Log-Scale Shrinkage Priors for Global-Local Shrinkage

Daniel F. Schmidt
Joint work with Enes Makalic

Department of Data Science and AI
Monash University

Virginia Tech
June 6th, 2024

Outline

Outline

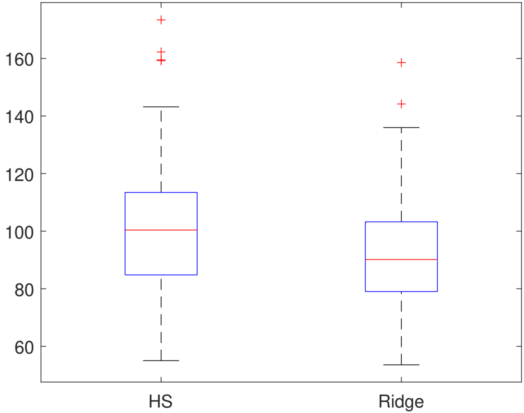
Problem Description

- Consider the problem of estimating a (potentially) sparse vector of parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$
- In particular, consider problems in which
  - $\beta_j = 0$ indicates the parameter has "no effect"
  - The effect of β_j increases with increasing $|\beta_j|$
- Encompasses a wide range of problems:
  - Linear models, GLMs, neural networks, matrix factorisation, etc.
- We make no assumption about the sparsity of configuration of the population vector $\boldsymbol{\beta}$
- Interested in Bayes procedures that are robust to degree of sparsity $\Longrightarrow$ yield small risk even if underlying vector happens to be dense

A Motivating Example

- I was fitting linear models to some genomic expression level data regarding eyes ($p = 200$ predictors).
- I assumed sparse horseshoe prior would give strong performance (conventional genomic wisdom)
- However, I decided to also test ridge regression (which assumes a dense vector) to see how much better horseshoe was doing
- I used cross-validation to estimate prediction error (100 repetitions)

A Motivating Example (2)



- Surprisingly, ridge regression was outperforming horseshoe by 10%
- Why? Almost every variable was weakly associated, so potential reduction in risk due remove variables was seemingly outweighed by the increase in risk due to trying to decide which variables to remove

Global-Local Shrinkage Hierarchies (1)

- The global-local shrinkage hierarchy
  $\Rightarrow$ generalises many popular Bayesian regression priors

$$
\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta} &\sim p(\mathbf{y} \,|\, \boldsymbol{\beta}, \ldots) d\mathbf{y}, \\
\beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j \\
\tau &\sim \pi(\tau) d\tau
\end{aligned}
$$

- Models priors for β_j as scale-mixtures of normals
  $\Rightarrow$ choice of $\pi(\lambda_j)$, $\pi(\tau)$ controls behaviour of the estimator

Global-Local Shrinkage Hierarchies (2)

- The global-local shrinkage hierarchy
  $\Rightarrow$ generalises many popular Bayesian regression priors

$$
\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta} &\sim p(\mathbf{y} \,|\, \boldsymbol{\beta}, \ldots) d\mathbf{y}, \\
\beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j \\
\tau &\sim \pi(\tau) d\tau
\end{aligned}
$$

- Local shrinkers λ_j control selection of variables; e.g.,
  $\lambda_j^2 \sim \mathrm{Exp}(1) \implies$ Bayesian lasso
  $\lambda_j \sim C^+(0,1) \implies$ Bayesian horseshoe
  $\lambda_j \sim \delta_1 \implies$ ridge regression (point mass at $\lambda_j = 1$)

Global-Local Shrinkage Hierarchies (3)

- The global-local shrinkage hierarchy
  $\Rightarrow$ generalises many popular Bayesian regression priors

$$
\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta} &\sim p(\mathbf{y} \,|\, \boldsymbol{\beta}, \ldots) d\mathbf{y}, \\
\beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j \\
\tau &\sim \pi(\tau) d\tau
\end{aligned}
$$

- Global shrinker τ controls overall shrinkage (and multiplicity)

Global-Local Shrinkage Hierarchies (4)

- To see how λ_j and τ affect estimation, consider the linear model

$$\mathbf{y} \mid \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n)$$

- If predictors are orthogonal, then conditional on $\lambda_1, \ldots, \lambda_p$, we have

$$\begin{aligned} \mathbb{E}\left[\beta_j \mid \lambda_j, \tau\right] &= \left(\frac{\lambda_j^2}{n + \lambda_j^2 \tau^2}\right) \hat{\beta}_j \\ &= (1 - \kappa_j)\hat{\beta}_j \end{aligned}$$

where $\hat{\beta}_j$ is the least-squares estimate; so

  - large λ_j (κ_j near zero) implies little shrinkage;
  - small λ_j (κ_j near one) implies a lot of shrinkage

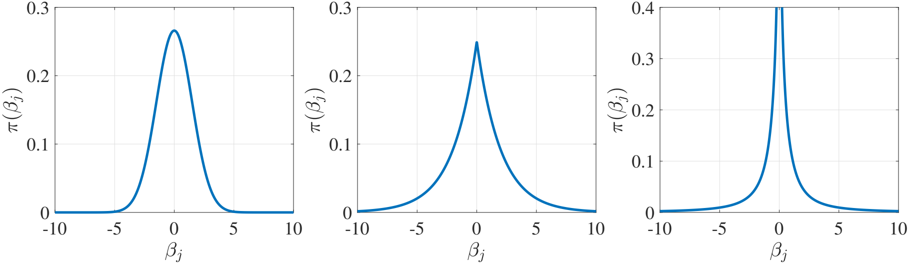
$\Longrightarrow$ setting τ affects the overall degree of shrinkage

Local Shrinkage Priors (1)

- Consider the marginal prior distribution for β_j:

$$p(\beta_j \,|\, \tau) = \int \left(\frac{1}{2\pi\lambda_j^2\tau^2} \right)^{1/2} \exp\left(-\frac{\beta_j^2}{2\lambda_j^2\tau^2} \right) \pi(\lambda_j) d\lambda_j$$

  - Concentration around $\beta_j = 0$ determines sparsity behaviour
  - Tails as $|\beta_j| \to \infty$ determine bias for large effects



- Ridge (ℓ_2, left), lasso (ℓ_1, centre) and horseshoe (right)

Local Shrinkage Priors (2)

- Carvalho, Polson and Scott (2010) proposed two desirable properties of $p(\beta_j \,|\, \tau, \sigma)$
  1. Should concentrate sufficient mass near $\beta_j = 0$ such that

     $$\lim_{\beta_j \to 0} p(\beta_j \,|\, \tau) \to \infty$$

     to guarantee fast rate of posterior contraction when $\beta_j = 0$ (KL supereffiency)
  2. Should have sufficiently heavy tails so that

     $$\mathbb{E}\left[\beta_j \,|\, \mathbf{y}\right] = \hat{\beta}_j + o_{\hat{\beta}_j}(1)$$

     to guarantee asymptotic (in effect-size) unbiasedness
- Neither ridge nor lasso satisfy these; the horseshoe satisfies both

Local Shrinkage Priors (2)

- Carvalho, Polson and Scott (2010) proposed two desirable properties of $p(\beta_j \mid \tau, \sigma)$
  1. Should concentrate sufficient mass near $\beta_j = 0$ such that
  $$\lim_{\beta_j \to 0} p(\beta_j \mid \tau) \to \infty$$
  to guarantee fast rate of posterior contraction when $\beta_j = 0$ (KL supereffiency)
  2. Should have sufficiently heavy tails so that
  $$\mathbb{E}\left[\beta_j \mid \mathbf{y}\right] = \hat{\beta}_j + o_{\hat{\beta}_j}(1)$$
  to guarantee asymptotic (in effect-size) unbiasedness
- Neither ridge nor lasso satisfy these; the horseshoe satisfies both

Outline

Log-Scale Shrinkage Priors (1)

- Observation: the local shrinkage parameters λ_j are scale parameters
- Scale parameters are often more clearly interpreted in log-space
  - Let $\xi_j = \log \lambda_j$, and $p(\xi_j)$ be the transformed density
- In a global-local hierarchy

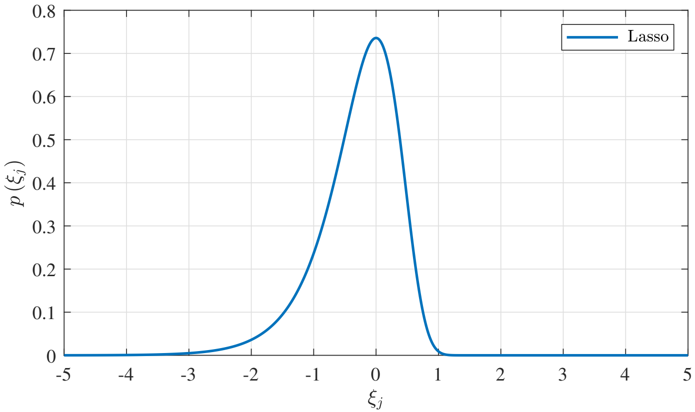
$$\beta_j \mid \lambda_j, \tau \sim N(0, \lambda_j^2 \tau^2)$$

  which implies that $\xi' = \log \lambda_j \tau$ follows

$$\xi' = \xi + \log \tau$$

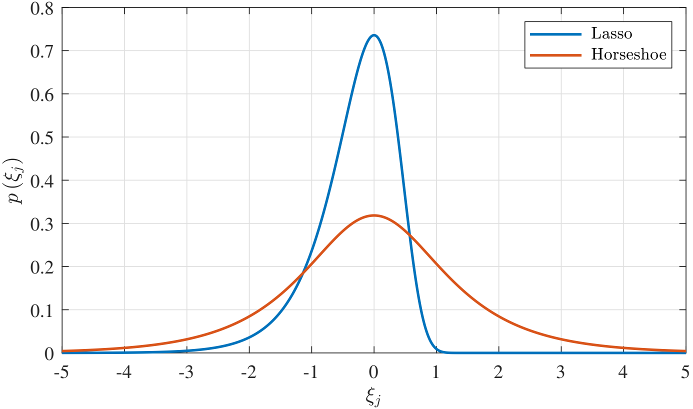
  so that scaling by the global shrinkage parameter induces a location transformation on the log-scales
- Let us examine lasso and horseshoe in log-scale space

Log-Scale Shrinkage Priors (2)



- If $\lambda_j^2 \sim \text{Exp}(1)$ the marginal prior is the Laplace (lasso)
- Lasso is known to overshrink effects if it tries to sparsify
- In $\xi = \log \lambda_j$ space, it's clear it has a strong preference for $\xi_j < 0$ and very light tails for $\xi_j > 0$ explain this

Log-Scale Shrinkage Priors (3)



- If $\lambda_j \sim \mathrm{C}^+(0,1)$ the marginal prior is the horseshoe
- It is clear it spreads probability mass more thinly over the ξ line
- This explains why the horseshoe is less biased, and more aggressive at shrinking than lasso

Log-Scale Shrinkage Priors (4)

- Observation: shrinkage hyperpriors can be thought of as controlling how tightly the shrinkage hyperparameters ξ_j are clustered around $\log \tau$ (the average shrinkage), i.e., a form of "meta-shrinkage"
- The more the probability is spread out, the more variation in shrinkage is allowed, and
  1. the more sparsity inducing it becomes;
  2. the less bias at estimating large effects
- In log-space, ridge regression implies a point mass at $\xi = \log \tau$, i.e., no variation in shrinkage

Log-Scale Shrinkage Priors (5)

- Let us introduce a scale hyperparameter ψ onto our log-scale distribution:

$$p(\xi \mid \log \tau, \psi) = \left(\frac{1}{\psi}\right) p\left(\frac{\xi - \log \tau}{\psi}\right)$$

  - Large ψ implies large variation in shrinkage
  - As $\psi \to 0$ this collapses to a point-mass at $\xi = \log \tau$

  $\implies$ so ψ controls variation from sparsity inducing to ridge regression

- Important: scale transformations in log-scale induce power transformations on the original scale
  - so ψ is controlling the tails of $p(\lambda_j)$

A Tunable Horseshoe (1)

- Let us apply this to the horseshoe; in log-space

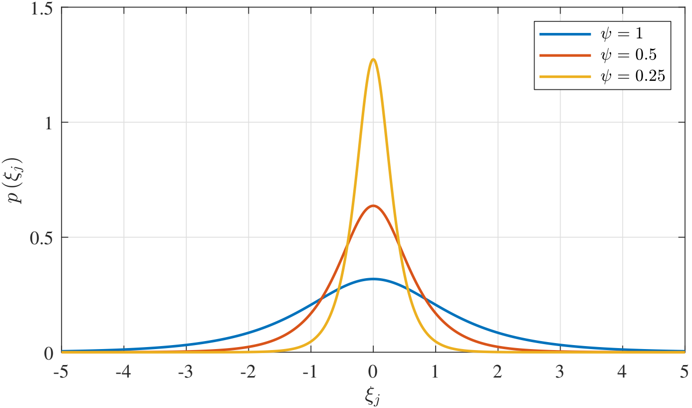
$$p_{\mathrm{HS}}(\xi \mid \psi) = \left(\frac{1}{\pi\psi}\right) \operatorname{sech}\left(\frac{\xi_j}{\psi}\right)$$

which transforms to

$$p_{\mathrm{HS}}(\lambda_j \mid \psi) = \frac{2\lambda_j^{1/\psi - 1}}{\pi\psi(\lambda_j^{2/\psi} + 1)}$$

- Question: can we make horseshoe like ridge regression by making ψ small, but preserve Property I and II?
- This would be useful as it would allow us to have a flexible prior distribution that could estimate dense coefficient vectors well but still retain low bias if the occasional effect was large

A Tunable Horseshoe (2)



- As ψ gets smaller, more probability is concentrated around $\xi = \log \tau$ (in this case $\tau = 0$)
- In the limit as $\psi \to 0$ we recover ridge regression (point mass at $\xi = \log \tau$

The log-Laplace Prior (1)

- Unfortunately, the answer to our question is no
- To prove this, we introduce a new shrinkage prior: the log-Laplace. Place an asymmetric double exponential on $\xi_j = \log \lambda_j$

$$\xi_j \mid \psi_1, \psi_2 \sim \mathrm{DE}(\psi_1, \psi_2)$$

with pdf

$$p(\xi_j \mid \psi_1, \psi_2) = \left(\frac{1}{2\psi(\xi_j)}\right) \exp\left(-\frac{|\xi_j|}{\psi(\xi_j)}\right).$$

where

$$\psi(\xi_j) = I(\xi_j < 0)\psi_1 + I(\xi_j \geq 0)\psi_2$$

- The asymmetric DE is two back-to-back exponential distributions centered at $\xi_j = 0$

The log-Laplace Prior (2)

- When transformed back to the shrinkage parameter λ_j

$$p_{\mathrm{LL}}(\lambda_j \,|\, \psi_1, \psi_2) = \begin{cases} \left(\dfrac{1}{2\psi_1}\right) \lambda_j^{-1+1/\psi_1} & 0 < \lambda_j \leq 1 \\ \left(\dfrac{1}{2\psi_2}\right) \lambda_j^{-1-1/\psi_2} & \lambda_j > 1 \end{cases} .$$

- Some basic properties:
    - Non-differentiable at $\lambda_j = 1$ if $\psi_1 = \psi_2$
    - Discontinuous at $\lambda_j = 1$ if $\psi_1 \neq \psi_2$
    - For $\lambda_j \in (0, 1)$ it is a beta-distribution
    - For $\lambda_j \in (1, \infty)$ it is a Pareto
- Why is this useful?

The log-Laplace Prior (3)

- It upper-bounds all log-concave prior distributions on ξ $\Longrightarrow$ this is effectively all commonly used shrinkage priors
- Specifically, if $f(\xi)$ is a unimodal density with log-linear tails then

$$f(\xi) \leq K\, p_{\mathrm{LL}}(\xi \,|\, \psi_1 = 1/g_1, \psi_2 = 1/g_2) \ \text{ for all } \xi \in \mathbb{R},$$

where $K < \infty$, and

$$g_1 = \lim_{\xi \to -\infty} \left\{ \frac{g(\xi)}{\xi} \right\} \ \text{ and } \ g_2 = \lim_{\xi \to \infty} \left\{ \frac{g(\xi)}{\xi} \right\}$$

where

$$g(\xi) = -\frac{d \log f(\xi)}{d\xi}$$

The log-Laplace Prior (4)

- For example

$$\left(\frac{2}{\pi}\right) p_{\mathrm{LL}}(\xi_j \mid \psi) \leq p_{\mathrm{HS}}(\xi_j \mid \psi) \leq \left(\frac{4}{\pi}\right) p_{\mathrm{LL}}(\xi_j \mid \psi)$$

- Let the prior for coefficient β_j be

$$\beta_j \sim N(0, e^{2\xi_j}) \equiv \beta_j \sim N(0, \lambda_j^2)$$

- Then, if $f(\xi_j)$ has log-linear tails then

$$\int \pi(\beta_j \mid \xi_j) f(\xi_j) d\xi_j \leq c \int \pi(\beta_j \mid \xi_j) p_{LL}(\xi_j \mid 1/g_1, 1/g_2) d\xi_j$$

by the monotone convergence theorem

The log-Laplace Prior (5)

- The marginal distribution, $p(\beta_j \mid \psi_1, \psi_2)$, for the log-Laplace prior is

$$\left(\frac{1}{\sqrt{32\pi}}\right)\left[\frac{1}{\psi_1}E_{\left(\frac{1+\psi_1}{2\psi_1}\right)}\left(\frac{\beta_j^2}{2}\right)+\frac{1}{\psi_2}\left(\frac{2}{\beta_j^2}\right)^{\left(\frac{1+\psi_2}{2\psi_2}\right)}\gamma\left(\frac{1+\psi_2}{2\psi_2},\frac{\beta_j^2}{2}\right)\right]$$

  where $E_n(\cdot)$ is the generalized exponential integral and $\gamma(s,x)$ is the incomplete lower-gamma function

- This nicely separates the effects of
  - the left-hand-tail (controlled by ψ_1), and
  - right-hand-tail (controlled by ψ_2)

  on the marginal distribution over β_j

- Most standard global-local shrinkage priors are bounded by this

The log-Laplace Prior (6)

- Asymptotic tail behaviour and concentration at $\beta_j = 0$
- Concentration: for all $\psi_2 > 0$ then as $|\beta_j| \to 0$ we have
  1. $\pi_{\text{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta_j|^{-1+1/\psi_1}\right)$ if $\psi_1 > 1$;
  2. $\pi_{\text{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(-\log|\beta_j|\right)$ if $\psi_1 = 1$;
  3. $\pi_{\text{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O(1)$ if $\psi_1 < 1$.

  $\implies$ so for $\psi_1 < 1$, prior loses KL superefficiency
- Tail behaviour: for all $\psi_1 > 0$

$$\pi_{\text{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta|^{-1-1/\psi_2}\right)$$

  as $|\beta| \to \infty$
- $\psi_1 = \psi_2 = 1$ is equivalent to horseshoe
- if $\psi_1 \to 0$ and $\psi_2 \to 0$ to mimic ridge regression, the log-Laplace (and therefore any log-linear log-scale prior) loses both Property (I) and Property (II)

The log-Laplace Prior (6)

- Asymptotic tail behaviour and concentration at $\beta_j = 0$
- Concentration: for all $\psi_2 > 0$ then as $|\beta_j| \to 0$ we have
  1. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta_j|^{-1+1/\psi_1}\right)$ if $\psi_1 > 1$;
  2. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(-\log|\beta_j|\right)$ if $\psi_1 = 1$;
  3. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O(1)$ if $\psi_1 < 1$.

  $\Longrightarrow$ so for $\psi_1 < 1$, prior loses KL superefficiency
- Tail behaviour: for all $\psi_1 > 0$

$$\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta|^{-1-1/\psi_2}\right)$$

  as $|\beta| \to \infty$
- $\psi_1 = \psi_2 = 1$ is equivalent to horseshoe
- if $\psi_1 \to 0$ and $\psi_2 \to 0$ to mimic ridge regression, the log-Laplace (and therefore any log-linear log-scale prior) loses both Property (I) and Property (II)

The log-Laplace Prior (6)

- Asymptotic tail behaviour and concentration at $\beta_j = 0$
- Concentration: for all $\psi_2 > 0$ then as $|\beta_j| \to 0$ we have
  1. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta_j|^{-1+1/\psi_1}\right)$ if $\psi_1 > 1$;
  2. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(-\log|\beta_j|\right)$ if $\psi_1 = 1$;
  3. $\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O(1)$ if $\psi_1 < 1$.

  $\Longrightarrow$ so for $\psi_1 < 1$, prior loses KL superefficiency
- Tail behaviour: for all $\psi_1 > 0$

$$\pi_{\mathrm{LL}}(\beta_j \,|\, \psi_1, \psi_2) = O\left(|\beta|^{-1-1/\psi_2}\right)$$

  as $|\beta| \to \infty$
- $\psi_1 = \psi_2 = 1$ is equivalent to horseshoe
- if $\psi_1 \to 0$ and $\psi_2 \to 0$ to mimic ridge regression, the log-Laplace (and therefore any log-linear log-scale prior) loses both Property (I) and Property (II)

Outline

Log-t Shrinkage Priors (1)

- We proposed a new class of shrinkage priors: the log-t

$$\xi_j \sim t_\alpha(\psi)$$

  where $\alpha > 0$ is the degrees-of-freedom

- Transforming $\lambda_j = e^{\xi_j}$ yields

$$p(\lambda_j \,|\, \alpha, \psi) \propto \lambda_j^{-1} \left(\frac{\log(\lambda_j)^2}{\alpha \psi^2} + 1 \right)^{-(\alpha+1)/2}$$

  This is a normal-Jeffrey's prior multiplied by a function of slow variation which renders it normalisable

- Importantly, the t-distribution is not log-linear in its tails

Log-t Shrinkage Priors (2)

- The marginal prior $p_t(\beta_j \mid \alpha, \psi)$ is very unpleasant
- However, we proved the following:
  - As $\psi \to 0$, the prior concentrates around $\xi = \log \lambda_j$ like ridge;
  - but, for all $\psi > 0$ and $\alpha > 0$ it satisfies

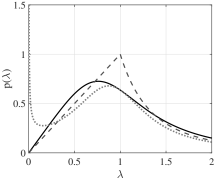
    $$\pi_t(\beta_j \mid \alpha, \psi) \to \infty \text{ as } |\beta_j| \to 0$$

    so it is always super-efficient at $\beta_j = 0$;
  - and the tails of the marginal prior satisfy

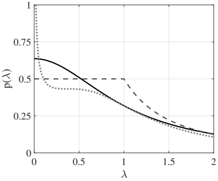
    $$\pi_t(\beta_j \mid \alpha, \psi) \asymp |\beta_j|^{-1}(\log |\beta_j|)^{-\alpha-1} \text{ as } |\beta_j| \to \infty$$

    so it always has low-bias for large effects
- Proof based on the fact that the t-distribution upperbounds any log-Laplace, followed by monotone convergence theorem $\implies$ the log-t can be tuned to estimate dense vectors while still retaining Properties (I) and (II)

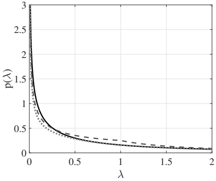
Log-t vs log-Laplace vs horseshoe (1)



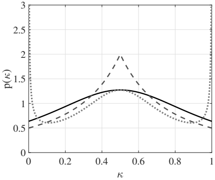
(a) $\psi = 1/2$



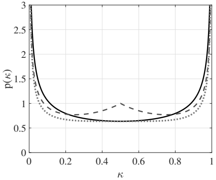
(b) $\psi = 1$



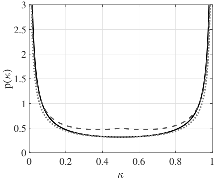
(c) $\psi = 2$



(d) $\psi = 1/2$



(e) $\psi = 1$



(f) $\psi = 2$

Prior probability density plots for the log-hyperbolic secant (solid), log-Laplace (dashed) and log-t with $\alpha = 1$ (dotted) distributions for the λ and $\kappa = 1/(1+\lambda)$.

Log-t vs log-Laplace vs horseshoe (2)

- We studied the sensitivity of GLS to τ for different priors
- The global-local shrinkage multiple-means hierarchy is

$$\begin{aligned} y_j \mid \beta_j &\sim N(\beta_j, 1) \\ \beta_j \mid \lambda_j &\sim N(0, e^{2\xi_j}) \\ \xi_j \mid \tau, \psi &\sim p(\xi_j - \log\tau \mid \psi) \end{aligned}$$

  where the global shrinkage hyperparameter is acts as a location parameter for $\xi_j = \log\lambda_j$.
- Under this hierachy, given τ, then posterior mean is

$$\begin{aligned} \mathbb{E}\left[\hat{\beta}_j \mid y_j, \tau\right] &= \left(\frac{\lambda_j^2}{1+\lambda_j^2\tau^2}\right) y_j \\ &= \left(1 - \mathbb{E}\left[\kappa_j \mid y_j, \tau\right]\right) y_j \end{aligned}$$

  where $\kappa_j = 1/(1+\lambda_j^2\tau^2)$ is the coefficient of shrinkage

Log-t vs log-Laplace vs horseshoe (3)

- Proposition. Let $p(\xi_j \mid \log\tau, \psi)$ be a unimodal, fully-supported prior probability distribution over $\xi_j \in (-\infty, \infty)$ with location $\log\tau$ and scale ψ. Then, under the previous hierarchy

$$\mathbb{P}(\kappa_j \leq \varepsilon \,|\, y_j, \tau, \psi) \leq \left(\frac{\int_{\xi(\varepsilon)}^{\infty} p(\xi \mid \log\tau, \psi) d\xi}{1 - \int_{\xi(\varepsilon)}^{\infty} p(\xi \mid \log\tau, \psi) d\xi} \right) e^{y_j^2/2}$$

- The term in the brackets is the prior-odds in favour of $\kappa_j \geq \varepsilon$.
- Therefore, the probability of no-shrinkage (κ_j close to zero) as a function of the global shrinkage parameter τ decreases at a rate determined by how quickly the odds tend to zero

Log-t vs log-Laplace vs horseshoe (4)

- Proposition. Under the previous hierarchy, with fixed y_j:
  1. if $\lambda_j^2 \sim \mathrm{Exp}(\tau^2)$ (lasso), then
  $$\mathbb{P}(\kappa_j \leq \varepsilon \,|\, y_j, \tau, \psi) = O\left(e^{y_j^2/2} \exp\left[-\frac{(1-\varepsilon)}{\varepsilon\tau^2}\right]\right)$$
  2. if $\lambda_j \sim \mathrm{LogHS}(\tau, \psi)$ (log-hyperbolic secant, i.e., horseshoe), then
  $$\mathbb{P}(\kappa_j \leq \varepsilon \,|\, y_j, \tau, \psi) = O\left(e^{y_j^2/2} \left(\frac{2}{\pi}\right) \left(\frac{\sqrt{\varepsilon}\tau}{\sqrt{1-\varepsilon}}\right)^{1/\psi}\right)$$
  3. if $\lambda_j \sim \mathrm{Log}\text{-}t(\alpha, \tau, \psi)$ (log-t), then
  $$\mathbb{P}(\kappa_j \leq \varepsilon \,|\, y_j, \tau, \psi) = O\left(e^{y_j^2/2} \left[\frac{\psi}{\log\left(\frac{1-\varepsilon}{\varepsilon\tau^2}\right)}\right]^{\alpha}\right)$$

  as $\tau \to 0$

  $\Longrightarrow$ log-t less sensitive to misspecification of τ than horseshoe or lasso

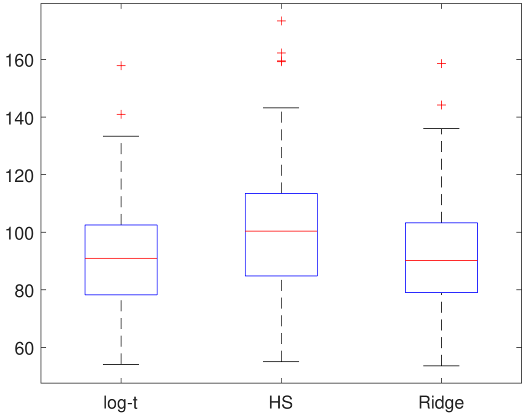
Adaptive log-t (1)

- How do we make our shrinkage priors adaptive?
- Observation: scale ψ on log-scale controls concentration and tails $\Longrightarrow$ one way is to put a prior on ψ and learn it
- We applied this to the log-t
- As ψ is a scale-parameter we chose a weakly-informative half-Cauchy

$$\psi \sim C^{+}(0, 1)$$

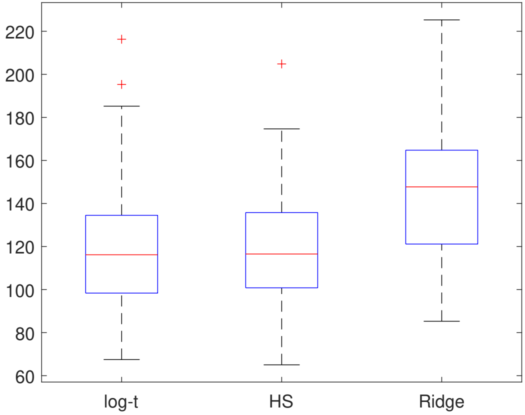
- We developed a simple, efficient Gibbs sampler to sample ψ given $\lambda_1, \ldots, \lambda_p$ and τ
  - We exploited the scale-mixture form of the t-distribution and some log-concavity properties
- We integrated it into efficient MCMC tools for generalized linear models

Adaptive log-t (2)



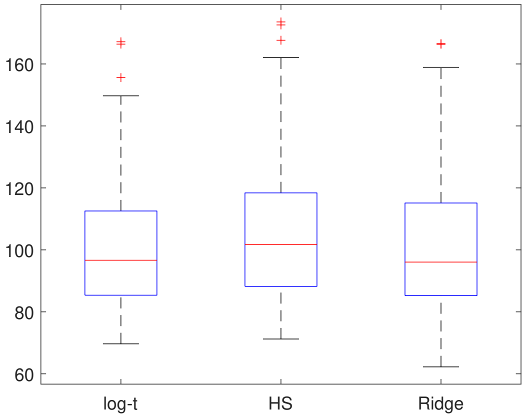
- Revisiting the eye data: log-t performs virtually the same as ridge $\implies$ estimated a $\psi \approx 0.1$

Adaptive log-t (3)



- A second test: kept the five most associated predictors and replaced the remaining 195 with noise $\Rightarrow$ now a sparse problem
- Log-t now performed virtually the same as horseshoe $\Longrightarrow$ estimated a $\psi \approx 0.65$

Adaptive log-t (4)



- A third test: kept the twenty-five most associated predictors and replaced the remaining 175 with noise $\Rightarrow$ now a sparse problem
- Log-t similar to ridge, but less variable (likely due to less overshrinkage)

Conclusion/Future Work

- Interestingly, if we do not bound τ then adaptive log-t can perform poorly; a similar problem observed with horseshoe
- Various experiments on linear models show the adaptive log-t is very robust to configuration of underlying parameter vectors
- I have interest to apply this to multi-layer neural networks
  - It is very unclear what distributions of weights should look like
  - They could vary from layer to layer
- If anyone is interested in working with me on these problems, feel free to contact me :)
- Manuscript this presentation is based on is in preprint form at `https://arxiv.org/abs/1801.02321` ("Log-Scale Shrinkage Priors and Adaptive Bayesian Global-Local Shrinkage Estimation")
- Thank you!