w\theta_2$ where $w \in [0,1]$. Define $h(w) = -f(\theta_w)$, to show that $f(\theta)$ is concave it suffices to see that $h(w)$ is convex for

arbitrary (w, θ_1, θ_2). Straightforward algebra shows that $\frac{\partial}{\partial w}h(w) = -g(\theta_w)(\theta_2 - \theta_1)$ and further derivation shows that

$$\frac{\partial^2}{\partial w^2}h(w) = -(\theta_2 - \theta_1)^T H(\theta_w)(\theta_2 - \theta_1) > 0,$$

since $H(\theta)$ is negative definite (≥ 0 for negative semidefinite).

The second derivative $\frac{\partial^2}{\partial w^2}h(w) > 0$ almost everywhere and the first derivative $\frac{\partial}{\partial w}h(w)$ is continuous, which implies that $\frac{\partial}{\partial w}h(w)$ is strictly increasing in w and hence $h(w)$ is strictly convex (non-strictly convex when $H(\theta)$ is negative semidefinite). □

Proof of Proposition 1, Part (i)

The gradient $g_1(\theta, \vartheta, \alpha)$ follows from straightforward algebra, which is obviously continuous with respect to $\vartheta \in \mathbb{R}^+$ and $\alpha \in [-1, 1]$. To see continuity of $g_1(\theta, \vartheta, \alpha)$ with respect to θ, consider increasing a single θ_j for some $j \in \{1, \ldots, p\}$ and fix the remaining elements in θ, which we denote $\theta_{(-j)}$. Also denote $x_{i(-j)}$ the subvector of x_i obtained by removing x_{ij}. Clearly, $\log L_1(\theta, \vartheta, \alpha)$ is quadratic in θ_j with coefficients that stay constant until θ_j increases beyond a value t such that an observation i^* is added to or removed from $A(\theta)$, *i.e.* $y_{i^*} < x_{i^*(-j)}^T \theta_{(-j)} + x_{i^*j}\theta_j$ for $\theta_j \leq t$ and $y_{i^*} > x_{i^*(-j)}^T \theta_{(-j)} + x_{i^*j}\theta_j$ for $\theta_j > t$. Taking the limit of the contribution of i^* to $\log\left(L_1(\theta, \vartheta, \alpha)\right)$ as either $\theta_j \to t^-$ or $\theta_j \to t^+$ we obtain

$$\lim_{\theta_j \to t^+} \frac{(y_{i^*} - x_{i^*}\theta_i)^2}{(1+\alpha)^2} = \lim_{\theta_j \to t^-} \frac{(y_{i^*} - x_{i^*}\theta_i)^2}{(1-\alpha)^2} = 0,$$

i.e. $\log\left(L_1(\theta, \vartheta, \alpha)\right)$ is continuous. Similarly, taking the limits for the contribution to the first partial derivative with respect to θ_j gives

$$\lim_{\theta_j \to t^+} \frac{2(y_{i^*} - x_{i^*}\theta_i)}{(1+\alpha)^2} = \lim_{\theta_j \to t^-} \frac{2(y_{i^*} - x_{i^*}\theta_i)}{(1-\alpha)^2} = 0,$$

which proves that $g_1(\theta, \vartheta, \alpha)$ is continuous.

Proof of Proposition 1, Part (ii)

The form of $H_1(\theta, \vartheta, \alpha)$ follows from easy algebra.

Proof of Proposition 1, Part (iii)

We start by noting that the maximum of the asymmetric-normal log-likelihood with respect to (θ, α) does not depend on ϑ, hence we simply need to see that

$$H = \begin{pmatrix} X^T W^2 X & 2X^T \overline{W}^3 (y - X\theta) \\ 2(y - X\theta)^T \overline{W}^3 X & 3(y - X\theta)^T W^2 (y - X\theta) \end{pmatrix}, \tag{2}$$

is positive definite for almost all (θ, α). Once we show this, by Part (i) and Lemma 1 we have that there is a unique maximum.

To see that H is positive definite, we shall show that all its leading principal minors are positive. Note that $X^T W^2 X$ is the gram matrix corresponding to WX and is hence positive definite when $\text{rank}(WX) = p$, or equivalently when $\text{rank}(X) = p$ given that the effect of W is to simply re-scale the rows of X. If $\text{rank}(WX) < p$ then $X^T W^2 X$ is positive semidefinite. Therefore, we just need to check that $\det(H) > 0$. Now, the usual formula for determinant based on submatrices gives that $\det(H) = \det(X^T W^2 X)\det(B)$, where $B =$

$$\begin{aligned} &3(y - X\theta)^T W^4 (y - X\theta) - 4(y - X\theta)^T \overline{W}^3 X (X^T W^2 X)^{-1} X^T \overline{W}^3 (y - X\theta) \\ &\quad = 3(y - X\theta)^T W^2 \left(I - \frac{4}{3} \overline{W} X (X^T W^2 X)^{-1} X^T \overline{W} \right) W^2 (y - X\theta), \end{aligned} \tag{3}$$

is a scalar, I is the $n \times n$ identity matrix, as usual W is an $n \times n$ diagonal matrix with entries $1/(1 \pm \alpha)^2$ where the $\pm$ depends on whether $i \in A(\theta)$ or $i \notin A(\theta)$, and similarly $\overline{W}$ is diagonal with entries $\pm(1 \pm \alpha)$. All that is left is to see that $B > 0$. For ease of notation let us define $Z = \overline{W} X$, given that $\overline{W}\overline{W} = \text{diag}(1/(1 \pm \alpha)^2) = W^2$ we can write

$$\begin{aligned} B = 3(y - X\theta)^T W^2 \left(I - \frac{4}{3} Z (Z^T Z)^{-1} Z^T \right) W^2 (y - X\theta) = \\ 4(y - X\theta)^T W^2 (I - Z(Z^T Z)^{-1} Z^T) W^2 (y - X\theta) - (y - X\theta)^T W^2 W^2 (y - X\theta) > 0 \\ \Leftrightarrow 4 \frac{(y - X\theta)^T W^2 (I - Z(Z^T Z)^{-1} Z^T) W^2 (y - X\theta)}{(y - X\theta)^T W^2 W^2 (y - X\theta)} - 1 > 0. \end{aligned} \tag{4}$$

To complete the proof, note that $a = W^2(y - X\theta) \in \mathbb{R}^n$ is simply a vector and that the hat matrix $Z(Z^TZ)^{-1}Z^T$ is symmetric and idempotent, which implies that it has $\text{rank}(Z)$ eigenvalues equal to 1 and $n - \text{rank}(Z)$ eigenvalues equal to 0. Thus $I - Z(Z^TZ)^{-1}Z^T$ has $n - \text{rank}(Z)$ eigenvalues equal to 1 and the remaining $\text{rank}(Z)$ eigenvalues equal to 0. Given that $n > \text{rank}(Z)$ by assumption, $I - Z(Z^TZ)^{-1}Z^T$ has at least one non-zero eigenvalue, which allows us to bound

$$\min_{a \in \mathbb{R}^n} \frac{a(I - Z(Z^TZ)^{-1}Z^T)a}{a^Ta} \geq 1,$$

which from (4) gives that $B \geq 3$ and hence that H is positive definite.

2.2 Proof of Proposition 2

Parts (i) and (ii) follow from straightforward algebra. For Part (iii) we first show that $\log L_2(\theta, \vartheta, \alpha)$ is (non-strictly) concave in (θ, α) and then that when $\text{rank}(X) = p$ it is strictly concave. To see non-strict concavity note that $-|y_i - x_i^T\theta|/(\sqrt{\vartheta}(1+\alpha)) = -\max\{y_i - x_i^T\theta, x_i^T\theta - y_i\}/(\sqrt{\vartheta}(1+\alpha))$ is the maximum of two (non-strictly) concave functions in (θ, α) and hence also concave, from which it follows that $L_3(\theta, \vartheta, \alpha)$ is a sum of concave functions and thus concave.

For ease of notation let $\eta = (\theta, \vartheta, \alpha)$, we now show that $\log L_2(\eta)$ is stricly concave at any arbitrary $\eta_1 = (\theta_1, \vartheta, \alpha_1)$ as long as $\text{rank}(X) = p$. It is useful to note that $H_2(\theta, \vartheta, \alpha)$ is strictly negative definite in α, as the corresponding minor $-2|W^3(y - X\theta)|/\sqrt{\vartheta} < 0$. From the definition of concavity and continuity of the log-likelihood, if $\log L_2(\eta)$ were concave but non-strictly concave at $\eta = \eta_1$ then for some $\eta_2 = (\theta_2, \vartheta, \alpha_2) \neq \eta_1$ we would have that $\log L_2(a\eta_1 + (1-a)\eta_2) = a \log L_2(\eta_1) + (1-a) \log L_2(\eta_2)$ for all $a \in [0, 1]$, *i.e.* $\log L_2(\eta)$ would be locally linear (in fact, constant) along the direction defined by $\eta_2 - \eta_1$, and in particular $\log L_2(\eta_1) = \log L_2(\eta_2)$. From its form

$$\log L_2(\eta) = -\frac{n}{2}\log(\vartheta) - \frac{1}{\vartheta}\left(\frac{\sum_{i \in A(\theta)} |y_i - x_i^T\theta|}{1+\alpha} + \frac{\sum_{i \notin A(\theta)} |y_i - x_i^T\theta|}{1-\alpha}\right),$$

is locally linear in θ but clearly non-linear in α, implying that $\alpha_2 = \alpha_1$. More formally, it is easy to see that for fixed $\theta_1 \neq \theta_2$ the roots of $\log L_2(\eta_1) = \log L_2(\eta_2)$ in terms of α_2 are given by the roots of a quadratic polynomial that are not linear in θ_2, thus the only possible linear solution is

$\alpha_2 = \alpha_1$. The problem is hence reduced to showing that there is no θ_2 sufficiently close to θ_1 such that

$$|W(y - X\theta_1)| = |W(y - X\theta_2)|, \tag{5}$$

where $|\cdot|$ denotes the L_1 norm and as usual W is a diagonal matrix with (i,i) element $(1+\alpha)^{-1}$ if $i \in A(\theta_1)$ and $(1-\alpha)^{-1}$ if $i \notin A(\theta_1)$, where we note that $A(\theta_2) = A(\theta_1)$ for θ_2 sufficiently close to θ_1 and thus the same weighting matrix W can be used in left and right hand sides of (5). Expression (5) is the L_1 error function featuring in median regression with re-scaled $\tilde{y} = Wy$ and $\tilde{X} = WX$, which is concave as long as $p = \text{rank}(WX) = \text{rank}(X)$, as we wished to prove.

2.3 Proof of Proposition 3

Two-piece normal errors ($k = 1$)

The proof strategy is as follows: we first show that the average log-likelihood $M_n(\theta_\gamma, \vartheta, \alpha) = \frac{1}{n} \log L_1(\theta_\gamma, \vartheta, \alpha)$ converges to its expected value $M(\theta_\gamma, \vartheta, \alpha)$ uniformly across $(\theta_\gamma, \vartheta, \alpha) \in \Gamma$, and later show that $M(\theta_\gamma, \vartheta, \alpha)$ has a unique maximum $(\theta^*_\gamma, \vartheta^*_\gamma, \alpha^*_\gamma)$, which jointly satisfy the conditions in Theorem 5.7 from van der Vaart (1998) for consistency of $(\widehat{\theta}_\gamma, \widehat{\vartheta}_\gamma, \widehat{\alpha}_\gamma) \xrightarrow{P} (\theta^*_\gamma, \vartheta^*_\gamma, \alpha^*_\gamma)$.

We remark that Condition A3 is met for instance by deterministic sequences $\{x_i\}$ satisfying the stated positive-definiteness condition and also by $x_i \overset{i.i.d.}{\sim} \Psi$ as long as $E(x_1 x_1^T) = \Sigma$ for some positive definite Σ, since then $n^{-1} X^T X \xrightarrow{a.s.} \Sigma$ by the strong law of large numbers, and given that eigenvalues are continuous functions of $X^T X$ by the continuous mapping theorem $X^T X$ is positive definite almost surely as $n \to \infty$. Finally, Γ is assumed to contain the maximizer $(\theta^*_\gamma, \vartheta^*_\gamma, \alpha^*_\gamma)$.

By the law of large numbers and the *i.i.d.* assumption, we have that $M_n(\theta_\gamma, \vartheta, \alpha) \xrightarrow{P} M(\theta_\gamma, \vartheta, \alpha)$,

for each $(\theta_\gamma, \vartheta, \alpha) \in \Gamma$. Next, we prove that the limit M is finite for all $(\theta_\gamma, \vartheta, \alpha) \in \Gamma$.

$$
\begin{aligned}
& |M(\theta_\gamma, \vartheta, \alpha)| = \left| \mathbb{E}\left[\log s_1(y_1 | x_1^T \theta_\gamma, \vartheta, \alpha) \right] \right| \leq \mathbb{E}\left[|\log s_1(y_1 | x_1^T \theta_\gamma, \vartheta, \alpha)| \right] \\
= & \int\int |\log s_1(y_1 | x_1^T \theta_\gamma, \vartheta, \alpha)| dS_0(y_1|x_1) d\Psi(x_1) \\
= & \int\int_{y_1 < x_1^T \theta_\gamma} \left| \log \frac{1}{\sqrt{\vartheta}} \phi\left(\frac{y_1 - x_1^T \theta_\gamma}{\sqrt{\vartheta}(1+\alpha)} \right) \right| dS_0(y_1|x_1) d\Psi(x_1) \\
+ & \int\int_{y_1 \geq x_1^T \theta_\gamma} \left| \log \frac{1}{\sqrt{\vartheta}} \phi\left(\frac{y_1 - x_1^T \theta_\gamma}{\sqrt{\vartheta}(1-\alpha)} \right) \right| dS_0(y_1|x_1) d\Psi(x_1).
\end{aligned}
$$

For the first term in the last inequality we obtain, by integrating over the whole space, assumption A4 with $j = 2$, and the triangle inequality, the following upper bound

$$
\begin{aligned}
& \int\int \left| \log \frac{1}{\sqrt{\vartheta}} \phi\left(\frac{y_1 - x_1^T \theta_\gamma}{\sqrt{\vartheta}(1+\alpha)} \right) \right| dS_0(y_1|x_1) d\Psi(x_1) \\
\leq & \ |\log \sqrt{2\pi\vartheta}| + \int\int \frac{(y_1 - x_1^T \theta_\gamma)^2}{2\vartheta(1+\alpha)^2} dS_0(y_1|x_1) d\Psi(x_1) < \infty.
\end{aligned}
$$

Analogously for the second term. Now, let $\vartheta = \vartheta^\star$ be an arbitrary fixed value for the (squared) scale parameter. The aim now is to first show that the average log-likelihood $M_n(\theta_\gamma, \vartheta^\star, \alpha) = n^{-1} \log L_1(\theta_\gamma, \vartheta^\star, \alpha)$ converges to its expected value $M(\theta_\gamma, \vartheta^\star, \alpha)$ uniformly in (θ_γ, α), which implies that $(\widehat{\theta}_\gamma, \widehat{\alpha}_\gamma) \xrightarrow{P} (\theta^*_\gamma, \alpha^*_\gamma)$, and to then exploit that $\widehat{\vartheta}_\gamma$ and ϑ^*_γ have simple expressions to show that $\widehat{\vartheta}_\gamma \xrightarrow{P} \vartheta^*_\gamma$. To see that $M_n(\theta_\gamma, \vartheta^\star, \alpha)$ converges to $M(\theta_\gamma, \vartheta^\star, \alpha)$ uniformly in (θ_γ, α) we use the result in Proposition 1 that for positive-definite $X^T X$ (which holds for $n > n_0$) we have that $M_n(\theta_\gamma, \vartheta^\star, \alpha)$ is a sequence of concave functions in (θ_γ, α), which by the convexity lemma in Pollard (1991) (see also Theorem 10.8 from Rockafellar (2015)) implies that

$$
\sup_{(\theta_\gamma, \alpha) \in K} |M_n(\theta_\gamma, \vartheta^\star, \alpha) - M(\theta_\gamma, \vartheta^\star, \alpha)| \xrightarrow{P} 0, \tag{6}
$$

for each compact set $K \subseteq \Gamma$, and also that $M(\theta_\gamma, \vartheta^\star, \alpha)$ is finite and concave in (θ_γ, α) and thus has a unique maximum $(\theta^*_\gamma, \alpha^*_\gamma)$. That is, for a distance measure $d()$ and every $\varepsilon > 0$ we have

$$
\sup_{d((\theta^*_\gamma, \vartheta^\star, \alpha^*_\gamma), (\theta, \vartheta^\star, \alpha)) \geq \varepsilon} M(\theta_\gamma, \vartheta^\star, \alpha) < M(\theta^*_\gamma, \vartheta^\star, \alpha^*_\gamma). \tag{7}
$$

The consistency of $(\widehat{\theta}_\gamma, \widehat{\alpha}_\gamma) \xrightarrow{P} (\theta^*_\gamma, \alpha^*_\gamma)$ follows directly from (6) and (7) together with Theorem 5.7 from van der Vaart (1998). To see that $\widehat{\vartheta}_\gamma \xrightarrow{P} \vartheta^*_\gamma$, note first that from

$$M(\theta_\gamma, \vartheta^\star, \alpha) = -\log\left(\sqrt{2\pi\vartheta^\star}\right) - \frac{1}{2\vartheta^\star}\int\left[\frac{(y_1 - x_1^T\theta_\gamma)^2}{(1+\alpha)^2} I(y_1 < x_1^T\theta_\gamma)\right. \tag{8}$$
$$\left. + \frac{(y_1 - x_1^T\theta_\gamma)^2}{(1-\alpha)^2} I(y_1 \geq x_1^T\theta_\gamma)\right] dS_0(y_1|x_1)d\Psi(x_1),$$

we see that $(\theta^*_\gamma, \alpha^*_\gamma)$ does not depend on $\vartheta^\star$, thus $(\theta^*_\gamma, \alpha^*_\gamma)$ is a global maximum. From (8) $M(\theta^*_\gamma, \vartheta, \alpha^*_\gamma)$ trivially has the maximizer

$$\vartheta^*_\gamma = \int\left[\frac{(y_1 - x_1^T\theta^*_\gamma)^2}{(1+\alpha^*_\gamma)^2} I(y_1 < x_1^T\theta^*_\gamma) + \frac{(y_1 - x_1^T\theta^*_\gamma)^2}{(1-\alpha^*_\gamma)^2} I(y_1 \geq x_1^T\theta^*_\gamma)\right] dS_0(y_1|x_1)d\Psi(x_1),$$

and from the likelihood equations we have that

$$\widehat{\vartheta}_\gamma = \frac{1}{n}\left(\sum_{i=1}^n \frac{(y_i - x_i^T\widehat{\theta}_\gamma)^2}{(1+\widehat{\alpha}_\gamma)^2} I(y_i \leq x_i^T\widehat{\theta}_\gamma) + \frac{(y_i - x_i^T\widehat{\theta}_\gamma)^2}{(1-\widehat{\alpha}_\gamma)^2} I(y_i > x_i^T\widehat{\theta}_\gamma)\right). \tag{9}$$

In order to simplify notation, let us define

$$\rho(y_i, x_i, \theta_\gamma, \alpha) = \frac{(y_i - x_i^T\theta_\gamma)^2}{(1+\alpha)^2} I(y_i \leq x_i^T\theta_\gamma) + \frac{(y_i - x_i^T\theta_\gamma)^2}{(1-\alpha)^2} I(y_i > x_i^T\theta_\gamma).$$

Then, by the triangle inequality

$$\left|\widehat{\vartheta}_\gamma - \vartheta^*_\gamma\right| \leq \left|\widehat{\vartheta}_\gamma - \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma)\right| + \left|\frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma) - \vartheta^*_\gamma\right|.$$

For the second term it follows, by the law of large numbers, that

$$\left|\frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma) - \vartheta^*_\gamma\right| \xrightarrow{P} 0.$$

For the first term we have

$$\begin{aligned}\left|\widehat{\vartheta}_\gamma - \frac{1}{n}\sum_{i=1}^{n}\rho(y_i,x_i,\theta^*_\gamma,\alpha^*_\gamma)\right| &= \left|M_n(\widehat{\theta}_\gamma,1/2,\widehat{\alpha}_\gamma) - M_n(\theta^*_\gamma,1/2,\alpha^*_\gamma)\right| \\ &\leq \left|M_n(\widehat{\theta}_\gamma,1/2,\widehat{\alpha}_\gamma) - M(\widehat{\theta}_\gamma,1/2,\widehat{\alpha}_\gamma)\right| \\ &+ \left|M(\widehat{\theta}_\gamma,1/2,\widehat{\alpha}_\gamma) - M(\theta^*_\gamma,1/2,\alpha^*_\gamma)\right| \\ &+ \left|M(\theta^*_\gamma,1/2,\alpha^*_\gamma) - M_n(\theta^*_\gamma,1/2,\alpha^*_\gamma)\right| \\ &\leq 2\sup_{(\theta_\gamma,\alpha)\in\Gamma}|M_n(\theta_\gamma,1/2,\alpha) - M(\theta_\gamma,1/2,\alpha)| \\ &+ \left|M(\widehat{\theta}_\gamma,1/2,\widehat{\alpha}) - M(\theta^*_\gamma,1/2,\alpha^*_\gamma)\right|.\end{aligned}$$

By using (6), the consistency of $(\widehat{\theta}_\gamma,\widehat{\alpha}_\gamma)$, and the continuous mapping theorem it follows that $\left|\widehat{\vartheta}_\gamma - \frac{1}{n}\sum_{i=1}^{n}\rho(y_i,x_i,\theta^*_\gamma,\alpha^*_\gamma)\right| \xrightarrow{P} 0$. Consequently, $\widehat{\vartheta} \xrightarrow{P} \vartheta^*_\gamma$, which completes the proof.

Two-piece Laplace errors ($k=2$)

The proof strategy is analogous to that with $k=1$. Denote $M_n(\theta_\gamma,\vartheta,\alpha) = \frac{1}{n}\log L_2(\theta_\gamma,\vartheta,\alpha)$. By the law of large numbers, we have that $M_n(\theta_\gamma,\vartheta,\alpha) \xrightarrow{P} M(\theta_\gamma,\vartheta,\alpha)$, for each $(\theta_\gamma,\vartheta,\alpha)\in\Gamma$. Moreover,

$$\begin{aligned}&|M(\theta_\gamma,\vartheta,\alpha)| = \left|\mathbb{E}\left[\log s_2(y_1|x_1^T\theta_\gamma,\vartheta,\alpha)\right]\right| \leq \mathbb{E}\left[|\log s_2(y_1|x_1^T\theta_\gamma,\vartheta,\alpha)|\right] \\ =\ &\int |\log s_2(y_1|x_1^T\theta_\gamma,\vartheta,\alpha)|dS_0(y_1|x_1)d\Psi(x_1) \\ =\ &\int_{y<x^T\theta_\gamma}\left|\log\frac{1}{\sqrt{\vartheta}}f\left(\frac{y_1-x_1^T\theta_\gamma}{\sqrt{\vartheta}(1+\alpha)}\right)\right|dS_0(y_1|x_1)d\Psi(x_1) \\ +\ &\int_{y_1\geq x_1^T\theta_\gamma}\left|\log\frac{1}{\sqrt{\vartheta}}f\left(\frac{y_1-x_1^T\theta_\gamma}{\sqrt{\vartheta}(1-\alpha)}\right)\right|dS_0(y_1|x_1)d\Psi(x_1),\end{aligned}$$

where $f(z) = 0.5\exp(-|z|)$. For the first term in the last inequality we have, by integrating over the whole space and the triangle inequality, the following upper bound

$$\begin{aligned}&\int\left|\log\frac{1}{\sqrt{\vartheta}}f\left(\frac{y_1-x_1^T\theta_\gamma}{\sqrt{\vartheta}(1+\alpha)}\right)\right|dS_0(y_1|x_1)d\Psi(x_1) \\ \leq\ &|\log 2\sqrt{\vartheta}| + \int\frac{|y_1-x_1^T\theta_\gamma|}{\sqrt{\vartheta}(1+\alpha)}dS_0(y_1|x_1)d\Psi(x_1) < \infty,\end{aligned}$$

where the finiteness follows by assumption A4 with $j = 1$. An analogous result is obtained for the second term. Now, let $\vartheta = \vartheta^\star$ be an arbitrary fixed value for the (squared) scale parameter. From Proposition 2, it follows that for positive-definite $X^T X$ (which is guaranteed by assumption A2, for $n > n_0$) we have that $M_n(\theta_\gamma, \vartheta^\star, \alpha)$ is concave in (θ, α), which by the convexity lemma in Pollard (1991) implies that

$$\sup_{(\theta_\gamma,\alpha)\in K} |M_n(\theta_\gamma, \vartheta^\star, \alpha) - M(\theta_\gamma, \vartheta^\star, \alpha)| \xrightarrow{P} 0, \tag{10}$$

for any compact set $K \subseteq \Gamma$, and also that $M(\theta_\gamma, \vartheta^\star, \alpha)$ is concave in (θ_γ, α) and thus has a unique maximum $(\theta^*_\gamma, \alpha^*_\gamma)$. That is, for a distance measure $d()$ and every $\varepsilon > 0$ we have

$$\sup_{d((\theta^*_\gamma,\vartheta^\star,\alpha^*_\gamma),(\theta_\gamma,\vartheta^\star,\alpha))\geq\varepsilon} M(\theta_\gamma, \vartheta^\star, \alpha) < M(\theta^*_\gamma, \vartheta^\star, \alpha^*_\gamma). \tag{11}$$

The consistency of $(\widehat{\theta}_\gamma, \widehat{\alpha}_\gamma) \xrightarrow{P} (\theta^*_\gamma, \alpha^*_\gamma)$ follows directly from (10) and (11) together with Theorem 5.7 from van der Vaart (1998). To see that $\widehat{\vartheta}_\gamma \xrightarrow{P} \vartheta^*_\gamma$, note first that from

$$\begin{aligned} M(\theta_\gamma, \vartheta^\star, \alpha) = &- \log\left(2\sqrt{\vartheta^\star}\right) - \frac{1}{\sqrt{\vartheta^\star}} \int \left[\frac{|y_1 - x_1^T\theta_\gamma|}{1+\alpha} I(y_1 < x_1^T\theta_\gamma) \right. \\ &\left. + \frac{|y_1 - x_1^T\theta_\gamma|}{1-\alpha} I(y_1 \geq x_1^T\theta_\gamma) \right] dS_0(y_1|x_1)\Psi(x_1), \end{aligned} \tag{12}$$

we see that $(\theta^*_\gamma, \alpha^*_\gamma)$ does not depend on $\vartheta^\star$, thus $(\theta^*_\gamma, \alpha^*_\gamma)$ is a global maximum. From (8) $M(\theta^*_\gamma, \vartheta, \alpha^*_\gamma)$ trivially has the maximizer

$$\vartheta^*_\gamma = \left\{ \int \left[\frac{|y_1 - x_1^T\theta^*_\gamma|}{1+\alpha^*_\gamma} I(y_1 < x_1^T\theta^*_\gamma) + \frac{|y_1 - x_1^T\theta^*_\gamma|}{1-\alpha^*_\gamma} I(y_1 \geq x_1^T\theta^*_\gamma) \right] dS_0(y_1|x_1) d\Psi(x_1) \right\}^2,$$

and from the likelihood equations we have that

$$\widehat{\vartheta}_\gamma = \left[\frac{1}{n} \left(\sum_{i=1}^n \frac{|y_i - x_i^T\widehat{\theta}_\gamma|}{1+\widehat{\alpha}_\gamma} I(y_i \leq x_i^T\widehat{\theta}_\gamma) + \frac{|y_i - x_i^T\widehat{\theta}_\gamma|}{1-\widehat{\alpha}_\gamma} I(y_i > x_i^T\widehat{\theta}_\gamma) \right) \right]^2. \tag{13}$$

Let us define

$$\rho(y_i, x_i, \theta_\gamma, \alpha) = \frac{|y_i - x_i^T\theta_\gamma|}{1+\alpha} I(y_i \leq x_i^T\theta_\gamma) + \frac{|y_i - x_i^T\theta_\gamma|}{1-\alpha} I(y_i > x_i^T\theta_\gamma).$$

Then, by the triangle inequality

$$\left|\sqrt{\widehat{\vartheta}_\gamma} - \sqrt{\vartheta^*_\gamma}\right| \leq \left|\sqrt{\widehat{\vartheta}_\gamma} - \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma)\right| + \left|\frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma) - \sqrt{\vartheta^*_\gamma}\right|.$$

For the second term in the right-hand side of the last equation, it follows, by the law of large numbers and the continuous mapping theorem, that

$$\left|\frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma) - \sqrt{\vartheta^*_\gamma}\right| \xrightarrow{P} 0.$$

For the first term we have

$$\begin{aligned}
\left|\sqrt{\widehat{\vartheta}_\gamma} - \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma)\right| &= \left|M_n(\widehat{\theta}_\gamma, 1, \widehat{\alpha}_\gamma) - M_n(\theta^*_\gamma, 1, \alpha^*_\gamma)\right| \\
&\leq \left|M_n(\widehat{\theta}_\gamma, 1, \widehat{\alpha}_\gamma) - M(\widehat{\theta}_\gamma, 1, \widehat{\alpha}_\gamma)\right| \\
&+ \left|M(\widehat{\theta}_\gamma, 1, \widehat{\alpha}_\gamma) - M(\theta^*_\gamma, 1, \alpha^*_\gamma)\right| \\
&+ \left|M(\theta^*_\gamma, 1, \alpha^*_\gamma) - M_n(\theta^*_\gamma, 1, \alpha^*_\gamma)\right| \\
&\leq 2 \sup_{(\theta_\gamma, \alpha)\in\Gamma} |M_n(\theta_\gamma, 1, \alpha) - M(\theta_\gamma, 1, \alpha)| \\
&+ \left|M(\widehat{\theta}_\gamma, 1, \widehat{\alpha}_\gamma) - M(\theta^*_\gamma, 1, \alpha^*_\gamma)\right|.
\end{aligned}$$

By using (10), the consistency of $(\widehat{\theta}_\gamma, \widehat{\alpha}_\gamma)$, and the continuous mapping theorem it follows that $\left|\sqrt{\widehat{\vartheta}_\gamma} - \frac{1}{n}\sum_{i=1}^n \rho(y_i, x_i, \theta^*_\gamma, \alpha^*_\gamma)\right| \xrightarrow{P} 0$. Consequently, $\widehat{\vartheta}_\gamma \xrightarrow{P} \vartheta^*_\gamma$, which completes the proof.

2.4 Proof of Proposition 4

Two-piece normal errors ($k = 1$)

The proof technique consists of showing first that $\dot{m}_\eta(y_1, x_1)$ is dominated by an L^2 function (square integrable), $K(y_1, x_1)$, for η in a neighborhood of η^*_γ. Then, we prove that the function Pm_η admits a second-order Taylor expansion at η^*_γ and that the matrix $V_{\eta^*_\gamma}$ is nonsingular. Finally, we appeal

to the consistency result in Proposition 3 in order to apply Theorem 5.23 of van der Vaart (1998) to prove the asymptotic normality of $\widehat{\eta}_\gamma$.

We first note that under assumptions A1–A4, where A4 is assumed to be satisfied for $j = 4$ throughout, Proposition 3 implies the existence and uniqueness of η_γ^*. The gradient of $m_\eta(y_1, x_1)$, which is given by (i) in Proposition 1 (with $n = 1$), is bounded for all $\eta \in \Gamma$ and for each (y_1, x_1), due to the compactness of Γ. Now, a direct application of the Minkowski inequality implies that $||\dot{m}_\eta(y_1, x_1)||$ is upper bounded by the sum of the absolute values of the entries of $\dot{m}_\eta(y_1, x_1)$. Let us now define $K(y_1, x_1) = \sup_{\eta \in \mathcal{B}_{\eta_\gamma^*}} ||\dot{m}_\eta(y_1, x_1)||$, where $\mathcal{B}_{\eta_\gamma^*} \subset \Gamma$ is any neighborhood of η_γ^*, whose projection over θ coincides with $\mathcal{B}_{\theta_\gamma^*}$. Thus, from the expression of $\dot{m}_\eta(y_1, x_1)$ together with assumption A4, it follows that

$$\int K(y_1, x_1)^2 dS_0(y_1|x_1) d\Psi(x_1) < \infty,$$

Then, by using the mean value theorem and the Cauchy-Schwartz inequality, it follows that for $\eta_1, \eta_2 \in \mathcal{B}_{\eta_0}$, with probability 1,

$$\begin{aligned} |m_{\eta_1}(y_1, x_1) - m_{\eta_2}(y_1, x_1)| &= |\dot{m}_{\eta_\star}(y_1, x_1)^T(\eta_1 - \eta_2)| \\ &\leq ||\dot{m}_{\eta_\star}(y_1, x_1)|| \cdot ||\eta_1 - \eta_2|| \\ &\leq K(y_1, x_1) \cdot ||\eta_1 - \eta_2||, \end{aligned}$$

where $\eta_\star = (1 - c)\eta_1 + c\eta_2$, for some $c \in (0, 1)$.

Now, for each x_1:

$$\begin{aligned} Pm_{\eta|x_1} = \mathbb{E}[m_\eta|x_1] &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\vartheta) \\ &- \frac{1}{2\vartheta(1+\alpha)^2} \int_{-\infty}^{x_1^T\theta_\gamma} (y_1 - x_1^T\theta_\gamma)^2 dS_0(y_1|x_1) \\ &- \frac{1}{2\vartheta(1-\alpha)^2} \int_{x_1^T\theta_\gamma}^{\infty} (y_1 - x_1^T\theta_\gamma)^2 dS_0(y_1|x_1). \end{aligned}$$

Thus, the gradient of $Pm_{\eta|x_1}$ is given by

$$\begin{aligned}
\frac{\partial}{\partial\theta_\gamma}Pm_{\eta|x_1} &= -\frac{x_1}{\vartheta(1+\alpha)^2}I_1 + \frac{x_1}{\vartheta(1-\alpha)^2}I_2,\\
\frac{\partial}{\partial\vartheta}Pm_{\eta|x_1} &= -\frac{1}{2\vartheta} + \frac{I_3}{2\vartheta^2(1+\alpha)^2} + \frac{I_4}{2\vartheta^2(1-\alpha)^2},\\
\frac{\partial}{\partial\alpha}Pm_{\eta|x_1} &= \frac{I_3}{\vartheta(1+\alpha)^3} - \frac{I_4}{\vartheta(1-\alpha)^3},
\end{aligned}$$

Then, the second derivative matrix is given by

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_\gamma^2}Pm_{\eta|x_1} &= -\frac{x_1x_1^T[(1+\alpha)^2 - 4\alpha S_0(x_1^T\theta_\gamma|x_1)]}{\vartheta(1-\alpha^2)^2},\\
\frac{\partial^2}{\partial\vartheta^2}Pm_{\eta|x_1} &= \frac{1}{2\vartheta^2} - \frac{I_3}{\vartheta^3(1+\alpha)^2} - \frac{I_4}{\vartheta^3(1-\alpha)^2},\\
\frac{\partial^2}{\partial\alpha^2}Pm_{\eta|x_1} &= -\frac{3I_3}{\vartheta(1+\alpha)^4} - \frac{3I_4}{\vartheta(1-\alpha)^4},\\
\frac{\partial^2}{\partial\vartheta\partial\theta_\gamma}Pm_{\eta|x_1} &= \frac{x_1}{\vartheta^2(1+\alpha)^2}I_1 - \frac{x_1}{\vartheta^2(1-\alpha)^2}I_2,\\
\frac{\partial^2}{\partial\alpha\partial\theta_\gamma}Pm_{\eta|x_1} &= \frac{2x_1}{\vartheta(1+\alpha)^3}I_1 + \frac{2x_1}{\vartheta(1-\alpha)^3}I_2,\\
\frac{\partial^2}{\partial\vartheta\partial\alpha}Pm_{\eta|x_1} &= -\frac{I_3}{\vartheta^2(1+\alpha)^3} + \frac{I_4}{\vartheta^2(1-\alpha)^3},
\end{aligned}$$

where $I_1 = \int_{-\infty}^{x_1^T\theta_\gamma} S_0(y_1|x_1)dy_1$, and $I_2 = \int_{x_1^T\theta_\gamma}^{\infty}[1 - S_0(y_1|x_1)]\,dy_1$, $I_3 = \int_{-\infty}^{x_1^T\theta_\gamma}(y_1 - x_1^T\theta_\gamma)^2 dS_0(y_1|x_1)$, and $I_4 = \int_{x_1^T\theta_\gamma}^{\infty}(y_1 - x_1^T\theta_\gamma)^2 dS_0(y_1|x_1)$. These entries are finite for all $\eta \in \Gamma$ by assumption A4. Note that $Pm_\eta = \mathbb{E}[Pm_{\eta|x_1}]$, where the expectation is taken over x_1. Assumptions A1–A4 together with Proposition (3) imply that Pm_η is finite and that this expectation is concave and has a unique maximum at η^*_γ. From assumption A5,

$$\begin{aligned}
\frac{\partial}{\partial\theta_\gamma}Pm_\eta\bigg|_{\eta=\eta^*_\gamma} &= \mathbb{E}\left[\frac{\partial}{\partial\theta_\gamma}Pm_{\eta|x_1}\right]\bigg|_{\eta=\eta^*_\gamma} = 0,\\
\frac{\partial}{\partial\alpha}Pm_\eta\bigg|_{\eta=\eta^*_\gamma} &= \mathbb{E}\left[\frac{\partial}{\partial\alpha}Pm_{\eta|x_1}\right]\bigg|_{\eta=\eta^*_\gamma} = 0,
\end{aligned}$$

which in turn imply that $\frac{\partial^2}{\partial\vartheta\partial\theta_\gamma}Pm_\eta = 0$ and $\frac{\partial^2}{\partial\vartheta\partial\alpha}Pm_\eta = 0$ at $\eta = \eta^*_\gamma$. Thus, the matrix of

second derivatives evaluated at η_γ^* has the following structure:

$$V_\eta = \begin{pmatrix} \frac{\partial^2}{\partial\theta_\gamma^2} Pm_\eta & 0 & \frac{\partial^2}{\partial\vartheta\partial\alpha} Pm_\eta \\ 0 & \frac{\partial^2}{\partial\vartheta^2} Pm_\eta & 0 \\ \frac{\partial^2}{\partial\vartheta\partial\alpha} Pm_\eta & 0 & \frac{\partial^2}{\partial\alpha^2} Pm_\eta \end{pmatrix}.$$

Consequently, the determinant of this matrix is given by

$$\det V_\eta = \frac{\partial^2}{\partial\vartheta^2} Pm_\eta \times \det \begin{pmatrix} \frac{\partial^2}{\partial\theta_\gamma^2} Pm_\eta & \frac{\partial^2}{\partial\vartheta\partial\alpha} Pm_\eta \\ \frac{\partial^2}{\partial\vartheta\partial\alpha} Pm_\eta & \frac{\partial^2}{\partial\alpha^2} Pm_\eta \end{pmatrix}.$$

The determinant on the right-hand side of this expression, evaluated at η_γ^*, is non-zero since the Pm_η is concave with respect to (θ_γ, α), as shown in Proposition 3. Moreover, the fact that the first derivative $\frac{\partial}{\partial\vartheta} Pm_\eta = 0$ at $\eta = \eta_\gamma^*$ together with the fact that η_γ^* is the unique maximizer implies that $\frac{\partial^2}{\partial\vartheta^2} Pm_\eta \neq 0$. Consequently, the matrix of second derivatives of Pm_η is nonsingular at η_γ^*. The asymptotic normality result follows by Theorem 5.23 from van der Vaart (1998).

Two-piece Laplace errors ($k = 2$)

First, we note that under assumptions A1–A4, where $j = 2$ in A4 throughout, Proposition 3 implies the existence and uniqueness of η_γ^*. The gradient of $m_\eta(y_1, x_1)$, which is given by (i) in Proposition 2 (with $n = 1$), is bounded for almost all $\eta \in \Gamma$ and for each (y_1, x_1), due to the compactness of Γ. Now, a direct application of the Minkowski inequality implies that $||\dot{m}_\eta(y_1, x_1)||$ is upper bounded almost surely by the sum of the absolute values of the entries of $\dot{m}_\eta(y_1, x_1)$. Let us now define $K(y_1, x_1) = \sup_{\eta \in \mathcal{B}_{\eta_\gamma^*}} ||\dot{m}_\eta(y_1, x_1)||$, where $\mathcal{B}_{\eta_\gamma^*} \subset \Gamma$ is any neighborhood of η_γ^*, whose projection over θ_γ coincides with $\mathcal{B}_{\theta_\gamma^*}$. Thus, from the expression of $\dot{m}_\eta(y_1, x_1)$ together with assumption A4, it follows that

$$\int K(y_1, x_1)^2 dS_0(y_1|x_1) d\Psi(x_1) < \infty,$$

Then, by using the mean value theorem and the Cauchy-Schwartz inequality, it follows that for $\eta_1, \eta_2 \in \mathcal{B}_{\eta_\gamma^*}$, with probability 1,

$$\begin{aligned} |m_{\eta_1}(y_1, x_1) - m_{\eta_2}(y_1, x_1)| &= |\dot{m}_{\eta_\star}(y_1, x_1)^T(\eta_1 - \eta_2)| \\ &\leq ||\dot{m}_{\eta_\star}(y_1, x_1)|| \cdot ||\eta_1 - \eta_2|| \\ &\leq K(y_1, x_1) \cdot ||\eta_1 - \eta_2||, \end{aligned}$$

where $\eta_\star = (1-c)\eta_1 + c\eta_2$, for some $c \in (0,1)$.

Now, for each x_1:

$$\begin{aligned} Pm_{\eta|x_1} = \mathbb{E}[m_\eta|x_1] &= -\log(2) - \frac{1}{2}\log(\vartheta) - \frac{1}{\sqrt{\vartheta}(1+\alpha)} \int_{-\infty}^{x_1^T\theta_\gamma} S_0(y_1|x_1)dy_1 \\ &- \frac{1}{\sqrt{\vartheta}(1-\alpha)} \int_{x_1^T\theta_\gamma}^{\infty} 1 - S_0(y_1|x_1)dy_1. \end{aligned}$$

Then, the gradient of $Pm_{\eta|x_1}$ is given by

$$\begin{aligned} \frac{\partial}{\partial\theta_\gamma} Pm_{\eta|x_1} &= -\frac{x_1 S_0(x_1^T\theta_\gamma|x_1)}{\sqrt{\vartheta}(1+\alpha)} + \frac{x_1[1 - S_0(x_1^T\theta_\gamma|x_1)]}{\sqrt{\vartheta}(1-\alpha)}, \\ \frac{\partial}{\partial\vartheta} Pm_{\eta|x_1} &= -\frac{1}{2\vartheta} + \frac{I_1}{2\vartheta^{3/2}(1+\alpha)} + \frac{I_2}{2\vartheta^{3/2}(1-\alpha)}, \\ \frac{\partial}{\partial\alpha} Pm_{\eta|x_1} &= \frac{I_1}{\sqrt{\vartheta}(1+\alpha)^2} - \frac{I_2}{\sqrt{\vartheta}(1-\alpha)^2}, \end{aligned}$$

where $I_1 = \int_{-\infty}^{x_1^T\theta_\gamma} S_0(y_1|x_1)dy$, and $I_2 = \int_{x_1^T\theta_\gamma}^{\infty} 1 - S_0(y_1|x_1)dy_1$, which are finite by assumption A4.

Then, the second derivative matrix is given by

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_\gamma^2}Pm_{\eta|x_1} &= -\frac{2x_1x_1^T s_0(x_1^T\theta_\gamma|x_1)}{\sqrt{\vartheta}(1-\alpha^2)},\\
\frac{\partial^2}{\partial\vartheta^2}Pm_{\eta|x_1} &= \frac{1}{2\vartheta^2}-\frac{3I_1}{4\vartheta^{5/2}(1+\alpha)}-\frac{3I_2}{4\vartheta^{5/2}(1-\alpha)},\\
\frac{\partial^2}{\partial\alpha^2}Pm_{\eta|x_1} &= -\frac{2I_1}{\sqrt{\vartheta}(1+\alpha)^3}-\frac{2I_2}{\sqrt{\vartheta}(1-\alpha)^3},\\
\frac{\partial^2}{\partial\vartheta\partial\theta_\gamma}Pm_{\eta|x_1} &= \frac{x_1S_0(x_1^T\theta_\gamma|x_1)}{2\vartheta^{3/2}(1+\alpha)}-\frac{x_1[1-S_0(x_1^T\theta_\gamma|x_1)]}{2\vartheta^{3/2}(1-\alpha)},\\
\frac{\partial^2}{\partial\alpha\partial\theta_\gamma}Pm_{\eta|x_1} &= \frac{x_1S_0(x_1^T\theta_\gamma|x_1)}{\sqrt{\vartheta}(1+\alpha)^2}+\frac{x_1[1-S_0(x_1^T\theta_\gamma|x_1)]}{\sqrt{\vartheta}(1-\alpha)^2},\\
\frac{\partial^2}{\partial\vartheta\partial\alpha}Pm_{\eta|x_1} &= -\frac{I_1}{2\vartheta^{3/2}(1+\alpha)^2}+\frac{I_2}{2\vartheta^{3/2}(1-\alpha)^2}.
\end{aligned}$$

These entries are finite for all $\eta\in\Gamma$ by assumption A4. Note that $Pm_\eta=\mathbb{E}[Pm_{\eta|x_1}]$, where the expectation is taken over x_1. Assumptions A1–A4 together with Proposition 3, imply that Pm_η is finite and that this expectation is concave and has a unique maximum at η_γ^*. From assumption A5,

$$\begin{aligned}
\frac{\partial}{\partial\theta_\gamma}Pm_\eta\bigg|_{\eta=\eta_\gamma^*} &= \mathbb{E}\left[\frac{\partial}{\partial\theta_\gamma}Pm_{\eta|x_1}\right]\bigg|_{\eta=\eta_\gamma^*}=0,\\
\frac{\partial}{\partial\alpha}Pm_\eta\bigg|_{\eta=\eta_\gamma^*} &= \mathbb{E}\left[\frac{\partial}{\partial\alpha}Pm_{\eta|x_1}\right]\bigg|_{\eta=\eta_\gamma^*}=0,
\end{aligned}$$

which in turn imply that $\dfrac{\partial^2}{\partial\vartheta\partial\theta_\gamma}Pm_\eta=0$ and $\dfrac{\partial^2}{\partial\vartheta\partial\alpha}Pm_\eta=0$ at $\eta=\eta_\gamma^*$. Thus, it follows that the matrix of second derivatives evaluated at η_γ^* has the structure:

$$V_\eta=\begin{pmatrix}
\dfrac{\partial^2}{\partial\theta_\gamma^2}Pm_\eta & 0 & \dfrac{\partial^2}{\partial\vartheta\partial\alpha}Pm_\eta\\
0 & \dfrac{\partial^2}{\partial\vartheta^2}Pm_\eta & 0\\
\dfrac{\partial^2}{\partial\vartheta\partial\alpha}Pm_\eta & 0 & \dfrac{\partial^2}{\partial\alpha^2}Pm_\eta
\end{pmatrix}.$$

Consequently, the determinant of this matrix is given by

$$\det V_\eta = \frac{\partial^2}{\partial \vartheta^2} Pm_\eta \times \det \begin{pmatrix} \frac{\partial^2}{\partial \theta_\gamma^2} Pm_\eta & \frac{\partial^2}{\partial \vartheta \partial \alpha} Pm_\eta \\ \frac{\partial^2}{\partial \vartheta \partial \alpha} Pm_\eta & \frac{\partial^2}{\partial \alpha^2} Pm_\eta \end{pmatrix}.$$

The determinant on the right-hand side of this expression, evaluated at η_γ^*, is non-zero since the Pm_η is concave with respect to (θ_γ, α), as shown in Proposition 3. Moreover, the fact that the first derivative $\frac{\partial}{\partial \vartheta} Pm_\eta = 0$ at $\eta = \eta_\gamma^*$ together with the fact that η_γ^* is the unique maximizer implies that $\frac{\partial^2}{\partial \vartheta^2} Pm_\eta \neq 0$. Consequently, the matrix of second derivatives of Pm_η is nonsingular at η_γ^*. The asymptotic normality result follows by Theorem 5.23 from van der Vaart (1998).

2.5 Proof of Corollary 1

The result when $\epsilon_i \sim L(0, \vartheta)$ follows directly from Pollard (1991) Theorem 1, hence it suffices to find the expression for f_0 under each assumed residual distribution. The median for a general two-piece distribution with a mode at 0 is given by $\sqrt{\vartheta}(1+\alpha)F^{-1}\left(\frac{1}{2(1+\alpha)}\right)$ if $\alpha > 0$ and $\sqrt{\vartheta}(1-\alpha)F^{-1}\left(\frac{(1-2\alpha)}{2(1-\alpha)}\right)$ if $\alpha \leq 0$, where $F(\cdot)$ is the cdf of the standard underlying distribution with mode 0, $\vartheta = 1$ (Arellano-Valle et al. (2005), Expression (9)).

When $\epsilon_i \sim N(0, \vartheta)$ we have $m = 0$ and hence $f_0 = N(0; 0, \vartheta) = 1/(\sqrt{2\pi\vartheta})$. When $\epsilon_i \sim \text{AN}(0, \vartheta, \alpha)$ we have $m = \sqrt{\vartheta}(1+\alpha)\Phi^{-1}(0.5/(1+\alpha))$ if $\alpha > 0$ and $m = \sqrt{\vartheta}(1-\alpha)\Phi^{-1}(0.5(1-2\alpha)/(1-\alpha))$ if $\alpha < 0$, where $\Phi^{-1}(\cdot)$ is the inverse standard cdf, and hence $f_0 = \exp\left\{-\frac{1}{2}\left(\Phi^{-1}\left(\frac{0.5}{1+|\alpha|}\right)\right)^2\right\}\frac{1}{\sqrt{2\pi\vartheta}}$. For the Laplace and Asymmetric Laplace, we note that the inverse cdf of the standard Laplace distribution evaluated at a quantile $q \in [0, 1]$ is $F^{-1}(q) = \log(2q)$ if $q < 0.5$ and $F^{-1}(q) = -\log(2(1-q))$ if $q \geq 0.5$. When $\epsilon_i \sim L(0, \vartheta)$ we have $m = 0$ and $f_0 = 1/(2\sqrt{\vartheta})$. Finally, when $\epsilon_i \sim \text{AL}(0, \vartheta, \alpha)$ we have $m = -\sqrt{\vartheta}(1+\alpha)\log(1+\alpha)$ if $\alpha > 0$ and $m = \sqrt{\vartheta}(1-\alpha)\log(1-\alpha)$ if $\alpha < 0$, from which it follows that $f_0 = \frac{1}{2\sqrt{\vartheta}}\exp\{-\log(1+|\alpha|)\} = \frac{1}{2\sqrt{\vartheta}(1+|\alpha|)}$.

The results for the true Normal model follows by using classic asymptotic results on least square estimators (see *e.g.* Newey and Powell (1987))

2.6 Proof of Proposition 5

We provide the proof for the asymmetric Normal and asymmetric Laplace ($\alpha \neq 0$), their symmetric counterparts follow as particular cases. $\eta_\gamma = (\theta_\gamma, \vartheta_\gamma, \alpha_\gamma)$ denotes the parameter vector under model γ, $\widehat{\eta}_\gamma$ the MLE and $\tilde{\eta}_\gamma$ the posterior mode for a given observed (y, X). Further, $M_k(\eta_\gamma) = E(\log L_k(\eta_\gamma))$ where the expectation is with respect to the data-generating truth and $\eta^*_\gamma = \arg\max_{\eta \in \Gamma_\gamma} M_k(\eta_\gamma)$ is the optimal parameter value under γ. We wish to characterize the asymptotic behaviour of the Laplace-approximated Bayes factors

$$\frac{\widehat{p}(y \mid \gamma)}{\widehat{p}(y \mid \gamma^*)} = e^{\log L_k(\tilde{\eta}_\gamma) - \log L_k(\tilde{\eta}_{\gamma^*})} \times \frac{p(\tilde{\eta}_\gamma \mid \gamma)}{p(\tilde{\eta}_{\gamma^*} \mid \gamma^*)} \times (2\pi)^{\frac{p_\gamma - p_{\gamma^*}}{2}} \times \frac{|H_k(\tilde{\eta}_{\gamma^*})|^{\frac{1}{2}}}{|H_k(\tilde{\eta}_\gamma)|^{\frac{1}{2}}}, \tag{14}$$

when (y, X) arise from the data-generating model in Condition A1, which may differ from the assumed model. The term $(2\pi)^{\frac{p_\gamma - p_{\gamma^*}}{2}}$ is a constant since p_γ and p_{γ^*} are fixed. The expression for H_1 is given by (24) and recall that for H_2 we are taking the asymptotic covariance in (26). Hence

$$|H_2(\tilde{\eta}_\gamma)| = n^{p_\gamma} \left| \frac{1}{n} H_2(\tilde{\eta}_\gamma) \right| = n^{p_\gamma} \begin{vmatrix} -\frac{1}{n} X^T X \frac{1}{\tilde{\vartheta}_\gamma (1 - \tilde{\alpha}_\gamma^2)} & \frac{\overline{x}}{\sqrt{\tilde{\vartheta}_\gamma}(1-\tilde{\alpha}_\gamma^2)} & 0 \\ \frac{\overline{x}}{\sqrt{\tilde{\vartheta}_\gamma}(1-\tilde{\alpha}_\gamma^2)} & -\frac{1}{4\tilde{\vartheta}_\gamma^2} & 0 \\ 0 & 0 & -\frac{2}{1-\tilde{\alpha}_\gamma^2} \end{vmatrix}.$$

The determinant converges in probability to a negative constant since $\tilde{\eta}_\gamma \xrightarrow{P} \eta^*_\gamma$ by Proposition 3, together with the continuous mapping theorem and the asymptotic Hessian (the limiting $-H_2$) being positive definite. An analogous argument applies to H_1, hence $n^{\frac{p_\gamma - p_{\gamma^*}}{2}} |H_k(\tilde{\eta}_{\gamma^*})|^{\frac{1}{2}} / |H_k(\tilde{\eta}_\gamma)|^{\frac{1}{2}} \xrightarrow{P} \tilde{a}_3$ for some constant $\tilde{a}_3 > 0$. In other words, $|H_k(\tilde{\eta}_{\gamma^*})|^{\frac{1}{2}} / |H_k(\tilde{\eta}_\gamma)|^{\frac{1}{2}} = O_p\left(n^{\frac{p_{\gamma^*} - p_\gamma}{2}}\right)$.

The proof strategy is to first show that when $M_k(\eta^*_\gamma) - M_k(\eta^*_{\gamma^*}) < 0$ (*i.e.* $\gamma^* \not\subset \gamma$) the log-first term of the right hand in (14) behaves asymptotically in probability as $-na_1$, for some constant $a_1 > 0$, and the logarithm of the second term converges in probability to a constant a_2. Thus,

$$\frac{1}{n} \log\left(\frac{\widehat{p}(y \mid \gamma)}{\widehat{p}(y \mid \gamma^*)}\right) = -a_1(1 + o_p(1)) + \frac{1}{n}\left(a_2 + \frac{p_\gamma - p_{\gamma^*}}{2}(a_3 - \log(n))\right) \xrightarrow{P} -a_1$$

where $a_3 = \log(2\pi) - \log(\tilde{a}_3)$, as we wish to prove. Subsequently we shall show that when $M_k(\eta^*_\gamma) - M_k(\eta^*_{\gamma^*}) = 0$ (the case $\gamma^* \subset \gamma$) the first term is essentially the likelihood ratio test statistic and is

$O_p(1)$, whereas, analogously to the results in Johnson and Rossell (2010) and Rossell and Telesca (2017), the second term converges to a positive constant under local priors, but it is $O_p(\tilde{b}_n)$ where $\tilde{b}_n = n^{p_{\gamma^*}-p_\gamma}$ under the pMOM prior and $\tilde{b}_n = e^{-c\sqrt{n}}$ for some $c > 0$ under the peMOM prior. This gives

$$\frac{\widehat{p}(y \mid \gamma)}{\widehat{p}(y \mid \gamma^*)} = e^{O_p(1)} O_p(\tilde{b}_n) O_p\left(n^{\frac{p_{\gamma^*}-p_\gamma}{2}}\right) = O_p(b_n)$$

where $b_n = n^{\frac{p_{\gamma^*}-p_\gamma}{2}}$ for local priors, $b_n = n^{3(p_{\gamma^*}-p_\gamma)/2}$ for the pMOM prior and $b_n = e^{-c\sqrt{n}} n^{\frac{p_{\gamma^*}-p_\gamma}{2}}$ for the peMOM prior, as we wish to prove.

Consider first the case when $\gamma^* \not\subset \gamma$, which implies $M_k(\eta_\gamma^*) - M_k(\eta_{\gamma^*}^*) < 0$. Then by continuity of $p(\eta_\gamma \mid \gamma)$ we have that $p(\tilde{\eta}_\gamma \mid \gamma) \xrightarrow{P} p(\eta_\gamma^* \mid \gamma) \geq 0$, and analogously $p(\tilde{\eta}_{\gamma^*} \mid \gamma^*) \xrightarrow{P} p(\eta_{\gamma^*}^* \mid \gamma^*) > 0$ (strict positivity is ensured by the assumption of prior positivity at $\eta_{\gamma^*}^*$). Hence $p(\tilde{\eta}_\gamma \mid \gamma)/p(\tilde{\eta}_{\gamma^*} \mid \gamma^*) \xrightarrow{P} a_2$ for some constant $a_2 \geq 0$. Note that $a_2 = 0$ when θ_γ^* contains some zeroes and hence a non-local prior would take the value $p(\eta_\gamma^* \mid \gamma) = 0$, but this gives even faster Bayes factor rates in favor of γ^*. Regarding $\log L_k(\tilde{\eta}_\gamma) - \log L_k(\tilde{\eta}_{\gamma^*})$, the law of large numbers and uniform convergence of $\log L_k$ to its expected value shown in Proposition 3 give that

$$\frac{1}{n}(\log L_k(\tilde{\eta}_\gamma) - \log L_k(\tilde{\eta}_{\gamma^*})) \xrightarrow{P} (M_k(\eta_\gamma^*) - M_k(\eta_{\gamma^*}^*)) < 0, \tag{15}$$

hence the constant a_1 defined above is $a_1 = M_k(\eta_{\gamma^*}^*) - M_k(\eta_\gamma^*) > 0$.

Next consider the case when $\gamma^* \subset \gamma$, which implies $M_k(\eta_\gamma^*) - M_k(\eta_{\gamma^*}^*) = 0$. Since $\tilde{\eta}_\gamma \xrightarrow{P} \eta_\gamma^*$ by Proposition 3, we have that under a local prior

$$\frac{p(\tilde{\eta}_\gamma \mid \gamma)}{p(\tilde{\eta}_{\gamma^*} \mid \gamma^*)} \xrightarrow{P} \frac{p(\eta_\gamma \mid \gamma)}{p(\eta_{\gamma^*} \mid \gamma^*)} > 0. \tag{16}$$

Under a non-local prior we still have $p(\eta_{\gamma^*} \mid \gamma^*) > 0$ but in contrast $p(\eta_\gamma \mid \gamma) = 0$. Thus, it is necessary to characterize the rate at which the latter term vanishes. Briefly, following the proof of Theorem 1 in Koenker and Bassett (1982), the fact that $\log L_k$ converges uniformly to its expectation (see the proof of our Proposition 3) and consistency of $\tilde{\eta}_\gamma \xrightarrow{P} \eta_\gamma^*$ give that $\log L_k$ can be approximated by a quadratic function plus a term that is $o_p(1)$. Then, the argument leading to Rossell and Telesca (2017), Proposition 2(i), gives that $\tilde{\theta}_{\gamma j} - \widehat{\theta}_{\gamma j} = O_p(1/n)$ and thus

$\tilde{\theta}_{\gamma j} = O_p(n^{-1/2})$ under the pMOM prior p_M, whereas $\tilde{\theta}_{\gamma j} = O_p(n^{-1/4})$ under the peMOM prior p_E. It follows that $\pi_M(\tilde{\theta}_\gamma) = O_p(1)\prod_{\theta^*_{\gamma j}\neq 0}\tilde{\theta}^2_{\gamma j} = O_p(n^{-(p_\gamma - p_{\gamma^*})})$, and $\pi_E(\tilde{\eta}) = O_p(1)\prod_{\theta^*_{\gamma j}\neq 0} e^{O_p(1)/\tilde{\theta}^2_\gamma} = O_p(e^{-c\sqrt{n}})$ for some $c > 0$, as desired.

To conclude the proof, since $\log L_k(\tilde{\eta}_\gamma) - \log L_k(\tilde{\eta}_{\gamma^*}) = \lambda(y) + o_p(1)$ where $\lambda(y) = \log L_k(\widehat{\eta}_\gamma) - \log L_k(\widehat{\eta}_{\gamma^*})$ is the likelihood ratio (LR) statistic, it only remains to show that $\lambda(y) = O_p(1)$. The strategy is to see that $\lambda(y) = \lambda(y;\vartheta^*_\gamma)(1+o_p(1))$, where $\lambda(y;\vartheta^*_\gamma) = \log L_k(\widehat{\theta}_\gamma, \vartheta^*_\gamma, \widehat{\alpha}_\gamma) - \log L_k(\widehat{\theta}_{\gamma^*}, \vartheta^*_\gamma, \widehat{\alpha}_\gamma)$ is the LR obtained by plugging in the oracle $\vartheta^*_\gamma = \vartheta^*_{\gamma^*}$, then use classical results to prove that $\lambda(y;\vartheta^*_\gamma) = O_p(1)$. Taking derivatives of the likelihoods (Expressions (3) and (4) in the main paper) shows that for $k = 1$ the MLE must satisfy

$$\widehat{\vartheta}_\gamma = \frac{1}{n}\left(\sum_{i\in A(\theta)}\frac{(y_i - x_i^T\widehat{\theta}_\gamma)^2}{(1+\hat{\alpha})^2} + \sum_{i\notin A(\theta)}\frac{(y_i - x_i^T\widehat{\theta}_\gamma)^2}{(1-\hat{\alpha})^2}\right) = \frac{1}{n}(y - X_\gamma\widehat{\theta}_\gamma)^T W^2_{\hat{\theta}_\gamma,\hat{\alpha}}(y - X_\gamma\widehat{\theta}_\gamma),$$

whereas for $k = 2$ it satisfies

$$\widehat{\vartheta}^{\frac{1}{2}}_\gamma = \frac{1}{n}\left(\sum_{i\in A(\theta)}\frac{|y_i - x_i^T\widehat{\theta}_\gamma|}{(1+\hat{\alpha})} + \sum_{i\notin A(\theta)}\frac{|y_i - x_i^T\widehat{\theta}_\gamma|}{(1-\hat{\alpha})}\right) = \frac{1}{n}|W^{\frac{1}{2}}_{\hat{\theta}_\gamma,\hat{\alpha}}(y - X_\gamma\widehat{\theta}_\gamma)|.$$

Plugging $\widehat{\vartheta}_\gamma$ into the likelihoods gives

$$\begin{aligned}\lambda(y) &= -\frac{n}{2}\log\left(\frac{\widehat{\vartheta}_\gamma}{\widehat{\vartheta}_{\gamma^*}}\right) = -\frac{n}{2}\log\left(1 + \frac{\widehat{\vartheta}_\gamma - \widehat{\vartheta}_{\gamma^*}}{\widehat{\vartheta}_{\gamma^*}}\right) = -\frac{n}{2}\frac{\widehat{\vartheta}_\gamma - \widehat{\vartheta}_{\gamma^*}}{\widehat{\vartheta}_{\gamma^*}}(1+o_p(1)) \\ &= -\frac{n}{2}\frac{\widehat{\vartheta}_\gamma - \widehat{\vartheta}_{\gamma^*}}{\vartheta^*_{\gamma^*}}(1+o_p(1)) = \lambda(y;\vartheta^*_\gamma)(1+o_p(1)) \qquad (17)\end{aligned}$$

since by Proposition 3 $\widehat{\vartheta}_{\gamma^*} \xrightarrow{P} \vartheta^*_{\gamma^*} > 0$ and $(\widehat{\vartheta}_\gamma - \widehat{\vartheta}_{\gamma^*})/\widehat{\vartheta}_{\gamma^*} \xrightarrow{P} 0$.

Finally we show that $\lambda(y;\vartheta^*_\gamma) = O_p(1)$, which implies $\lambda(y;\vartheta^*_\gamma)(1+o_p(1)) = O_p(1)$ and completes the proof. For ease of notation when $k = 1$ define $Z_n(\gamma) = (y - X_\gamma\widehat{\theta}_\gamma)^T W^2_{\hat{\theta}_\gamma,\hat{\alpha}}(y - X_\gamma\widehat{\theta}_\gamma)$ and $Z(\gamma) = (y - X_\gamma\widehat{\theta}_\gamma)^T W^2_{\theta^*_\gamma,\alpha^*}(y - X_\gamma\widehat{\theta}_\gamma)$, and when $k = 2$ let $Z_n(\gamma) = |W^{\frac{1}{2}}_{\hat{\theta}_\gamma,\hat{\alpha}}(y - X_\gamma\widehat{\theta}_\gamma)|$, $Z(\gamma) = |W^{\frac{1}{2}}_{\theta^*_\gamma,\alpha^*}(y - X_\gamma\widehat{\theta}_\gamma)|$. Then by definition

$$\lambda(y;\vartheta^*_\gamma) = \frac{Z_n(\gamma^*) - Z_n(\gamma)}{2\vartheta^*_\gamma}. \qquad (18)$$

Now, note that $Z_n(\gamma) = Z(\gamma) + Z(\gamma)(Z_n(\gamma) - Z(\gamma))/Z(\gamma) = Z(\gamma)(1 + o_p(1))$, since Proposition 3 gives that $\frac{1}{n}Z_n(\gamma) \xrightarrow{P} \vartheta_\gamma$, $\frac{1}{n}Z(\gamma) \xrightarrow{P} \vartheta_\gamma$ and hence $(Z_n(\gamma) - Z(\gamma))/Z(\gamma) \xrightarrow{P} 0$. Following the same argument $Z_n(\gamma^*) = Z(\gamma^*)(1 + o_p(1))$, hence

$$\lambda(y; \vartheta^*_\gamma) = \frac{Z(\gamma^*) - Z(\gamma) + o_p(Z(\gamma^*) - Z(\gamma))}{2\vartheta^*_\gamma} = \frac{Z(\gamma^*) - Z(\gamma)}{2\vartheta^*_\gamma}(1 + o_p(1)). \tag{19}$$

The term $(Z(\gamma^*) - Z(\gamma))/\vartheta^*_\gamma$ is the LR test statistic for fixed $(\vartheta^*_\gamma, \alpha^*_\gamma)$ comparing γ and $\gamma^* \subset \gamma$.

When $k = 2$ this is a quantile regression LR test statistic, which Koenker and Bassett (1982) showed to be asymptotically $\chi^2_{p_\gamma - p_{\gamma^*}}$ (after rescaling by a constant) precisely under our Conditions A2-A3. When $k = 1$, $(Z(\gamma^*) - Z(\gamma))/\vartheta^*_\gamma$ is the LR test statistic for a weighted least squares problem regressing $\tilde{y} = W_{\theta^*,\alpha^*}y$ on $\widetilde{X} = W_{\theta^*,\alpha^*}X$, which can be shown to be $O_p(1)$ under the conditions in Proposition 4. Briefly, as usual for any γ the total sum of squares can be decomposed as $\tilde{y}^T\tilde{y} = \widehat{\theta}^T_\gamma \widetilde{X}^T_\gamma \widetilde{X}_\gamma \widehat{\theta}_\gamma + (\tilde{y} - \widetilde{X}_\gamma\widehat{\theta}_\gamma)^T(\tilde{y} - \widetilde{X}_\gamma\widehat{\theta}_\gamma)$, hence $Z(\gamma^*) - Z(\gamma) =$

$$(\tilde{y} - \widetilde{X}_{\gamma^*}\widehat{\theta}_{\gamma^*})^T(\tilde{y} - \widetilde{X}_{\gamma^*}\widehat{\theta}_{\gamma^*}) - (\tilde{y} - \widetilde{X}_\gamma\widehat{\theta}_\gamma)^T(\tilde{y} - \widetilde{X}_\gamma\widehat{\theta}_\gamma) = \widehat{\theta}^T_\gamma \widetilde{X}^T_\gamma \widetilde{X}_\gamma \widehat{\theta}_\gamma - \widehat{\theta}^T_{\gamma^*} \widetilde{X}^T_{\gamma^*} \widetilde{X}_{\gamma^*} \widehat{\theta}_{\gamma^*}. \tag{20}$$

Without loss of generality let $\widetilde{X}_\gamma = (\widetilde{X}_{\gamma^*}, \widetilde{X}_{\gamma\setminus\gamma^*})$, where $\widetilde{X}_{\gamma\setminus\gamma^*}$ are the columns in $\widetilde{X}_\gamma$ not contained in $\widetilde{X}_{\gamma^*}$. Let $R = (I - \widetilde{X}_{\gamma^*}(\widetilde{X}^T_{\gamma^*}\widetilde{X}_{\gamma^*})^{-1}\widetilde{X}_{\gamma^*})\widetilde{X}_{\gamma\setminus\gamma^*}$ be orthogonal to the projection of $\widetilde{X}_\gamma$ onto $\widetilde{X}_{\gamma^*}$, then clearly $\widetilde{X}^T_{\gamma^*}R = 0$ and $(\widetilde{X}_{\gamma^*}, R)$ span the column space of $\widetilde{X}_\gamma$. Hence $\widehat{\theta}^T_\gamma \widetilde{X}^T_\gamma \widetilde{X}_\gamma \widehat{\theta}_\gamma = \widehat{\theta}^T_{\gamma^*} \widetilde{X}^T_{\gamma^*} \widetilde{X}_{\gamma^*} \widehat{\theta}_{\gamma^*} + \widehat{\theta}^T_R R^T R \widehat{\theta}_R$, where $\widehat{\theta}_R = (R^TR)^{-1}R^Ty$, giving that $Z(\gamma^*) - Z(\gamma) = \widehat{\theta}^T_R R^T R \widehat{\theta}_R$. By Proposition 4, $\sqrt{n}\widehat{\theta}_R \xrightarrow{D} N(0, \vartheta^*_\gamma V)$ for a fixed positive-definite matrix V.

To conclude, our Conditions A3-A4 guarantee $\frac{1}{n}R^TR \xrightarrow{P} \Sigma_R$ for some fixed Σ_R and by the continuous mapping theorem $\sqrt{n}\Sigma^{\frac{1}{2}}_R\widehat{\theta}_R \xrightarrow{D} N(0, \vartheta^*_\gamma\Sigma^{\frac{1}{2}}_R V \Sigma^{\frac{1}{2}}_R)$. Hence $\frac{n}{\vartheta^*_\gamma}\widehat{\theta}^T_R\Sigma_R\widehat{\theta}_R \xrightarrow{D} Q$, where $Q = O_p(1)$ is a sum of re-scaled central chi-square random variables with 1 degree of freedom. By Slutsky's theorem $\frac{Z(\gamma^*)-Z(\gamma)}{\vartheta^*_\gamma} = \frac{n}{\vartheta^*_\gamma}\widehat{\theta}^T_R(\frac{1}{n}R^TR)\widehat{\theta}_R \xrightarrow{D} Q$, as we wished to prove.

2.7 Proof of Corollary 2

The proof runs analogous to Rossell and Telesca (2017), Proposition 3(ii). Briefly, the BMA estimate is $E(\theta_i \mid y) =$

$$E(\theta_i \mid \gamma^*, y)p(\gamma^* \mid y) + \sum_{\gamma^* \subset \gamma} E(\theta_i \mid \gamma, y)p(\gamma \mid y) + \sum_{\gamma^* \not\subset \gamma} E(\theta_i \mid \gamma, y)p(\gamma \mid y). \tag{21}$$

Suppose that $\theta_i^* \neq 0$. From Proposition 4, the difference between the MLE under γ and θ_i^* is $O_p(1/\sqrt{n})$, and it can be shown that the difference between a Laplace approximation to $E(\theta_i \mid \gamma, y)$ and the MLE is $O_p(1/\sqrt{n})$ hence $E(\theta_i \mid \gamma, y) - \theta_i^* = O_p(1/\sqrt{n})$. Since $p(\gamma^* \mid y) \xrightarrow{P} 1$ by Proposition 5, we have that $E(\theta_i \mid \gamma^*, y)p(\gamma^* \mid y) = \theta_i^* + O_p(1/\sqrt{n})$. If $\theta_i^* = 0$ then by definition $E(\theta_i \mid \gamma^*, y)p(\gamma^* \mid y) = 0$.

Consider the second term in (21) where $\gamma^* \subset \gamma$,

$$p(\gamma \mid y) \leq 1/(1 + B_{\gamma^*,\gamma}p(\gamma^*)/p(\gamma)) < B_{\gamma,\gamma^*}p(\gamma)/p(\gamma^*) = O_p(b_n^{(k)})p(\gamma)/p(\gamma^*) \leq O_p(b_n^{(k)})r^+,$$

where $B_{\gamma^*,\gamma}$ is the Bayes factor between γ^* and γ. From Proposition 5, we have that $b_n^{(k)} = n^{-(p_\gamma - p_{\gamma^*})/2}$ for a local prior, $b_n^{(k)} = n^{-3(p_\gamma - p_{\gamma^*})/2}$ for the pMOM prior, and $b_n^{(k)} = e^{-c\sqrt{n}}$, for some $c > 0$, for the peMOM and piMOM priors. Also, $E(\theta_i \mid \gamma, y) = \theta_i^* + O_p(1/\sqrt{n})$. Therefore, if $\theta_i^* \neq 0$, we have $E(\theta_i \mid \gamma, y)p(\gamma \mid y) = O_p(b_n^{(k)})r^+$. If $\theta_i^* = 0$, then $E(\theta_i \mid \gamma, y)p(\gamma \mid y) = O_p(b_n^{(k)}/\sqrt{n})p(\gamma)/p(\gamma^*) \leq O_p(b_n^{(k)}/\sqrt{n})r^+$. The case for $\gamma^* \not\subset \gamma$ proceeds similarly by noting that by Proposition 5 we have $B_{\gamma,\gamma^*}r^- = O_p(e^{-cn})r^- = O_p(b_n^{(k)})$ for some $c > 0$, since $e^{-cn}r^- = O(b_n^{(k)})$ by assumption.

Combining the previous results it follows that, if $\theta_i^* \neq 0$, then

$$E(\theta_i \mid y) = \theta_i^* + O_p(1/\sqrt{n}) + O_p(b_n^{(k)})r^+ = \theta_i^* + O_p(1/\sqrt{n}), \tag{22}$$

since $b_n^{(k)}r^+ = O_p(1/\sqrt{n})$ by the assumption that r^+ does not increase with n. Conversely if $\theta_i^* = 0$, then

$$E(\theta_i \mid y) = O_p(b_n^{(k)}/\sqrt{n})r^+, \tag{23}$$

giving the desired result.

3. APPROXIMATIONS TO THE INTEGRATED LIKELIHOOD

For ease of notation, we drop the subindex k denoting the set of active variables and let $\theta = (\theta_1, \ldots, \theta_{|k|})$ be their coefficients. Both the Laplace and Importance Sampling approximations require maximizing and evaluating the hessian of $h_l(\theta, \vartheta, \tilde{\alpha}) = \log L(\theta, \tilde{\vartheta}, \tilde{\alpha}) + \log p(\theta, \tilde{\vartheta}, \tilde{\alpha})$, where $L(\cdot)$ and $p(\cdot)$ are the appropriate likelihood and prior density. Denote by $g_l(\theta, \tilde{\vartheta}, \tilde{\alpha})$ the gradient of $h_l(\cdot)$ and by $H_l(\theta, \tilde{\vartheta}, \tilde{\alpha})$ its hessian, Algorithm 1 finds the posterior mode.

Algorithm 1. Posterior mode via Newton-Raphson

1. *Initialize* $(\theta^{(0)}, \tilde{\vartheta}^{(0)}, \tilde{\alpha}^{(0)}) = (\widehat{\theta}, \log(\widehat{\vartheta}), atanh(\widehat{\alpha}))$ *where* $(\widehat{\theta}, \widehat{\vartheta}, \widehat{\alpha})$ *is the MLE given by Algorithm 1. Set* $t = 1$ *and repeat Steps 2-3 until* e *is below some small tolerance (default* 10^{-5}*).*

2. *Update* $(\theta^{(t)}, \tilde{\vartheta}^{(t)}, \tilde{\alpha}^{(t)}) =$

$$(\theta^{(t-1)}, \tilde{\vartheta}^{(t-1)}, \tilde{\alpha}^{(t-1)}) - H_l^{-1}(\theta^{(t-1)}, \tilde{\vartheta}^{(t-1)}, \tilde{\alpha}^{(t-1)}) g_l(\theta^{(t-1)}, \tilde{\vartheta}^{(t-1)}, \tilde{\alpha}^{(t-1)}).$$

3. *Compute* $e = ||(\theta^{(t)}, \tilde{\vartheta}^{(t)}, \tilde{\alpha}^{(t)}) - (\theta^{(t-1)}, \tilde{\vartheta}^{(t-1)}, \tilde{\alpha}^{(t-1)})||^{\infty}$ *where* $||\mathbf{z}||^{\infty}$ *is the largest element of* $\mathbf{z}$ *in absolute value. Set* $t = t + 1$*.*

As usual, in the event that $(\theta^{(t)}, \tilde{\vartheta}^{(t)}, \tilde{\alpha}^{(t)})$ does not increase $h_l(\cdot)$, Step 2 can be adjusted by adding a constant λ to the diagonal of $H_l(\cdot)$, which for large λ gives the direction of the gradient and is guaranteed to decrease $h_l(\cdot)$. However, we observed that this is extremely rare in practice. Usually, the simple Newton step increases $h_l(\cdot)$ at each iteration and converges to the maximum in a few iterations.

Both $g_l(\cdot)$ and $H_l(\cdot)$ are the sum of a term coming from the log-likelihood plus a term coming from the log-prior density. The exact expressions are given below separately.

As an alternative to Algorithm 1, we also provide Algorithm 2 based on Coordinate Descent (*i.e.* successive univariate optimization). Note that the Newton steps to update θ_j and α are in the direction of the gradient and are hence guaranteed to increase the objective function for small enough λ. Step 2 takes advantage of the fact that the maximizer with respect to $\tilde{\vartheta}$ for fixed (θ, α) is available in closed form.

Algorithm 2. Posterior mode via CDA

1. *Initialize* $\theta^{(0)}$ *to the least squares estimate,* $\alpha^{(0)} = 0$, $t = 0$.

2. *For the MOM prior set* $\tilde{\vartheta}^{(t)} = \log\left(s/(n + p + 3a_\vartheta)\right)$, *where*

$$s = \left(b_\vartheta + \theta^{(t)^T}\theta^{(t)} + \sum_{i \in A(\theta)} \frac{(y_i - x_i^T\theta^{(t)})^2}{(1+\alpha^{(t)})^2} + \sum_{i \notin A(\theta)} \frac{(y_i - x_i^T\theta^{(t)})^2}{(1-\alpha^{(t)})^2} \right).$$

   For eMOM and iMOM use a Newton-Raphson step.

3. *For* $j = 1, \ldots, p$

   (a) *Set* $\lambda = 1$ *and* $\theta^* = \theta_j^{(t-1)} - \lambda g^*/h^*$, *where* g^* *and* h^* *are the first and second derivatives of* $f(\theta_j) = \log L_1(\theta_1^{(t-1)}, \ldots, \theta_{j-1}^{(t-1)}, \theta_j, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}, \vartheta^{(t)}, \alpha) + \log p(\theta_j \mid \vartheta)$ *evaluated at* $\theta_j = \theta_j^{(t-1)}$.

   (b) *If* $f(\theta^*) > f(\theta_j^{(t-1)})$ *set* $\theta_j^{(t)} = \theta^*$, *else set* $\lambda = 0.5\lambda$ *and repeat Step 3-(1).*

4. *Let* $\tilde{\alpha}^* = \tilde{\alpha}^{(t-1)} - \lambda g^*/h^*$, *where* g^* *and* h^* *are the first and second derivatives of* $f(\tilde{\alpha}) = \log L_1(\theta^{(t)}, \vartheta^{(t)}, \tilde{\alpha}) + \log p(\tilde{\alpha})$ *at* $\tilde{\alpha} = \tilde{\alpha}^{(t-1)}$. *If* $f(\alpha^*) > f(\alpha^{(t-1)})$ *set* $\alpha^{(t)} = \alpha^*$, *else set* $\lambda = 0.5\lambda$ *and repeat Step 4.*

5. *Compute* $e = \max|(\theta^{(t)}, \tilde{\vartheta}^{(t)}, \tilde{\alpha}^{(t)}) - (\theta^{(t-1)}, \tilde{\vartheta}^{(t-1)}, \tilde{\alpha}^{(t-1)})|$. *If* $e < 10^{-5}$ *stop, else set* $t = t+1$ *and go back to Step 1.*

3.1 Derivatives of the log-likelihood

Two-piece Normal Under the re-parameterization $\tilde{\vartheta} = \log(\vartheta)$, $\tilde{\alpha} = \text{atanh}(\alpha)$ the two-piece Normal log-likelihood (3) has gradient

$$\begin{pmatrix} \frac{1}{\exp(\tilde{\vartheta})} X^T W (y - X\theta) \\ -\frac{n}{2} + \frac{1}{2\exp(\tilde{\vartheta})} (y - X\theta)^T W (y - X\theta) \\ \frac{1}{2\exp(\tilde{\vartheta})} (y - X\theta)^T W^\star (y - X\theta) \end{pmatrix},$$

where as usual $W = \text{diag}(w)$, $w_i = [1+\tanh(\tilde{\alpha})]^{-2}$ if $i \in A(\theta)$ and $w_i = [1-\tanh(\tilde{\alpha})]^{-2}$ if $i \notin A(\theta)$, and $W^\star = \text{diag}(w^*)$ with $w_i^\star = -\frac{2\text{sech}^2(\tilde{\alpha})}{(\tanh(\tilde{\alpha})+1)^3}$ if $i \in A(\theta)$ and $w_i^\star = \frac{2\text{sech}^2(\tilde{\alpha})}{(1-\tanh(\tilde{\alpha}))^3}$ if $i \notin A(\theta)$. Its Hessian is given by

$$-e^{-\tilde{\vartheta}} \begin{pmatrix} X^TWX & X^TW(y-X\theta) & X^TW^\star(y-X\theta) \\ & \frac{1}{2}(y-X\theta)^TW(y-X\theta) & -\frac{1}{2}(y-X\theta)^TW^\star(y-X\theta) \\ & & \frac{1}{2}(y-X\theta)^TW^{\star\star}(y-X\theta) \end{pmatrix}, \tag{24}$$

where $W^{\star\star} = \text{diag}(w^{\star\star})$, with $w_i^{\star\star} = 2e^{-4\tilde{\alpha}}\left(e^{2\tilde{\alpha}}+2\right)$ if $i \in A(\theta)$ and $w_i^{\star\star} = 2e^{2\tilde{\alpha}} + 4e^{4\tilde{\alpha}}$ if $i \notin A(\theta)$.

Two-piece Laplace The asymmetric Laplace $\log L_2(\theta, \tilde{\vartheta}, \tilde{\alpha})$, where $\tilde{\vartheta} = \log(\vartheta)$, $\tilde{\alpha} = \text{atanh}(\alpha)$ has gradient

$$\begin{pmatrix} -e^{-\tilde{\vartheta}/2}X^T\overline{w} \\ -\frac{n}{2} + \frac{1}{2}e^{-\tilde{\vartheta}/2}w^T|y-X\theta| \\ e^{-\tilde{\vartheta}/2}|y-X\theta|^T\overline{w}^* \end{pmatrix},$$

and hessian

$$e^{-\tilde{\vartheta}/2} \times \begin{pmatrix} 0 & \frac{1}{2}X^T\overline{w} & X^Tw^* \\ \frac{1}{2}\overline{w}^TX & -\frac{1}{4}w^T|y-X\theta| & -\frac{1}{2}|y-X\theta|^T\overline{w}^* \\ (X^Tw^*)^T & -\frac{1}{2}|y-X\theta|^T\overline{w}^* & -2|y-X\theta|^Tw^* \end{pmatrix}, \tag{25}$$

where $w_i = \overline{w}_i = (1+\alpha)^{-1}$, $w_i^* = \overline{w}_i^* = e^{-2\alpha}$ if $i \in A(\theta)$, and $w_i = (1-\alpha)^{-1}$, $\overline{w}_i = -w_i$ $w_i^* = e^{2\alpha}$, $\overline{w}_i^* = -w_i^*$ if $i \notin A(\theta)$. Naturally, symmetric Laplace errors are the particular case $\alpha = 0$ and give $w_i = w_i^* = 1$.

Expected two-piece Laplace log-likelihood We derive $\overline{L}_2 = E(\log L_2(\eta))$, where $\eta = (\theta, \vartheta, \alpha)$ and its derivatives under the data-generating model $y_i = x_i^T\theta_0 + \epsilon_i$, for some $\theta_0 \in \mathbb{R}^p$ where ϵ_i are independent across $i = 1, \ldots, n$ and arise from an arbitrary probability density function $s_0(y_i|x_i)$. After some algebra and noting that $\epsilon_i = y_i - x_i^T\theta_0$ gives

$$\overline{L}_2 = \int \log L_2(\eta) s_0(\epsilon|x) d\epsilon = -n\log(2) - \frac{n}{2}\log(\vartheta) \quad - \sum_{i=1}^n \frac{1}{\sqrt{\vartheta}(1+\alpha)} \int_{-\infty}^{x_i^T(\theta-\theta_0)} S_0(\epsilon_i) d\epsilon_i \\ - \sum_{i=1}^n \frac{1}{\sqrt{\vartheta}(1-\alpha)} \int_{x_i^T(\theta-\theta_0)}^{\infty} 1 - S_0(\epsilon_i) d\epsilon_i,$$

where $S_0(\epsilon_i) = S_0(\epsilon_i|0)$ is the cumulative probability function associated to $s_0(\epsilon_i) = s_0(\epsilon_i|0)$, where 0 indicates a zero covariate vector. Then taking derivatives we obtain

$$\begin{aligned}
\frac{\partial}{\partial\theta}\overline{L}_2 &= \sum_{i=1}^n -\frac{x_i S_0(x_i^T(\theta-\theta_0))}{\sqrt{\vartheta}(1+\alpha)} + \frac{x_i[1-S_0(x_i^T(\theta-\theta_0))]}{\sqrt{\vartheta}(1-\alpha)}, \\
\frac{\partial}{\partial\vartheta}\overline{L}_2 &= \sum_{i=1}^n -\frac{1}{2\vartheta} + \frac{I_{i1}}{2\vartheta^{3/2}(1+\alpha)} + \frac{I_{i2}}{2\vartheta^{3/2}(1-\alpha)}, \\
\frac{\partial}{\partial\alpha}\overline{L}_2 &= \sum_{i=1}^n \frac{I_{i1}}{\sqrt{\vartheta}(1+\alpha)^2} - \frac{I_{i2}}{\sqrt{\vartheta}(1-\alpha)^2},
\end{aligned}$$

where $I_{i1} = \int_{-\infty}^{x_i^T(\theta-\theta_0)} S_0(\epsilon_i) d\epsilon_i$, $I_{i2} = \int_{x_i^T(\theta-\theta_0)}^{\infty} 1 - S_0(\epsilon_i) d\epsilon_i$. The second derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\overline{L}_2 &= -\sum_{i=1}^n \frac{2x_i x_i^T s_0(x_i^T(\theta-\theta_0))}{\sqrt{\vartheta}(1-\alpha^2)}, \\
\frac{\partial^2}{\partial\vartheta^2}\overline{L}_2 &= \sum_{i=1}^n \frac{1}{2\vartheta^2} - \frac{3I_{i1}}{4\vartheta^{5/2}(1+\alpha)} - \frac{3I_{i2}}{4\vartheta^{5/2}(1-\alpha)}, \\
\frac{\partial^2}{\partial\alpha^2}\overline{L}_2 &= -\sum_{i=1}^n \frac{2I_{i1}}{\sqrt{\vartheta}(1+\alpha)^3} - \frac{2I_{i2}}{\sqrt{\vartheta}(1-\alpha)^3}, \\
\frac{\partial^2}{\partial\vartheta\partial\theta}\overline{L}_2 &= \sum_{i=1}^n \frac{x_i S_0(x_i^T(\theta-\theta_0))}{2\vartheta^{3/2}(1+\alpha)} - \frac{x_i[1-S_0(x_i^T(\theta-\theta_0))]}{2\vartheta^{3/2}(1-\alpha)}, \\
\frac{\partial^2}{\partial\alpha\partial\theta}\overline{L}_2 &= \sum_{i=1}^n \frac{x_i S_0(x_i^T(\theta-\theta_0))}{\sqrt{\vartheta}(1+\alpha)^2} + \frac{x_i[1-S_0(x_i^T(\theta-\theta_0))]}{\sqrt{\vartheta}(1-\alpha)^2}, \\
\frac{\partial^2}{\partial\vartheta\partial\alpha}\overline{L}_2 &= -\sum_{i=1}^n \frac{I_{i1}}{2\vartheta^{3/2}(1+\alpha)^2} + \frac{I_{i2}}{2\vartheta^{3/2}(1-\alpha)^2}.
\end{aligned}$$

Simple inspection reveals that $(\partial/\partial\theta)\overline{L}_2 = 0$ implies $(\partial^2/\partial\theta\partial\vartheta)\overline{L}_2 = 0$, and likewise $(\partial/\partial\alpha)\overline{L}_2 = 0$ implies $(\partial^2/\partial\theta\partial\alpha)\overline{L}_2 = 0$. Since the maximum likelihood estimator $(\widehat{\theta}, \widehat{\vartheta}, \widehat{\alpha})$ converges in prob-

ability to the maximizer of $\overline{L}_2$, these second derivatives evaluated at $(\widehat{\theta}, \widehat{\vartheta}, \widehat{\alpha})$ also converge in probability to 0.

We wish to find an asymptotic expression for the remaining second derivatives evaluated at $(\widehat{\theta}, \widehat{\vartheta}, \widehat{\alpha})$ when the data-generating truth is $\epsilon_i \sim \text{AL}(x_i^T\theta_0, \vartheta_0, \alpha_0)$ for some $(\theta_0, \vartheta_0, \alpha_0)$. Given that $(\widehat{\theta}, \widehat{\vartheta}, \widehat{\alpha}) \xrightarrow{P} (\theta_0, \vartheta_0, \alpha_0)$, the expressions above require evaluating the density of an asymmetric Laplace $s_0(0) = 1/(2\sqrt{\vartheta_0})$ and its cumulative probability function $S_0(0) = (1+\alpha_0)/2$. Similarly, direct integration gives $I_{i1} = \sqrt{\vartheta_0}(1+\alpha_0)^2/2$ and $I_{i2} = \sqrt{\vartheta_0}(1-\alpha_0)^2/2$.

$$
\begin{aligned}
&\frac{\partial^2}{\partial\theta^2}\overline{L}_2 \xrightarrow{P} -X^TX\frac{1}{\vartheta_0(1-\alpha_0^2)},\\
&\frac{\partial^2}{\partial\vartheta^2}\overline{L}_2 \xrightarrow{P} \frac{n}{2\vartheta_0^2} - \frac{3n(1+\alpha_0)}{8\vartheta_0^2} - \frac{3(1-\alpha_0)}{8\vartheta_0^2} = -\frac{n}{4\vartheta_0^2},\\
&\frac{\partial^2}{\partial\alpha^2}\overline{L}_2 \xrightarrow{P} -\frac{n}{1+\alpha_0} - \frac{n}{1-\alpha_0} = -\frac{2n}{1-\alpha_0^2},\\
&\frac{\partial^2}{\partial\alpha\partial\theta}\overline{L}_2 \xrightarrow{P} \frac{n\overline{x}}{\sqrt{\vartheta_0}}\left(\frac{1}{2(1+\alpha_0)} + \frac{1}{2(1-\alpha_0)}\right) = \frac{n\overline{x}}{\sqrt{\vartheta_0}(1-\alpha_0^2)}.
\end{aligned}
\tag{26}
$$

3.2 Derivatives of the log-prior density

The log-prior density is $\log p(\theta, \tilde{\vartheta}) = \log p(\theta \mid \tilde{\vartheta}) + \log p(\tilde{\vartheta})$ when $\tilde{\alpha} = 0$ under the assumed model and $\log p(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \log p(\theta, \tilde{\vartheta}) + \log p(\tilde{\alpha})$ when $\tilde{\alpha} \neq 0$, where $p(\theta \mid \tilde{\vartheta})$ and $p(\tilde{\alpha})$ are the pMOM, piMOM or peMOM priors and $p(\tilde{\vartheta}) = \text{IG}(e^{\tilde{\vartheta}}; a_\vartheta/2, b_\vartheta/2)e^{\vartheta}$. For ease of notation let θ^{-a} be the vector with elements θ_i^{-a} for $i = 1, \ldots, |k|$.

pMOM prior Straightforward algebra gives

$$\nabla \log p_M(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \begin{pmatrix} 2\theta^{-1} - \theta e^{-\tilde{\vartheta}}/g_\theta \\ -\frac{3|k|+a_\vartheta}{2} + (\theta^T\theta/g_\theta + b_\vartheta)e^{-\tilde{\vartheta}}/2 \\ 2\tilde{\alpha}^{-1} - \tilde{\alpha} g_\alpha^{-1} \end{pmatrix},$$

$$\nabla^2 \log p_M(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \begin{pmatrix} \mathrm{diag}(-2\theta^{-2} - e^{-\tilde{\vartheta}}/g_\theta) & \theta e^{-\tilde{\vartheta}}/g_\theta & 0 \\ \theta^T e^{-\tilde{\vartheta}}/g_\theta & -e^{-\tilde{\vartheta}}(\theta^T\theta/g_\theta + b_\vartheta)/2 & 0 \\ 0 & 0 & -2\tilde{\alpha}^{-2} - g_\alpha^{-1} \end{pmatrix},$$

piMOM prior We obtain

$$\nabla \log p_I(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \begin{pmatrix} -2\theta^{-1} + 2g_\theta e^{\tilde{\vartheta}}\theta^{-3} \\ (|k| - a_\vartheta)/2 + b_\vartheta e^{-\tilde{\vartheta}}/2 - g_\theta e^{\tilde{\vartheta}} \sum_i \theta_i^{-2} \\ -2\tilde{\alpha}^{-1} - 2g_\alpha \tilde{\alpha}^{-3} \end{pmatrix},$$

$$\nabla^2 \log p_I(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \begin{pmatrix} \mathrm{diag}(2\theta^{-2} - 6g_\theta e^{\tilde{\vartheta}}\theta^{-4}) & 2g_\theta e^{\tilde{\vartheta}}\theta^{-3} & 0 \\ (-2g_\theta e^{\tilde{\vartheta}}\theta^{-3})^T & -b_\vartheta e^{-\tilde{\vartheta}}/2 - e^{\tilde{\vartheta}} g_\theta \sum_i \theta_i^{-2} & 0 \\ 0 & 0 & 2\tilde{\alpha}^{-2} + 6g_\alpha \tilde{\alpha}^{-4} \end{pmatrix}.$$

peMOM prior We obtain

$$\nabla \log p_E(\theta, \tilde{\vartheta}, \tilde{\alpha}) = \begin{pmatrix} 2g_\theta e^{\tilde{\vartheta}}\theta^{-3} - \theta e^{-\tilde{\vartheta}} g_\theta^{-1} \\ -(|k| + a_\vartheta)/2 + (b_\vartheta + \theta^T\theta/g_\theta)e^{-\tilde{\vartheta}}/2 - g_\theta e^{\tilde{\vartheta}} \sum_i \theta_i^{-2} \\ 2g_\alpha \tilde{\alpha}^{-3} - \tilde{\alpha} g_\alpha^{-1} \end{pmatrix},$$

and $\nabla^2 \log p_E(\theta, \tilde{\vartheta}, \tilde{\alpha}) =$

$$\begin{pmatrix} \mathrm{diag}(-6g_\theta e^{\tilde{\vartheta}}\theta^{-4} - e^{-\tilde{\vartheta}} g_\theta^{-1}) & 2g_\theta e^{\tilde{\vartheta}}\theta^{-3} + \theta e^{-\tilde{\vartheta}} g_\theta^{-1} & 0 \\ (2g_\theta e^{\tilde{\vartheta}}\theta^{-3} + \theta e^{-\tilde{\vartheta}} g_\theta^{-1})^T & -(b_\vartheta + \theta^T\theta/g_\theta)e^{-\tilde{\vartheta}}/2 - e^{\tilde{\vartheta}} g_\theta \sum_i \theta_i^{-2} & 0 \\ 0 & 0 & -6g_\alpha \tilde{\alpha}^{-4} - g_\alpha^{-1} \end{pmatrix}.$$

3.3 Quadratic approximation to asymmetric Laplace log-likelihood

The goal is to approximate the curvature of the one-dimensional function $f(\lambda) = \log L_2(\theta_\lambda, \widehat{\vartheta}, \widehat{\alpha})$ around $\lambda = 0$, where $\theta_\lambda = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{j-1}, \widehat{\theta}_j + \lambda, \widehat{\theta}_{j+1}, \ldots, \widehat{\theta}_p)$ is fixed to the maximum likelihood estimator except for the j^{th} regression parameter, which is a function of $\lambda \in \mathbb{R}$. Given that $f(0)$ is known and that its derivative at $\lambda = 0$ is 0 ($\widehat{\theta}$ is a maximum) we seek $h_j^* < 0$ such that $f(\lambda) - f(0) \approx 0.5 h_j^* \lambda^2$. Our strategy is to evaluate $f(\lambda_k)$ on a grid λ_k for $k = 1, \ldots, K$ and use the least-squares estimate $h_j^* = 2 \sum_{k=1}^K \lambda_k^2 (f(\lambda_k) - f(0)) / \sum_{k=1}^K \lambda_k^4$, where the form of $\log L_2$ gives the simple expression

$$f(\lambda_k) - f(0) = -\frac{1}{\sqrt{\widehat{\vartheta}}} \sum_{i=1}^n |r_i - \lambda_k x_{ij}| \left(\frac{\mathrm{I}(r_i \leq \lambda_k x_{ij})}{1 + \widehat{\alpha}} + \frac{\mathrm{I}(r_i > \lambda_k x_{ij})}{1 - \widehat{\alpha}} \right),$$

and $r_i = y_i - x_i^T \widehat{\theta}$. Once $h_1^*, \ldots, h_p^*$ have been obtained we let $D = \mathrm{diag}(h_1^*/\bar{h}_{11}, \ldots, h_p^*/\bar{h}_{pp})$ where $\overline{H} = (X^T X)/(\widehat{\vartheta}(1 - \widehat{\alpha}^2))$ is the asymptotic hessian under asymmetric Laplace errors, and we approximate the hessian of $\log L_2(\theta, \widehat{\vartheta}, \widehat{\alpha})$ around $\theta = \widehat{\theta}$ with $H^* = D^{\frac{1}{2}} \overline{H} D^{\frac{1}{2}}$. The construction ensures that the diagonal elements in H^* are $h_1^*, \ldots, h_p^*$, i.e. the quadratic approximation matches the actual curvature of $\log L_2$ along each canonical axis. From Section 4 the correlation structure borrowed from $\overline{H}$ remains asymptotically valid as long as the residuals are independent and identically distributed, however in our experience the approximation usually suffices for practical purposes even when these assumptions is violated.

The problem has been thus reduced to choosing the grid $\lambda_1, \ldots, \lambda_K$. One naive option is to take the n points of non-differentiability $\lambda = r_i / x_{ij}$, however, by the nature of least squares, this strategy tends to approximate better $f(\lambda)$ for large λ^2 and we are interested in local approximations around $\lambda = 0$, further evaluating $f(\lambda)$ at n points requires $O(n^2)$ operations for each $j = 1, \ldots, p$ and is thus computationally costly. Instead we evaluate $f(\lambda)$ only at the $K = 2$ points given by the endpoints of the asymptotic 95% confidence interval $\lambda = \{-1.96\overline{v}_j, 1.96\overline{v}_j)$ where $\overline{v}_j$ is the j^{th} diagonal element in $\overline{H}^{-1}$. This simple strategy ensures that the approximation holds locally around $\lambda = 0$ in the sense of having non-negligible likelihood, requires only $O(n)$ operations and we have observed to deliver reasonably accurate approximations in practice. Our approximation is similar in spirit to the rank-based score test inversion used to obtain confidence intervals in quantile

regression, which has been amply described to deliver fairly precise intervals, with the important difference that rank inversion requires an ordering of observations that scales poorly with p and n.

Supplementary Figure 1S shows an example with the likelihood L_2 (scaled to $(0, 1)$) and the two quadratic approximations based on the asymptotic covariance and its least-squares adjustment for an intercept-only model ($p = 1$) and $n = 200$. When residuals were truly generated from an asymmetric Laplace (left panel) the two quadratic approximations were essentially identical, however under truly normally distributed residuals the asymptotic covariance over-estimated the curvature.

[Figure 1 about here.]

4. SUPPLEMENTARY RESULTS

[Table 1 about here.]

[Table 2 about here.]

4.1 Simulation study with identically distributed errors

[Table 3 about here.]

[Table 4 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

[Table 5 about here.]

[Table 6 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

We assessed the sensitivity of the results of the $p = 6$ simulation study in Section 6.1 of the main paper to the prior on the asymmetry coefficient by setting g_α such that $P(|\alpha| > 0.1) = 0.99$. Supplementary Table 3S summarizes the inference on the error distribution and Supplementary Figure 2S the marginal variable inclusion probabilities. The latter were virtually identical to those in Figure 2 obtained under g_α such that $P(|\alpha| > 0.2) = 0.99$, showing that variable inclusion is robust to moderate changes in g_α.

We also assessed the accuracy of the Laplace approximations to the integrated likelihood $p(y \mid \gamma)$ by comparing the results with those obtained with the importance sampling estimates with $B = 10,000$ draws described in Section 5 of the main paper. Supplementary Figure 3S displays the results for $g_\alpha = 0.357$. These are extremely similar to those based on Laplace approximation in Figure 2.

Supplementary Figure 6S shows analogous results for $p = 100$, with $g_\alpha = 0.357$ and $p(y \mid \gamma)$ estimated via Laplace approximations.

4.2 Simulation study with non-identically distributed errors

[Table 7 about here.]

[Table 8 about here.]

[Figure 8 about here.]

[Table 9 about here.]

Supplementary Table 7S shows the mean average posterior probability assigned to the Normal, asymmetric Normal, Laplace and asymmetric Laplace models under the heteroskedastic simulation (Section 6.2, main manuscript).

Supplementary Figure 8S shows marginal variable inclusion probabilities under the hetero-asymmetric simulation.

Supplementary Table 9S reports true and false positives for our simulation study mimicking Grünwald and van Ommen (2014) described in Section 6.2 of the main manuscript.

4.3 DLD data

[Table 10 about here.]

[Table 11 about here.]

Supplementary Table 10S shows the six genes with largest marginal inclusion probabilities $p(\gamma_j = 1 \mid y)$ when conditioning on Normal errors and when inferring the error distribution. The figures were similar for the four top genes, but the Normal model assigned somewhat higher probability to FBXL19 substantially lower probability to MTMR1.

REFERENCES

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.

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.

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.

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

Newey, W. K., and Powell, J. L. (1987), "Asymmetric Least Squares Estimation and Testing," *Econometrica*, 55(4), 819–847.

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

Rockafellar, R. (2015), *Convex analysis*, Princeton: Princeton university press.

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

van der Vaart, A. (1998), *Asymptotic statistics*, New York: Cambridge University Press.

List of Figures

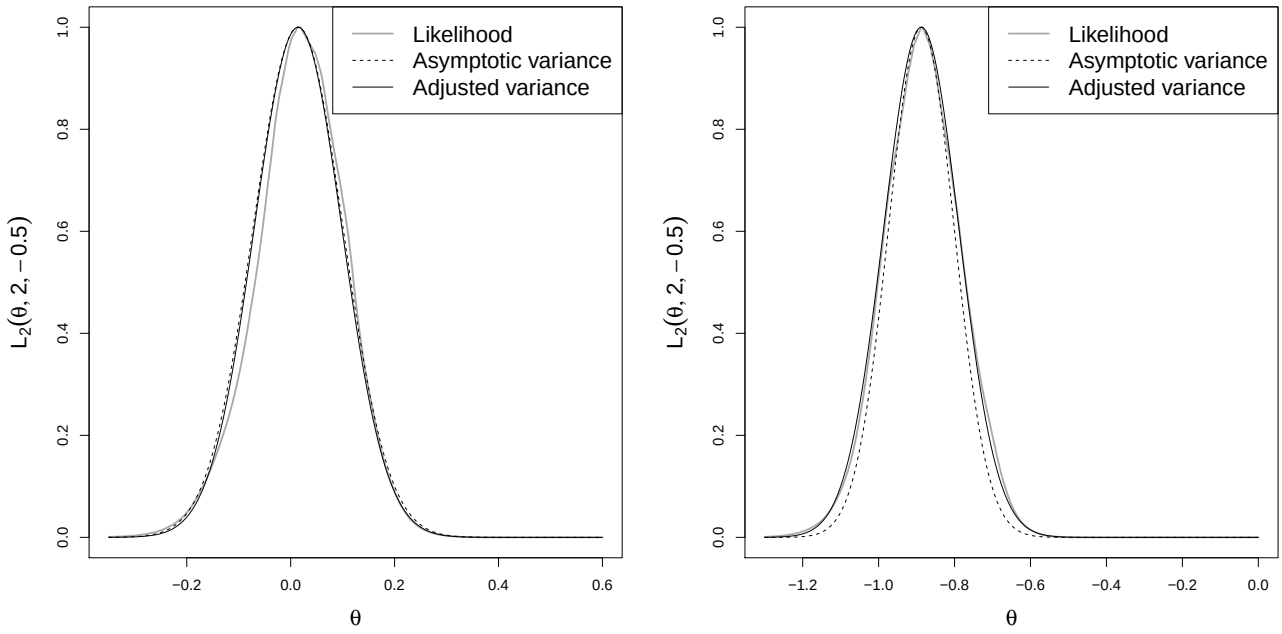

Figure 1S: Quadratic approximation to L_2 (solid grey) with $p = 1, n = 200$ from asymptotic covariance (dotted black) and least-squares adjustment (solid black). Left: $\epsilon_i \sim \text{AL}(0, 2, -0.5)$; Right: $\epsilon_i \sim N(0, 2)$.

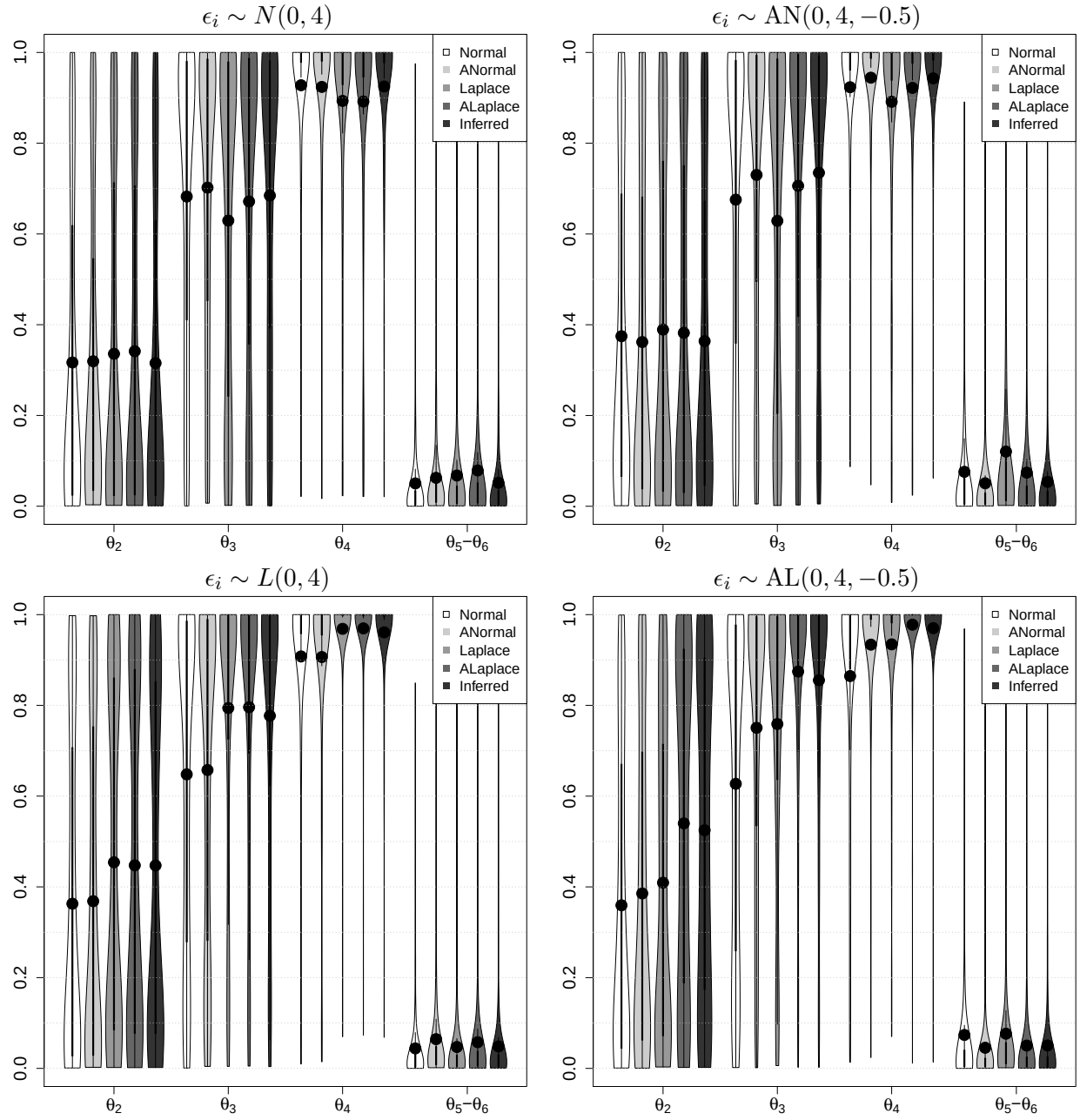

Figure 2S: Sensitivity analysis with $g_\alpha = 0.087$. $P(\theta_i \neq 0 \mid y)$ for $p = 5$, $\vartheta = 2$, $\theta = (0.5, 1, 1.5, 0, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

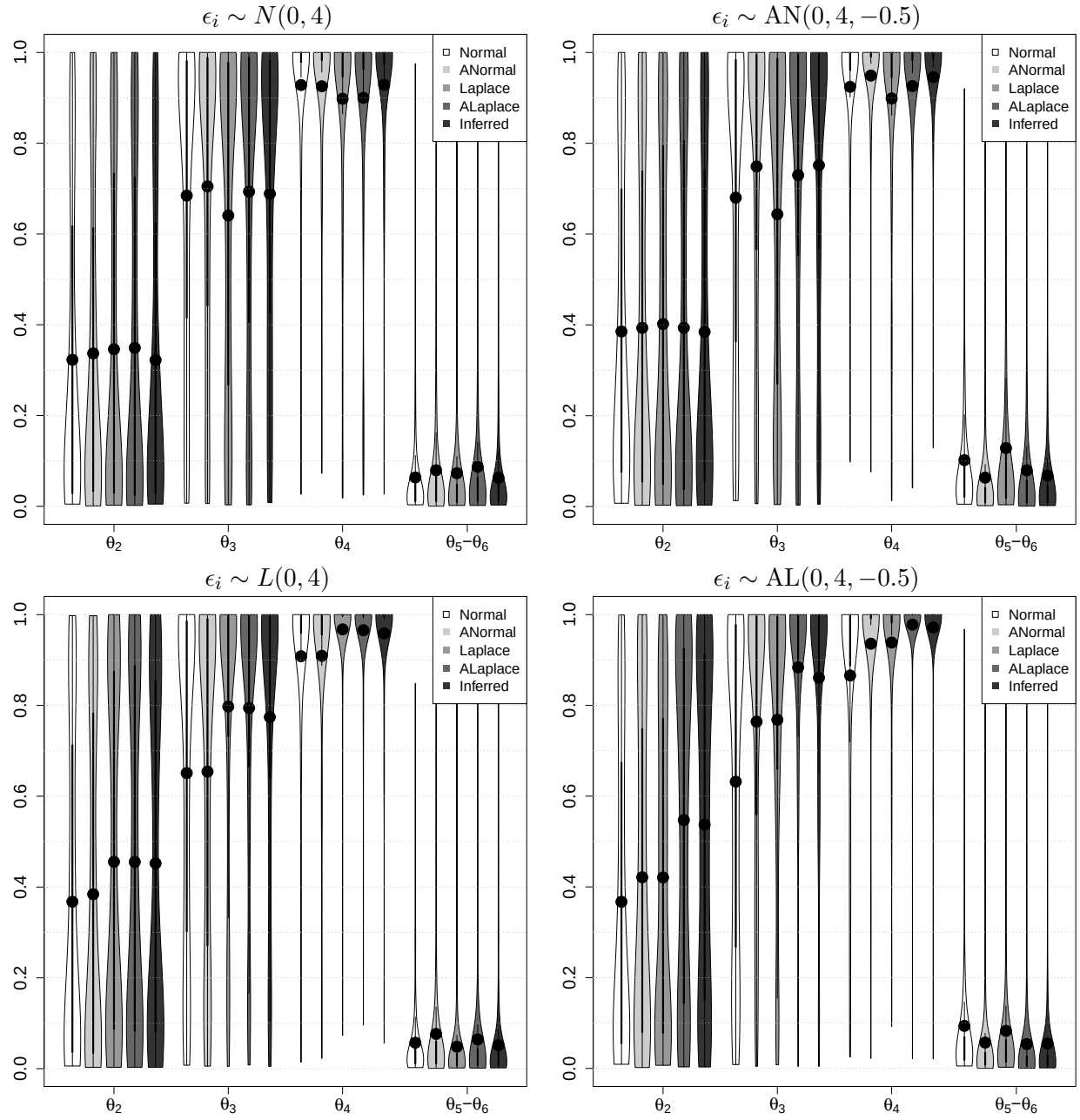

Figure 3S: Monte Carlo estimates ($B = 10,000$) under $g_\alpha = 0.357$. $P(\theta_i \neq 0 \mid y)$ for $p = 5$, $\vartheta = 2$, $\theta = (0.5, 1, 1.5, 0, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

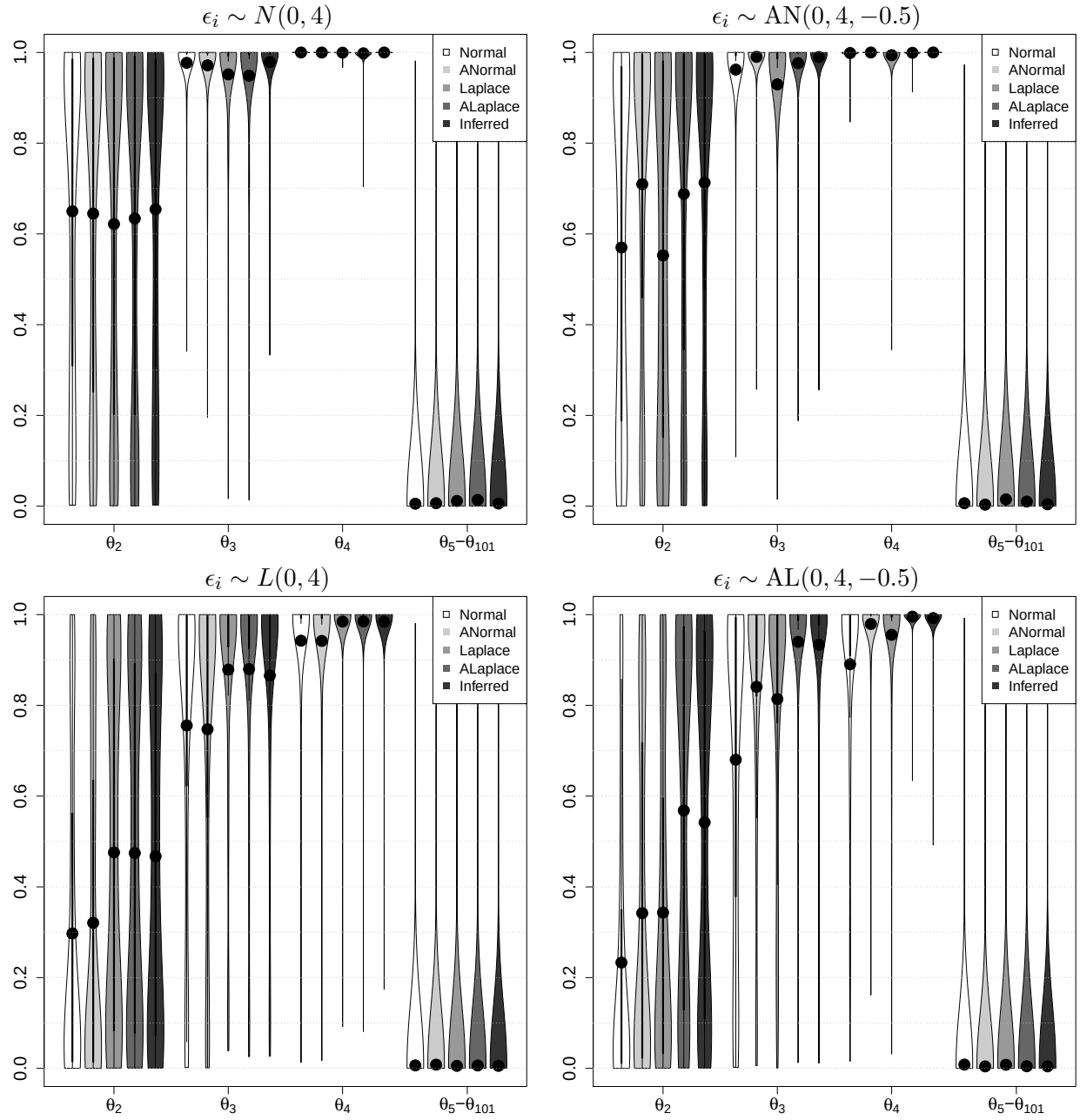

Figure 4S: $P(\theta_i \neq 0 \mid y)$ for $p = 100$, $\vartheta = 1$, $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

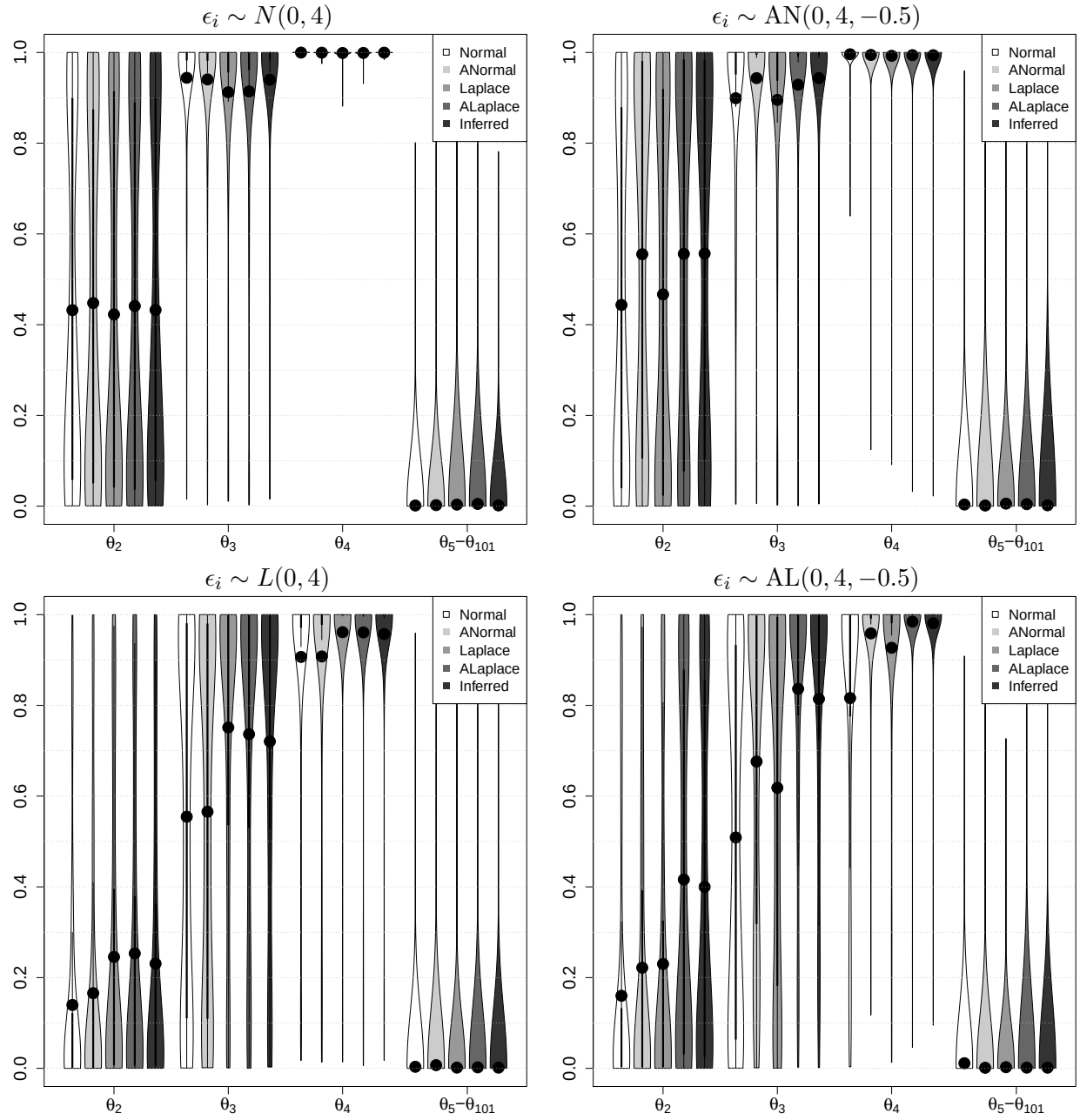

Figure 5S: $P(\theta_i \neq 0 \mid y)$ for $p = 500$, $\vartheta = 1$, $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

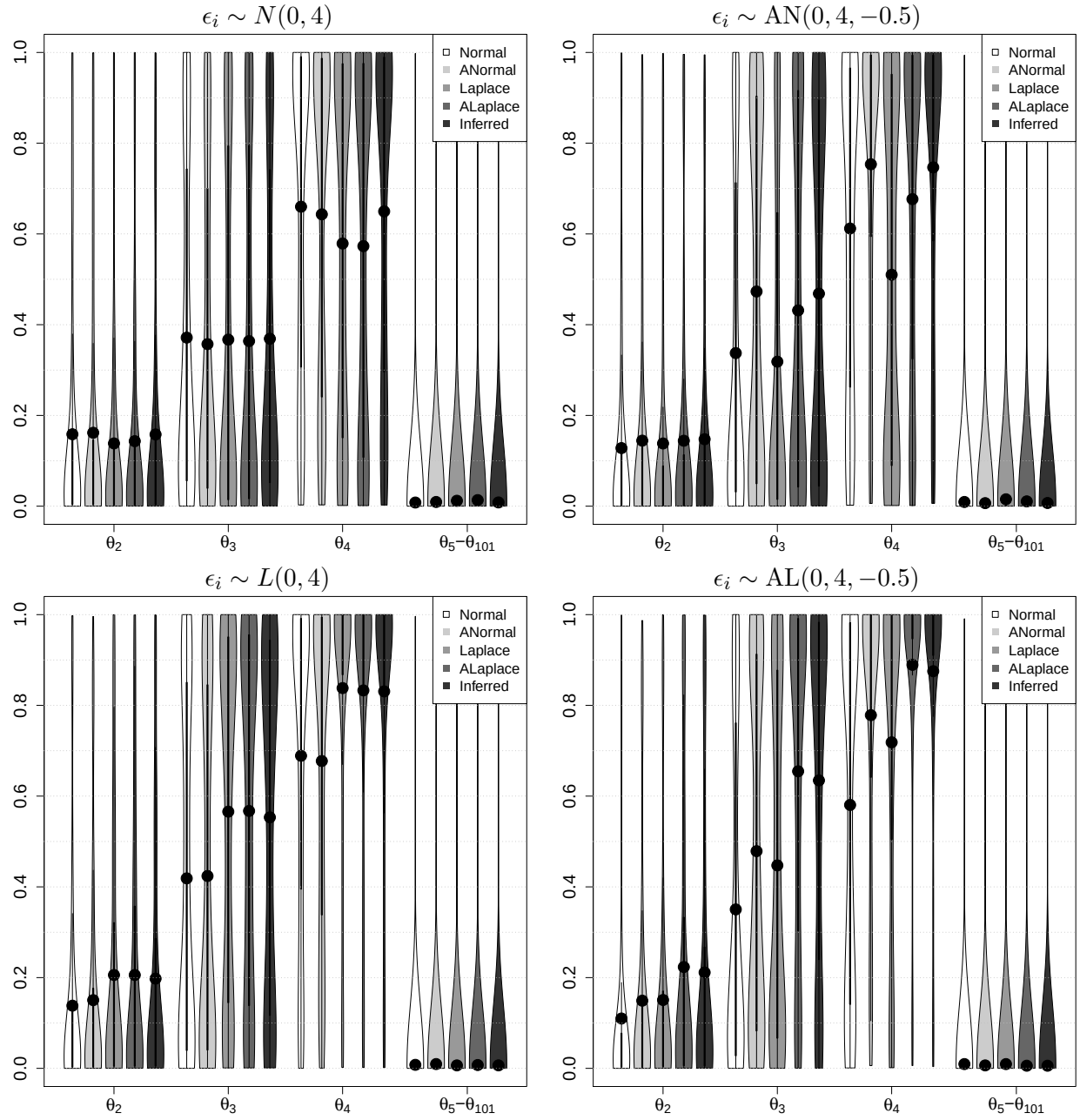

Figure 6S: $P(\theta_i \neq 0 \mid y)$ for $p = 100$, $\vartheta = 2$, $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

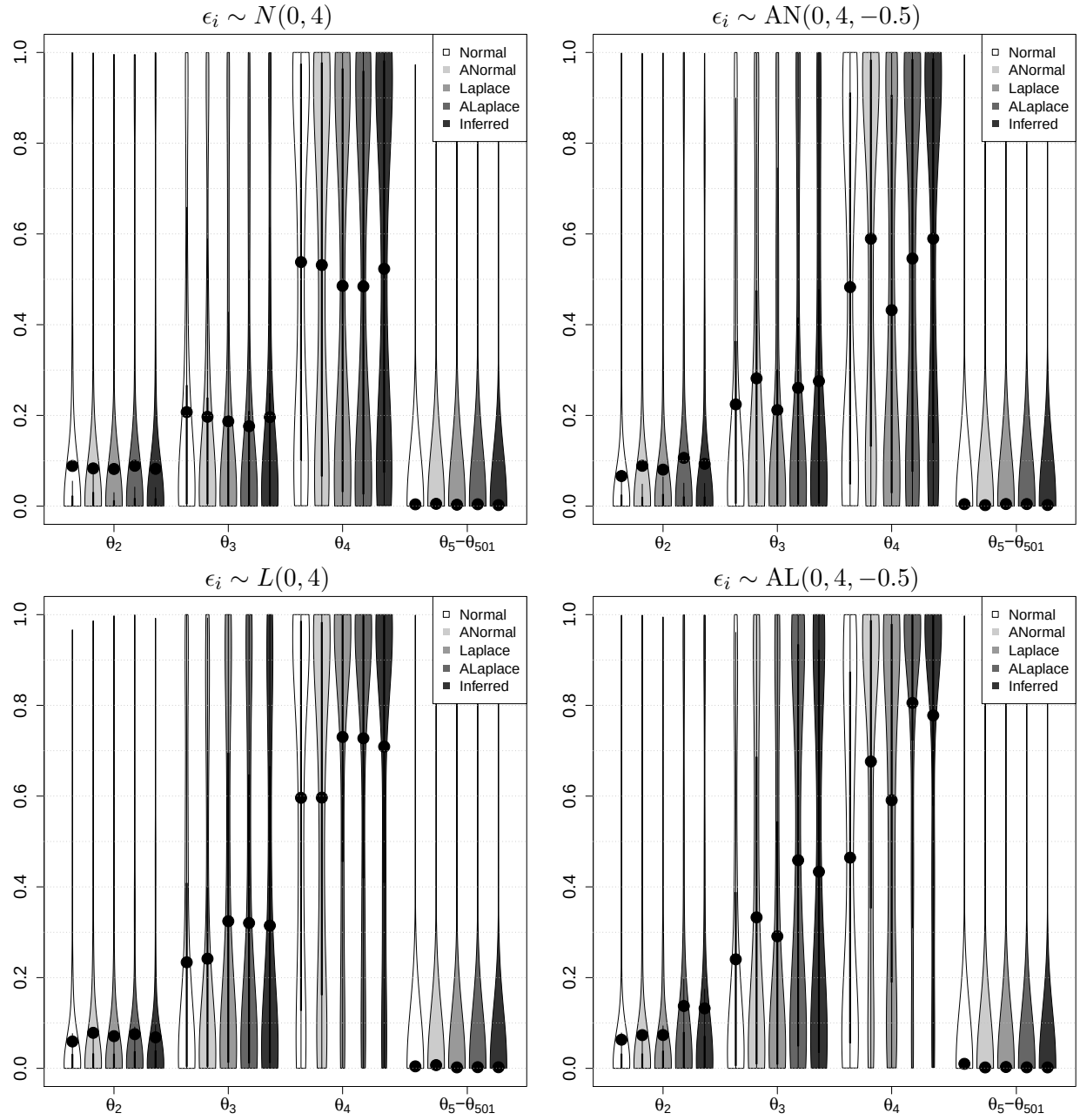

Figure 7S: $P(\theta_i \neq 0 \mid y)$ for $p = 500$, $\vartheta = 2$, $\theta = (0, 0.5, 1, 1.5, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

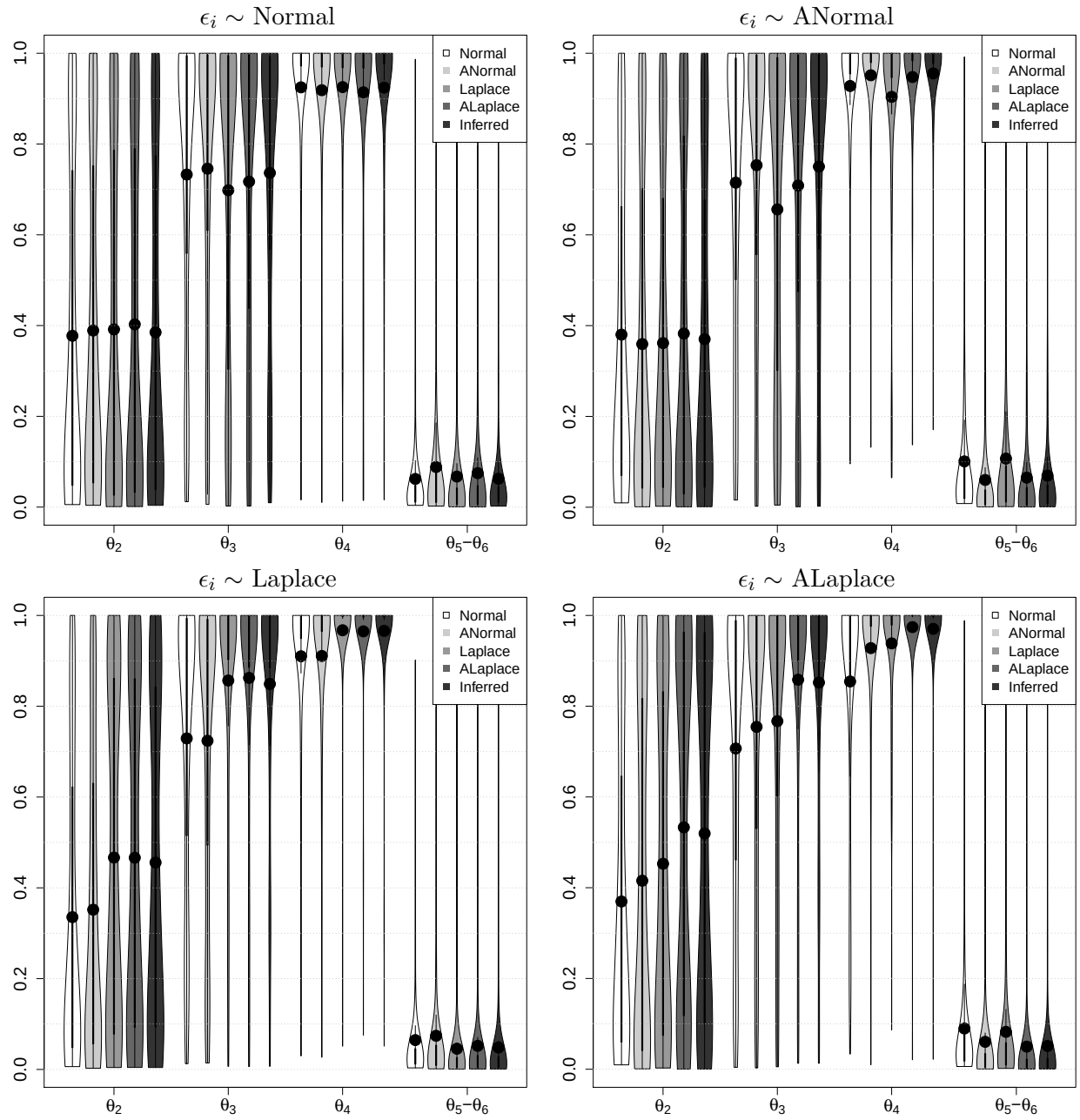

Figure 8S: $P(\theta_i \neq 0 \mid y)$ for simulation with constant $\vartheta = 0$ and varying $\tanh(\alpha_i) \sim N(\text{atanh}(\bar{\alpha}, 1/4^2))$, where $\bar{\alpha} = 0$ for Normal and Laplace and $\bar{\alpha} = -0.5$ for ANormal and ALaplace. $P(\theta_i \neq 0 \mid y)$ for $p = 6$, $\theta = (0, 0.5, 1, 1.5, 0, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Black circles show the mean.

List of Tables

Simulation truth $\epsilon_i \sim \text{AN}(0, 4, \alpha)$				
Fitted model	$\alpha = 0$	$\alpha = -0.25$	$\alpha = -0.5$	$\alpha = -0.75$
Normal	76.99	98.84	103.84	101.25
ANormal	92.08	86.10	102.60	115.64
Laplace	90.58	92.84	97.90	93.13
ALaplace	122.64	121.12	124.69	131.50
Simulation truth $\epsilon_i \sim \text{AL}(0, 4, \alpha)$				
Fitted model	$\alpha = 0$	$\alpha = -0.25$	$\alpha = -0.5$	$\alpha = -0.75$
Normal	76.62	96.05	99.85	97.78
ANormal	81.77	82.74	92.67	104.76
Laplace	90.29	93.42	92.88	91.25
ALaplace	117.30	113.69	115.08	122.79

Table 1S: CPU time (10^{-4} seconds) on 3.4GHz Intel i7, 32Gb RAM, Windows 10. $p = 6$, $\vartheta = 4$, $\theta = (0, 0.5, 0.75, 1, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$.

	Simulation truth			
	$N(0,4)$	AN$(0,4,-0.5)$	$L(0,4)$	AL$(0,4,-0.5)$
Normal	6.9	29.7	6.4	32.9
ANormal	52.9	21.9	41.3	22.2
Laplace	17.0	28.2	14.6	26.6
ALaplace	57.7	26.4	26.7	22.4
Inferred	6.1	22.6	13.5	23.0

Table 2S: CPU time (seconds) on 8GB RAM Mac laptop with 1.6GHz Intel i5 processors running OS X 10.11.6 $p = 100$, $\vartheta = 2$, $\theta = (0, 0.5, 0.75, 1, 0, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$.

Truth	Average $p(\gamma_{p+1}, \gamma_{p+2} \mid y)$			
	$\gamma_{p+1} = \gamma_{p+2} = 0$	$\gamma_{p+1} = 1, \gamma_{p+2} = 0$	$\gamma_{p+1} = 0, \gamma_{p+2} = 1$	$\gamma_{p+1} = \gamma_{p+2} = 0$
	$p = 6$, $g_\alpha = 0.357$, Laplace $p(\gamma \mid y)$			
$N(0, 2)$	0.91	0.02	0.06	0.00
$\text{AN}(0, 2, -0.5)$	0.11	0.81	0.01	0.06
$L(0, 2)$	0.14	0.00	0.84	0.02
$\text{AL}(0, 2, -0.5)$	0.02	0.12	0.01	0.85
	$p = 6$, $g_\alpha = 0.357$, Monte Carlo $p(\gamma \mid y)$			
$N(0, 2)$	0.91	0.02	0.06	0.00
$\text{AN}(0, 2, -0.5)$	0.11	0.81	0.01	0.07
$L(0, 2)$	0.12	0.01	0.85	0.02
$\text{AL}(0, 2, -0.5)$	0.02	0.12	0.01	0.85
	$p = 6$, $g_\alpha = 0.087$, Laplace $p(\gamma \mid y)$			
$N(0, 2)$	0.87	0.07	0.06	0.01
$\text{AN}(0, 2, -0.5)$	0.07	0.86	0.01	0.07
$L(0, 2)$	0.13	0.01	0.79	0.07
$\text{AL}(0, 2, -0.5)$	0.01	0.13	0.01	0.85

Table 3S: Simulation study for $p = 6$. Posterior probability of the 4 error distributions under $\vartheta = 2$, $\theta = (0, 0.5, 1, 1.5, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$.

Truth	Average $p(\gamma_{p+1}, \gamma_{p+2} \mid y)$			
	$\gamma_{p+1} = \gamma_{p+2} = 0$	$\gamma_{p+1} = 1, \gamma_{p+2} = 0$	$\gamma_{p+1} = 0, \gamma_{p+2} = 1$	$\gamma_{p+1} = \gamma_{p+2} = 0$
		$p = 101$, $\vartheta = 1$		
$N(0, 2)$	0.91	0.01	0.08	0.00
$\mathrm{AN}(0, 2, -0.5)$	0.03	0.86	0.00	0.11
$L(0, 2)$	0.15	0.01	0.83	0.02
$\mathrm{AL}(0, 2, -0.5)$	0.00	0.13	0.01	0.86
		$p = 101$, $\vartheta = 2$		
$N(0, 2)$	0.89	0.01	0.10	0.00
$\mathrm{AN}(0, 2, -0.5)$	0.02	0.89	0.00	0.09
$L(0, 2)$	0.15	0.01	0.82	0.02
$\mathrm{AL}(0, 2, -0.5)$	0.00	0.16	0.01	0.83
		$p = 501$, $\vartheta = 1$		
$N(0, 2)$	0.85	0.00	0.14	0.00
$\mathrm{AN}(0, 2, -0.5)$	0.01	0.85	0.01	0.14
$L(0, 2)$	0.18	0.00	0.80	0.02
$\mathrm{AL}(0, 2, -0.5)$	0.00	0.15	0.00	0.84
		$p = 501$, $\vartheta = 2$		
$N(0, 2)$	0.83	0.00	0.16	0.00
$\mathrm{AN}(0, 2, -0.5)$	0.00	0.87	0.00	0.12
$L(0, 2)$	0.19	0.00	0.79	0.01
$\mathrm{AL}(0, 2, -0.5)$	0.00	0.22	0.00	0.77

Table 4S: Simulation study for $p = 101, 501$. Posterior probability of the 4 error distributions under $g_\alpha = 0.357$, $\theta = (0, 0.5, 1, 1.5, \ldots, 0)$, $n = 100$, $\rho_{ij} = 0.5$. Laplace approximation to $p(y \mid \gamma)$ was used.

	$p = 100$				$p = 500$			
	$p(\gamma_0 \mid y)$	$p(\widehat{\gamma} = \gamma_0)$	FP	TP	$p(\gamma_0 \mid y)$	$p(\widehat{\gamma} = \gamma_0)$	FP	TP
	Truly $\epsilon \sim N(0, 1)$							
Normal	0.46	0.63	0.1	2.7	0.26	0.37	0.2	2.4
Two-piece Normal	0.43	0.63	0.2	2.7	0.24	0.38	0.3	2.4
Laplace	0.26	0.42	0.5	2.6	0.12	0.19	0.8	2.3
Two-piece Laplace	0.23	0.39	0.7	2.6	0.12	0.21	0.9	2.3
Inferred	0.45	0.62	0.2	2.7	0.25	0.37	0.2	2.4
LASSO-LS		0.00	12.4	3.0		0.00	20.4	2.9
LASSO-LAD		0.00	10.2	2.9		0.00	18.7	2.6
LASSO-QR		0.00	10.2	2.9		0.00	18.7	2.6
SCAD		0.07	4.2	2.9		0.01	7.3	2.8
	Truly $\epsilon \sim AN(0, 1, -0.5)$							
Normal	0.38	0.55	0.2	2.6	0.21	0.34	0.5	2.4
Two-piece Normal	0.59	0.73	0.1	2.8	0.40	0.55	0.4	2.6
Laplace	0.20	0.35	0.7	2.5	0.07	0.14	1.2	2.4
Two-piece Laplace	0.33	0.48	0.5	2.7	0.18	0.32	1.1	2.5
Inferred	0.57	0.72	0.1	2.8	0.38	0.52	0.4	2.6
LASSO-LS		0.00	12.4	3.0		0.00	21.9	2.9
LASSO-LAD		0.00	9.8	2.8		0.00	18.1	2.6
LASSO-QR		0.00	9.0	2.9		0.00	15.1	2.7
SCAD		0.07	4.0	2.9		0.03	7.3	2.8
	Truly $\epsilon \sim L(0, 1)$							
Normal	0.11	0.14	0.3	2.0	0.03	0.02	0.6	1.6
Two-piece Normal	0.11	0.15	0.3	2.1	0.04	0.04	1.1	1.7
Laplace	0.29	0.38	0.2	2.4	0.13	0.19	0.4	2.0
Two-piece Laplace	0.28	0.35	0.3	2.4	0.12	0.18	0.5	2.0
Inferred	0.28	0.38	0.2	2.4	0.12	0.18	0.4	2.0
LASSO-LS		0.00	11.3	2.8		0.00	21.4	2.5
LASSO-LAD		0.01	9.7	2.8		0.00	17.8	2.5
LASSO-QR		0.01	9.7	2.8		0.00	17.8	2.5
SCAD		0.02	5.0	2.7		0.00	9.0	2.4
	Truly $\epsilon \sim AL(0, -0.5)$							
Normal	0.07	0.10	0.4	1.9	0.02	0.02	1.1	1.5
Two-piece Normal	0.21	0.27	0.2	2.2	0.11	0.15	0.3	2.0
Laplace	0.16	0.19	0.4	2.1	0.05	0.07	0.7	1.8
Two-piece Laplace	0.43	0.51	0.2	2.5	0.27	0.34	0.4	2.3
Inferred	0.41	0.48	0.2	2.5	0.25	0.33	0.4	2.2
LASSO-LS		0.00	11.6	2.8		0.00	20.1	2.5
LASSO-LAD		0.00	9.9	2.7		0.00	17.5	2.3
LASSO-QR		0.00	9.0	2.8		0.00	15.2	2.5
SCAD		0.01	5.2	2.6		0.01	9.4	2.3

Table 5S: Simulation results under $\vartheta = 1$. γ_0: true predictors. $\widehat{\gamma}$: selected variables. CC: number of correctly classified variables ($\sum_{j=1}^{p} \mathrm{I}(\widehat{\gamma}_j = \gamma_{0j})$). FP: number of false positives; TP: number of true positives. LASSO-LAD and LASSO-QR are equivalent when $\alpha = 0$

	$p = 100$				$p = 500$			
	$p(\gamma_0 \mid y)$	$p(\widehat{\gamma} = \gamma_0)$	FP	TP	$p(\gamma_0 \mid y)$	$p(\widehat{\gamma} = \gamma_0)$	FP	TP
	Truly $\epsilon \sim N(0, 1)$							
Normal	0.01	0.01	0.4	1.2	0.00	0.00	0.8	0.9
Two-piece Normal	0.01	0.01	0.5	1.2	0.00	0.00	0.9	0.8
Laplace	0.00	0.00	0.7	1.1	0.00	0.00	1.0	0.8
Two-piece Laplace	0.00	0.01	0.8	1.1	0.00	0.00	1.1	0.8
Inferred	0.01	0.01	0.5	1.2	0.00	0.00	0.7	0.9
LASSO-LS		0.00	11.9	2.5		0.00	18.0	2.0
LASSO-LAD		0.00	8.9	2.0		0.00	15.6	1.4
LASSO-QR		0.00	8.9	2.0		0.00	15.6	1.4
SCAD		0.00	6.3	2.3		0.01	10.4	1.8
	Truly $\epsilon \sim AN(0, 1, -0.5)$							
Normal	0.00	0.00	0.5	1.2	0.00	0.00	0.7	0.9
Two-piece Normal	0.01	0.01	0.4	1.4	0.00	0.01	0.7	1.1
Laplace	0.00	0.00	0.9	1.0	0.00	0.00	1.4	0.7
Two-piece Laplace	0.01	0.01	0.7	1.2	0.00	0.00	1.5	1.0
Inferred	0.01	0.01	0.4	1.4	0.00	0.01	0.9	1.0
LASSO-LS		0.00	11.0	2.4		0.00	19.4	1.9
LASSO-LAD		0.00	8.6	1.8		0.00	15.3	1.4
LASSO-QR		0.00	8.1	2.1		0.00	12.8	1.5
SCAD		0.00	6.1	2.1		0.00	10.1	1.8
	Truly $\epsilon \sim L(0, 1)$							
Normal	0.01	0.01	0.4	1.3	0.00	0.00	0.8	0.9
Two-piece Normal	0.01	0.01	0.5	1.3	0.00	0.00	0.9	1.0
Laplace	0.05	0.06	0.4	1.7	0.01	0.01	0.7	1.2
Two-piece Laplace	0.05	0.07	0.4	1.7	0.01	0.01	0.8	1.2
Inferred	0.04	0.04	0.3	1.7	0.01	0.01	0.7	1.2
LASSO-LS		0.00	10.8	2.5		0.00	20.4	2.0
LASSO-LAD		0.01	9.3	2.5		0.00	17.1	2.0
LASSO-QR		0.01	9.3	2.5		0.00	17.1	2.0
SCAD		0.00	5.9	2.2		0.00	10.3	1.8
	Truly $\epsilon \sim AL(0, -0.5)$							
Normal	0.00	0.00	0.5	1.1	0.00	0.00	0.9	0.8
Two-piece Normal	0.02	0.01	0.4	1.5	0.01	0.01	0.6	1.2
Laplace	0.02	0.01	0.6	1.3	0.00	0.01	0.8	1.0
Two-piece Laplace	0.09	0.12	0.3	1.9	0.04	0.05	0.7	1.5
Inferred	0.09	0.10	0.3	1.8	0.04	0.05	0.6	1.4
LASSO-LS		0.00	10.9	2.3		0.00	18.0	1.8
LASSO-LAD		0.00	9.4	2.3		0.00	15.6	1.7
LASSO-QR		0.00	8.3	2.5		0.00	14.0	2.0
SCAD		0.01	5.7	2.1		0.00	10.2	1.6

Table 6S: Simulation results under $\vartheta = 2$. γ_0: true predictors. $\widehat{\gamma}$: selected variables. CC: number of correctly classified variables ($\sum_{j=1}^{p} \mathrm{I}(\widehat{\gamma}_j = \gamma_{0j})$). FP: number of false positives; TP: number of true positives. LASSO-LAD and LASSO-QR are equivalent when $\alpha = 0$

Truth	Average $p(\gamma_{p+1}, \gamma_{p+2} \mid y)$			
	$\gamma_{p+1} = \gamma_{p+2} = 0$	$\gamma_{p+1} = 1, \gamma_{p+2} = 0$	$\gamma_{p+1} = 0, \gamma_{p+2} = 1$	$\gamma_{p+1} = \gamma_{p+2} = 0$
$\mathrm{N}(0, \vartheta_i)$	0.000	0.000	0.914	0.086
$\mathrm{AN}(0, \vartheta_i, -0.5)$	0.000	0.003	0.096	0.901
$\mathrm{L}(0, \vartheta_i)$	0.000	0.000	0.906	0.094
$\mathrm{AL}(0, \vartheta_i, -0.5)$	0.000	0.000	0.053	0.947

Table 7S: Inference on the error distribution under the $p = 6$ simulation and heteroskedastic $\vartheta_i \propto e^{x_i^T \theta}$ errors

	$P(\gamma_2 = 1 \mid \mathbf{y})$	$P(\gamma_3 = 1 \mid \mathbf{y})$	$P(\gamma_4 = 1 \mid \mathbf{y})$	$P(\gamma_5 = 1 \mid \mathbf{y})$	$P(\gamma_6 = 1 \mid \mathbf{y})$
$q = 0.05$	0.425	0.834	0.961	0.017	0.015
$q = 0.25$	0.751	0.950	0.996	0.016	0.015
$q = 0.5$	0.796	0.970	0.999	0.020	0.016
$q = 0.75$	0.769	0.969	0.999	0.016	0.012
$q = 0.95$	0.473	0.912	0.987	0.016	0.016

Table 8S: Average marginal $P(\gamma_j = 1 \mid \mathbf{y})$ at multiple quantiles $q = 0.05, 0.25, 0.5, 0.75, 0.95$ (i.e. conditioning on asymmetric Laplace errors with fixed $\alpha = 2q - 1$) under the $p = 6$ simulation and heteroskedastic $\epsilon_i \sim N(0, \vartheta_i)$, $\vartheta_i \propto e^{x_i^T \theta}$ errors. Simulation truth is $\theta = (0, 0.5, 1, 1.5, 0, 0)$

	TP	FP
Zellner, Normal errors	2.8	21.3
pMOM, Normal errors	3.0	12.0
pMOM, inferred errors	2.8	10.5
peMOM, Normal errors	1.9	2.9

Table 9S: Number of true and false positives in non-id example with 0.5 probability of degenerate $(y_i, x_i) = (0, \ldots, 0)$. $p = n = 50$, $\theta^* = (0.1, 0.1, 0.1, 0.1, 0.1, 0, \ldots, 0)$, $\vartheta^* = 2$

Gene symbol	Normal	Inferred
C6orf226	1.000	1.000
ECH1	1.000	1.000
CSF2RA	1.000	1.000
RRP1B	0.944	0.999
FBXL19	0.993	0.658
MTMR1	0.183	0.467
SLC35B4	0.209	0.332
RAB3GAP2	0.007	0.040

Table 10S: Six genes with largest $p(\gamma_j = 1 \mid y)$ in the DLD dataset under assumed normality and inferred error distribution.

$\alpha = -0.5$	
Model	$P(\gamma \mid y)$
C6orf226, ECH1, CSF2RA, FBXL19, RRP1B	0.384
SLC35B4, C6orf226, ECH1, CSF2RA, RRP1B	0.349
SLC35B4, C6orf226, MTMR1, ECH1, CSF2RA, RRP1B	0.127
C6orf226, MTMR1, ECH1, CSF2RA, FBXL19, RRP1B	0.049
C6orf226, MTMR1, RAB3GAP2, ECH1, CSF2RA, RRP1B	0.023
$\alpha = 0$	
Model	$P(\gamma \mid y)$
C6orf226, MTMR1, ECH1, CSF2RA, FBXL19, RRP1B	0.454
C6orf226, ECH1, CSF2RA, FBXL19, RRP1B	0.258
SLC35B4, C6orf226, MTMR1, ECH1, CSF2RA, RRP1B	0.108
SLC35B4, C6orf226, ECH1, CSF2RA, RRP1B	0.061
C6orf226, MTMR1, RAB3GAP2, ECH1, CSF2RA, RRP1B	0.016
$\alpha = 0.5$	
Model	$P(\gamma \mid y)$
C6orf226, ECH1, CSF2RA, FBXL19, RRP1B	0.399
SLC35B4, C6orf226, ECH1, CSF2RA, RRP1B	0.359
SLC35B4, C6orf226, MTMR1, ECH1, CSF2RA, RRP1B	0.120
C6orf226, MTMR1, ECH1, CSF2RA, FBXL19, RRP1B	0.051
SLC35B4, C6orf226, RAB3GAP2, ECH1, CSF2RA, RRP1B	0.008

Table 11S: DLD data. Top 5 models when conditioning on asymmetric Laplace residuals and fixed $\alpha = -0.5, 0, 0.5$