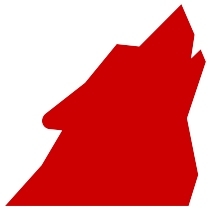
Bayesian variable selection methods

Eric Yanchenko

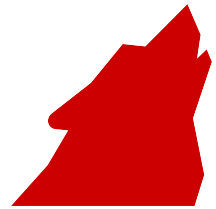
ST740 Guest Lecture

Motivation



- Consider gene assocation study
  - Effects of different genes on a particular disease
- Recruit $n = 100$ patients for the study
- Measure $p = 100,000$ genes for each patient
- Which genes are relevant to the disease?
- What are some challenges/problems associated with this analysis?

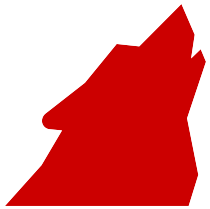
Motivation



1. Classical analysis assumes $p << n$
2. Majority of genes are (likely) not relevant to the response

- What are frequentist solutions to this problem?

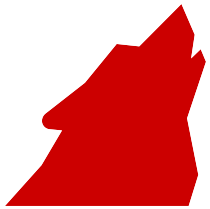
Motivation



1. Classical analysis assumes $p << n$
2. Majority of genes are (likely) not relevant to the response

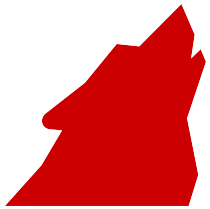
- What are frequentist solutions to this problem?
- Penalization (LASSO, Ridge), forward/backward model selection, ...

Bayesian solution



- Sparsity! Enforce in priors
- Assume that most fixed effects, β_j, are zero
- Large prior mass near 0, heavy tails
- Why Bayes?
  - Uncertainty quantification
  - Intuitive penalization
  - Tuning hyperparameters
  - Frequentist solution is often posterior mode of Bayesian solution

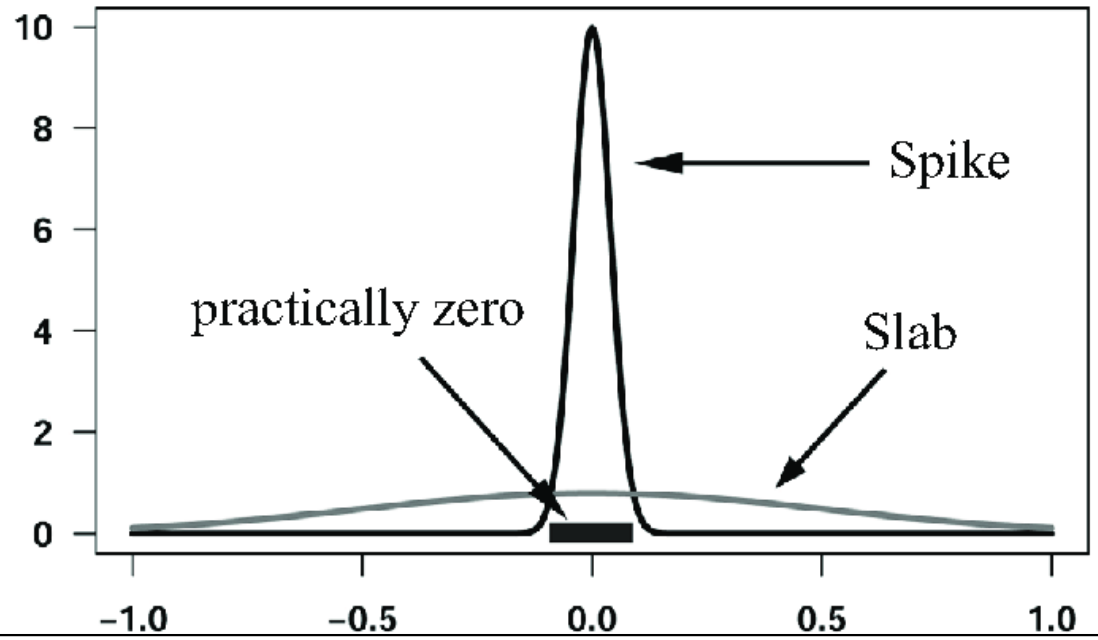
Outline



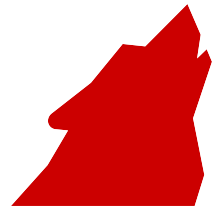
- Discrete
  - Spike-and-slab
- Continuous (shrinkage)
  - Horseshoe
  - Dirichlet-Laplace
  - R2D2 (best method)
- Other
  - Penalized complexity

Spike-and-slab

- George and McCulloch (1997)
- Idea: mixture of two priors
  1. Highly concentrated mass at zero (spike)
  2. Vague prior (slab)



Spike-and-slab

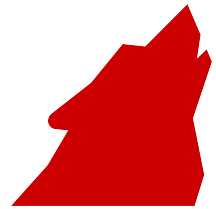


$$
\begin{aligned}
Y_i|\boldsymbol{\beta} &\sim \mathbf{X}_i\boldsymbol{\beta} + \mathsf{N}(0, \sigma^2) \\
\beta_j|\lambda_j &\sim \lambda_j \mathsf{N}(0, \sigma_1^2) + (1 - \lambda_j)\mathsf{N}(0, \sigma_0^2) \\
\lambda_j &\sim \mathsf{Bern}(\pi_j)
\end{aligned}
$$

where $\sigma_0^2 << \sigma_1^2$.

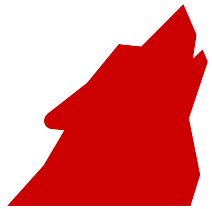
- $\lambda_j = 1$ if variable is included, 0 otherwise
- π_j is probability of including this effect in the model
- Gibbs sampler
- Slow convergence

Shrinkage priors



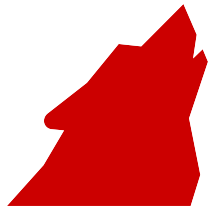
- Discrete form slows convergence
- Continuous notion of inclusion/exclusion for each variable
- *Global-local shrinkage priors*

Shrinkage priors



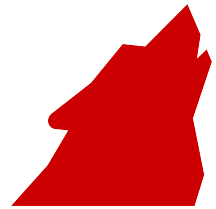
- $\beta_j|\sigma, \gamma_0, \gamma_j \sim N(0, (\sigma\gamma_0\gamma_j)^2)$
- γ_0 controls *global shrinkage*
- γ_j controls *local shrinkage*
- Want β priors to have:
  - Large mass near 0 (most effects are non-significant)
  - Heavy tails (capture significant effects)

Bayesian LASSO



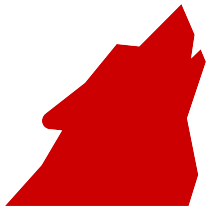
- Recall $\beta_j \sim \text{DE}(\tau)$
- Large prior mass at 0
- τ is KNOWN and/or FIXED
- Just global variance, no local

Horseshoe prior



- Carvalho et al. (2009)
- $\gamma_0 \sim \text{HalfCauchy}(1)$
- $\gamma_j \sim \text{HalfCauchy}(1)$
- Recall that if $X \sim \text{Cauchy}(0, 1)$, then $f(x) = \{\pi(1 + x^2)\}^{-1}$.
- If $Y_j|\beta_j \sim \text{N}(\beta_j, 1)$ and $\beta_j \sim \text{N}(0, \gamma_j^2)$, find $\text{E}(\beta_j|Y_j)$.

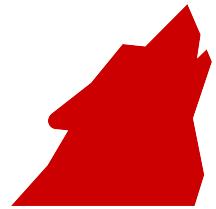
Horseshoe prior



$$\mathsf{E}(\beta_j|Y_j) = \frac{\gamma_j^2}{1+\gamma_j^2}Y_j + \frac{1}{1+\gamma_j^2}0 = \frac{\gamma_j^2}{1+\gamma_j^2}Y_j := (1-\kappa_j)Y_j$$

- κ_j controls the shrinkage for effect j
  - No shrinkage if 0
  - Total shrinkage if 1
- Exercise (in pairs): If $\gamma_j \sim$ HalfCauchy(1) and $\kappa_j = (1+\gamma_j^2)^{-1}$, find the distribution of κ_j.

Horseshoe prior

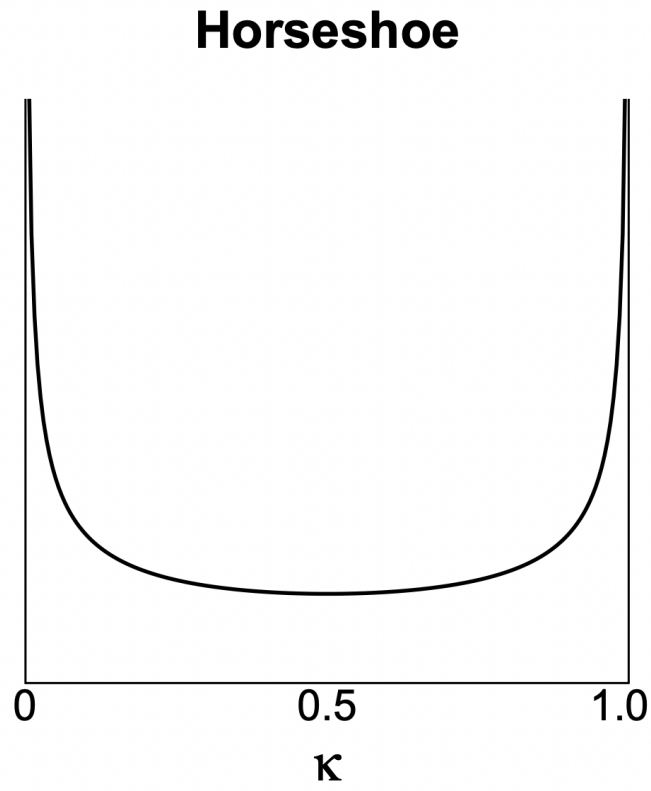


(Do on board.)
Let $X = \gamma_j \sim \text{HalfCauchy}(1)$. Then
$Y = \kappa_j = 1/(1+X^2) \implies X = \{(1-Y)/Y\}^{1/2}$.

$$\left|\frac{dX}{dY}\right| = \frac{1}{2y^{3/2}(1-y)^{1/2}}$$

$$\begin{aligned} f_Y(y) &= f_X(\{(1-Y)/Y\}^{1/2}) \times \frac{1}{2y^{3/2}(1-y)^{1/2}} \\ &\propto y^{-1/2}(1-y)^{-1/2} \\ &\sim \text{Beta}(1/2, 1/2) \end{aligned}$$

Horseshoe prior



- Beta$(1/2, 1/2)$ has “horseshoe" shape
- Unbounded at 0 and 1
- High prior mass on no shrinkage and total shrinkage

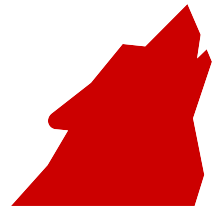
Dirichlet-Laplace prior

- Bhattacharya et al. (2015)
- $\beta_j | \phi, \tau \sim \text{DE}(0, \phi_j \tau)$
- $\phi \sim \text{Dirichlet}(a, \ldots, a)$
- $\tau \sim \text{Gamma}(na, 1/2)$
- τ is “global" component
- ϕ is “local" component
  - Apportions variance to each fixed effect
- Which parameter seems like it will be particularly difficult to sample from? Why?
-

Dirichlet-Laplace prior

- Bhattacharya et al. (2015)
- $\beta_j|\phi,\tau \sim \text{DE}(0,\phi_j\tau)$
- $\phi \sim \text{Dirichlet}(a,\ldots,a)$
- $\tau \sim \text{Gamma}(na, 1/2)$
- τ is “global" component
- ϕ is “local" component
  - Apportions variance to each fixed effect
- Which parameter seems like it will be particularly difficult to sample from? Why?
- ϕ because it’s constrained. And not conjugate.

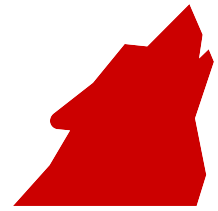
Clever Gibbs trick



If $X \sim \text{giG}(\lambda, \rho, \chi)$ (generalized inverse Gaussian distribution), then

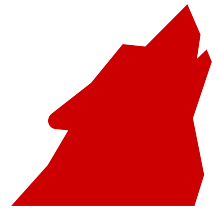
$$f(x) \propto x^{\lambda-1} e^{-0.5(\rho x + \chi/y)}.$$

Clever Gibbs trick



Theorem 2.1 (Bhattacharya et al., 2015)
The joint posterior of $\phi|\beta$ has the same distribution as $(T_1/T, \ldots, T_p/T)$ where $T_j \sim giG(a-1, 1, 2|\beta_j|)$ and $T = \sum_{j=1}^{p} T_i$.

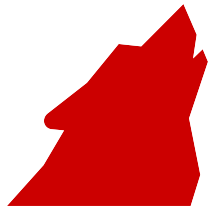
Proof of Theorem 2.1



(Do on board).
For simplicity, assume $Y_i = \beta_i + \epsilon_i$, i.e., no covariates.
(1) Integrate out τ

$$\pi(\phi_1, \ldots, \phi_{n-1}|\boldsymbol{\beta}) \propto \prod_{j=1}^{n} \phi_j^{a-1} \frac{1}{\phi_j} \int_{\tau=0}^{\infty} e^{-\tau/2} \tau^{\lambda-n-1} e^{-\sum_{i=1}^{n} |\beta_j|/(\tau\phi_j)} d\tau$$

Proof of Theorem 2.1

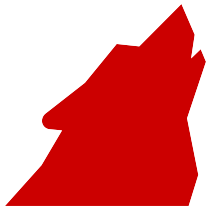


(2) Result from Kruijer, Rousseau, and van der Vaart (2010).
Let $T_1, \ldots, T_n$ be independent random variables on $(0, \infty)$ where $T_j \sim f_j$. Let $\phi_j = T_j/T$ where $T = \sum_{i=1}^{n} T_j$. Then the joint density of f of $(\phi_1, \ldots, \phi_{n-1})$ on simplex $\mathcal{S}^{n-1}$ has the form

$$f(\phi_1, \ldots, \phi_{n-1}) = \int_{t=0}^{\infty} t^{n-1} \prod_{i=1}^{n} f_j(\phi_j t)\, dt$$

where $\phi_n = 1 - \sum_{i=1}^{n-1} \phi_i$

Proof of Theorem 2.1



(3) In our case,

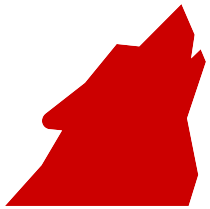
$$f_j(x) \propto x^{-\delta} e^{-|\beta_j|/x} e^{-x/2}$$

so

$$f(\phi_1, \ldots, \phi_{n-1}) = \left(\prod_{i=1}^{n} \phi_j^{-\delta} \right) \int_{t=0}^{\infty} e^{-t/2} t^{n-1-n\delta} e^{-\sum_{j=1}^{n} |\beta_j|/(\phi_j t)} dt$$

Set equal to our parameters and observe that f_j is a giG. $\square$

R2D2 prior



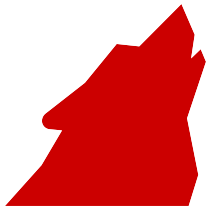
- Recall: “uninformative" priors may be quite “informative" for transformation of the parameter
  - logit of p in binomial model
- We will derive a Bayesian R^2. Assume for a moment that it is similar to usual R^2
  - Terrible fit, $R^2 \to 0$. Great fit, $R^2 \to 1$.
- What do you think the prior distribution of R^2 is for: Vague, HS and DL?

R2D2 prior

Method	Mean	St. Dev
Vague	0.990	0
DL	0.990	0.001
HS	0.901	0.212

Prior R^2 distribution for different priors

R2D2 prior



- Zhang et al. (2022)
- Seek to place a prior on Bayesian R^2
- Recall classical R^2

$$\text{classical } R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - mean(\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

- What are some issues with this measure in a Bayesian context?
-
-

R2D2 prior

- Zhang et al. (2022)
- Seek to place a prior on Bayesian R^2
- Recall classical R^2

$$\text{classical}R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - mean(\hat{y}_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

- What are some issues with this measure in a Bayesian context?
- Does not incorporate posterior uncertainty in fitted values
- Strong prior and weak data, fitted variance could be greater than total variance, i.e., $R^2 > 1$.

Bayesian R^2

- Derive from first principles, i.e., squared correlation between observed and modeled value
- In pairs: if $y = \boldsymbol{x}^T\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$ and $x \perp \epsilon$, find $Cor^2(y, x^T\beta)$.

Bayesian R^2

(On board).

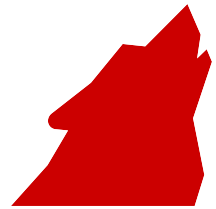
$$R^2 = \frac{cov^2(y, x^T\beta)}{var(y)var(x^T\beta)} = \frac{cov^2(x^T\beta + \epsilon, x^T\beta)}{var(x^T\beta + \epsilon)var(x^T\beta)} = \frac{var(x^T\beta)}{var(x^T\beta) + \sigma^2}$$

Marginal R^2

- Consider prior such that $E(\beta) = 0$ and $Var(\beta) = \sigma^2\Lambda$ where $\Lambda = diag(\lambda_1, \ldots, \lambda_p)$

In pairs: find $var(x^T\beta)$. Remember that x is treated as random with mean 0 and covariance Σ such that $diag(\Sigma) = (1, \ldots, 1)$.

Marginal R^2



(On board).

$$Var(x^T\beta) = E_x(Var_\beta(x^T\beta|x)) + Var_x(E_\beta(x^T\beta|\beta))$$
$$= E_x(\sigma^2 x^T\Lambda x) + Var_x(0) = \sigma^2 E_x(tr(x^T\Lambda x))$$
$$= \sigma^2 tr(\Lambda E_x(x^T x)) = \sigma^2 \sum_{j=1}^{p} \lambda_j$$

Marginal R^2

Thus,

$$R^2 = \frac{var(x^T\beta)}{var(x^T\beta) + \sigma^2} = \frac{\sigma^2 \sum_{j=1}^{p} \lambda_j}{\sigma^2 \sum_{j=1}^{p} \lambda_j + \sigma^2} := \frac{W}{W+1}$$

Marginal R^2

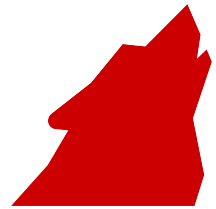
Thus,

$$R^2 = \frac{var(x^T\beta)}{var(x^T\beta) + \sigma^2} = \frac{\sigma^2 \sum_{j=1}^{p} \lambda_j}{\sigma^2 \sum_{j=1}^{p} \lambda_j + \sigma^2} := \frac{W}{W+1}$$

If $R^2 \sim \text{Beta}(a, b)$, then $W \sim \text{BetaPrime}(a, b)$

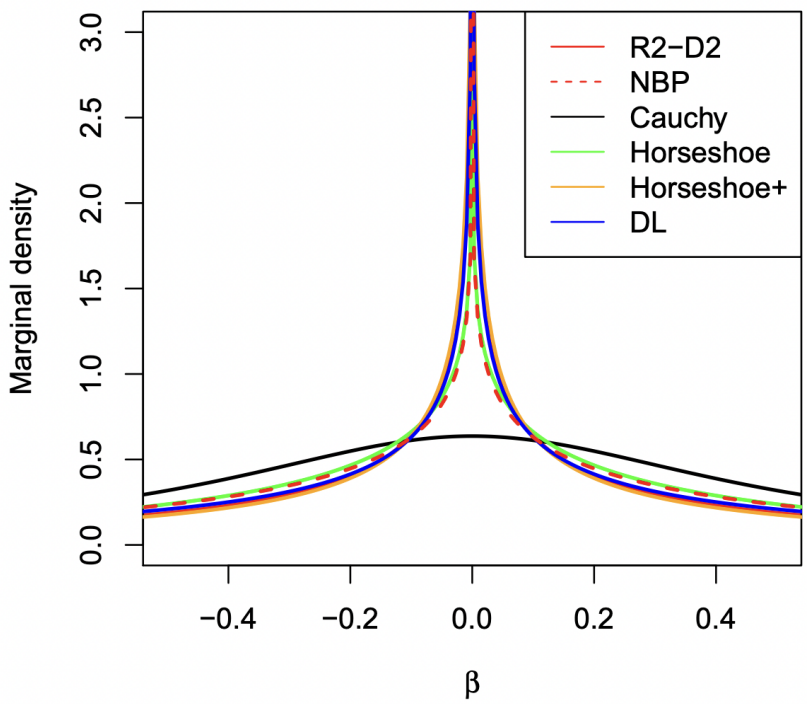
$$f(x) = \frac{1}{B(a,b)} \frac{x^{a-1}}{(1+x)^{a+b}}, \; x \geq 0.$$

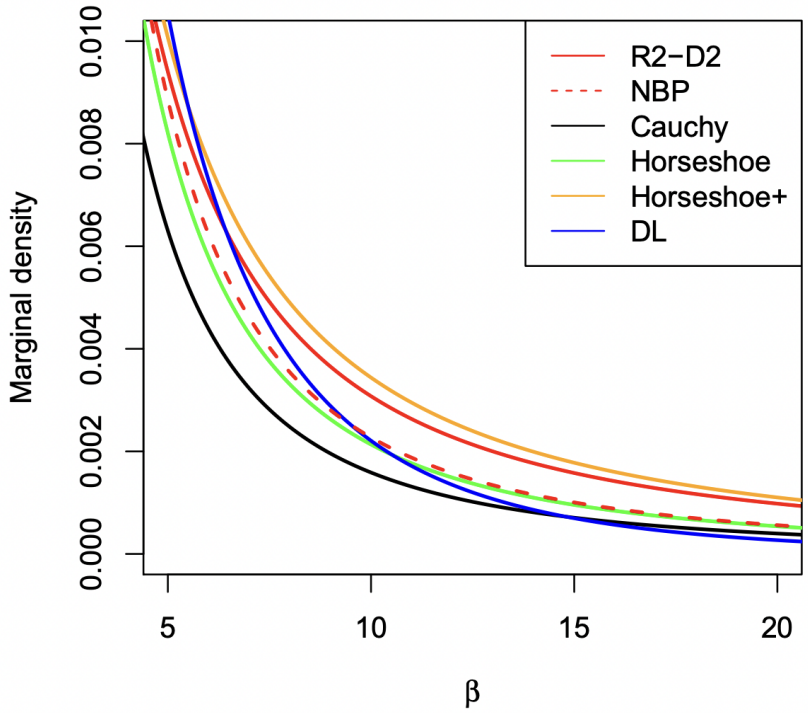
R2D2 prior



- $\beta_j|\sigma^2, \phi_j, W \sim \text{DE}(0, \sigma^2\phi_j W)$
- $W \sim \text{BetaPrime}(a, b)$ (global component)
- $(\phi_1, \ldots, \phi_p) \sim \text{Dirichlet}(a_\pi, \ldots, a_\pi)$ (local component)
- R^2 dirichlet-decomposition (R2D2)
- Extended to generalized linear mixed models in Yanchenko et al. (2021)

R2D2 is the best



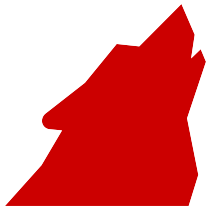


R2D2 is the best

- Recall: shrinkage prior should have large mass near 0 and heavy tails
- Slow decay in tails, high concentration at 0.

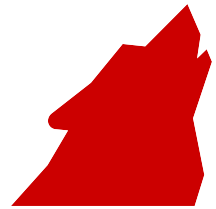
	Tail Decay	Concentration at 0
Horseshoe	$O(\frac{1}{\beta^2})$	$O(\log \frac{1}{\lvert\beta\rvert})$
DL	$O(\frac{\lvert\beta\rvert^{a_*/2-3/4}}{\exp(\sqrt{2\lvert\beta\rvert})})$	$O(\frac{1}{\lvert\beta\rvert^{1-a_*}})$
R2D2	$O(\frac{1}{\lvert\beta\rvert^{1+2b}})$	$O(\frac{1}{\lvert\beta\rvert^{1-2a_\pi}})$

Comparing priors



- Discrete is most intuitive
- GL framework is better computationally (normal mixtures)
- ϕ parameter in DL and R2D2 allows easier comparison of local variance of each fixed effect
- DL and R2D2 advanced theory
- R2D2 framework allows you to more intuitively incorporate domain information than HS or DL
    - Arguably more principled as well

Posterior contraction results

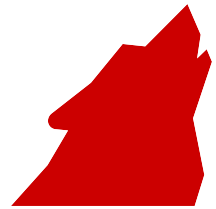


- One of main theoretical goal for shrinkage priors
- Want to show that as $n \to \infty$ AND $p_n \to \infty$, the posterior of β “contracts" around the “true" value, β_0

For any $\epsilon > 0$,

$$P_{\beta_0}\{\pi_n(\beta_n : ||\beta_n - \beta_0|| > \epsilon | \boldsymbol{Y}_n) \to 0\} \to 1 \text{ as } n \to \infty.$$

Posterior contraction results



- Song and Liang (2017) give conditions which yield posterior contraction

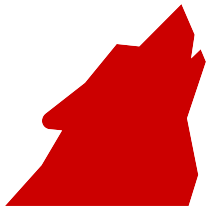
Let $\pi(\boldsymbol{\beta}|\sigma) = \prod_{j=1}^{p} g(\beta_j/\sigma)\sigma,\ \sigma \sim \text{InvGamma}(a_0, b_0)$. Let $k_n, E_n \to 0$, u be a constant

1.
$$1 - \int_{-k_n}^{k_n} g(x)dx \leq p_n^{-(1+u)}$$
(sufficient mass at origin)

2.
$$-\log\left(\inf_{x\in(-E_n,E_n)} g(x)\right) = O(\log p_n)$$
(sufficient mass in tails)

HS and DL meet these criteria.

Simulation study



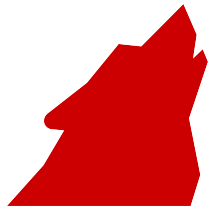
- GOAL: compare the performance of vague and Horseshoe prior
- Let $n = 60$ and $p = 20, 50, 100$ for

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} + \epsilon_i$$

  where $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\epsilon_i \sim N(0, 1)$
- Generate $\beta_j^0 \sim N(1, 1)$ if $j = 1, \ldots, 0.1p$ and $\beta_j^0 = 0$ otherwise
  - Only 10 % of covariates are significant
- Vague ($\beta_j \sim N(0, 100)$) and Horseshoe prior
- Compute average MSE

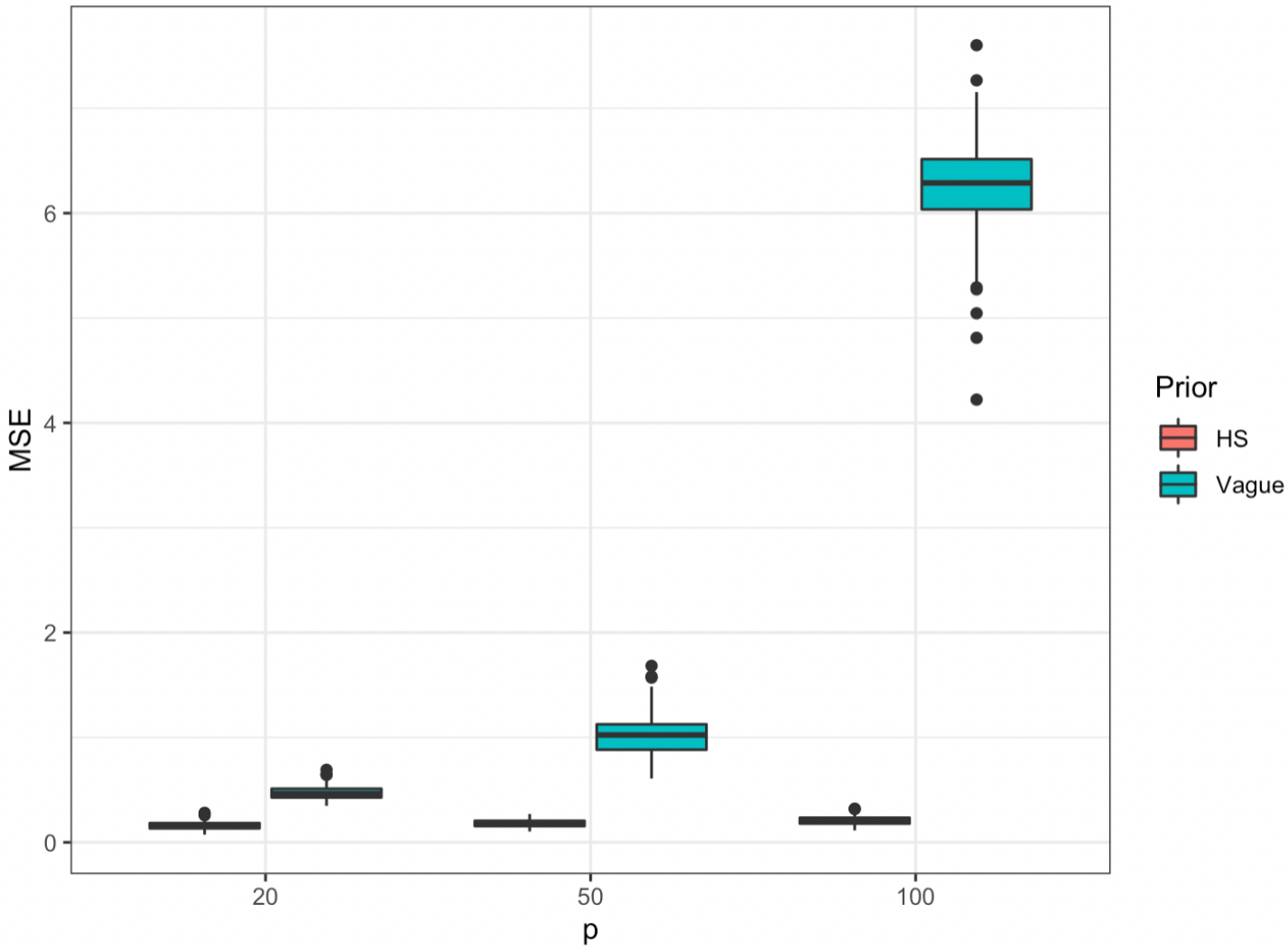
$$rMSE = \left(\frac{1}{pS} \sum_{j=1}^{p} \sum_{s=1}^{S} (\beta_j^{(s)} - \beta_j^0)^2 \right)^2$$

Simulation study

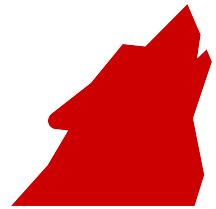


- Recall: 10% of β_j's are significant and on order of 1
- For $p = 20$, the methods will perform the same?
  A. TRUE
  B. FALSE
- For $p = 50$, the vague prior will have rMSE around
  A. 0.5
  B. 1
  C. 5
  D. 20
- For $p = 100$, the vague prior will have rMSE around
  A. 0.5
  B. 1
  C. 5
  D. 20

Simulation study

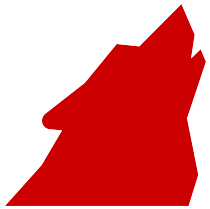


Penalized complexity priors



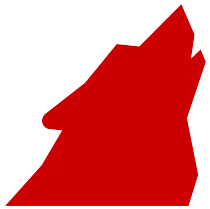
- Shrinkage priors “shrink" the parameter estimates towards a “null" or “baseline" model
- What is this baseline model?

Baseline models



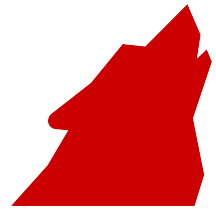
- Shrinkage priors “shrink" the parameter estimates towards a “null" or “baseline" model
- What is this baseline model?
  - Intercept only model
- Horseshoe and DL shrink individual parameters
- R2D2 shrinks the entire model towards $R^2 = 0$ which induces shrinkage on fixed effects
  - R2D2 can be thought of as prior on model fit

Penalized complexity priors



- Simpson et al. (2017) consider prior on model fit
- Prior on KL divergence between model and “baseline" model
- “Penalize the complexity" of the model (PC priors)
- Large prior mass near $KLD = 0$ (and other considerations) yields $\sqrt{2KLD} \sim Exp(\lambda)$.
  - Put on natural distance scale

Penalized complexity priors



- Example: Gaussian random effects
- $Y_i = \beta_0 + X_i\boldsymbol{\beta} + Z_i\boldsymbol{\alpha} + \epsilon_i$
- Random effect: $\boldsymbol{\alpha} \sim \mathrm{N}(0, \tau^{-1}\mathbf{I}_L)$ (easy to extend to other covariances like spatial)
- Base model: no random effect $\iff \tau = \infty$ (zero variance/constant intercept)
- PC prior for τ is

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}e^{-\lambda\tau^{-1/2}}, \ \tau \geq 0 \text{ (Type 2 Gumbel)}$$

Derivation of PC prior

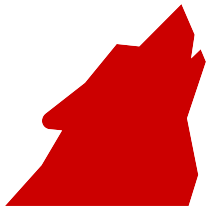
(On board).
(1) Find KL divergence between base and flexible model

$$KL(N(0, \tau^{-1}\mathbf{I}_L)||N(0, \tau_0^{-1}\mathbf{I}_L)) = \frac{1}{2}\left\{\frac{\tau_0}{\tau}L - L - \log(\frac{\tau}{\tau_0})\right\}$$
$$= \frac{1}{2}\frac{\tau_0}{\tau}L\left\{1 - \frac{\tau}{\tau_0} - \frac{\tau}{\tau_0}\log(\frac{\tau}{\tau_0})\right\}$$

As $\tau_0 \to \infty$,

$$KLD \to \frac{L}{2}\frac{\tau_0}{\tau}$$

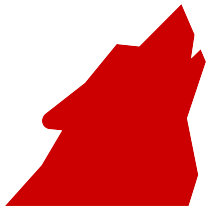
Derivation of PC prior



By definition of the PC prior,
$\sqrt{2KLD} = \sqrt{L\tau_0/\tau} \sim Exp(\lambda = \theta/\sqrt{L\tau_0})$.

$$f_\tau(\tau) = f_{Exp(\lambda)}(\sqrt{L\tau_0/\tau}) \times \frac{1}{2}\sqrt{L\tau_0}\tau^{-3/2} = \frac{1}{2}\theta\tau^{-3/2}e^{-\theta\tau^{-1/2}} \checkmark$$

Summary



- Looked at different ways to enforce sparsity into our models when $p > n$
- Discrete vs. continuous
- Can shrink parameters individually or model as a whole
- Great research area