Bayesian shrinkage methods: Priors, considerations, and possibilities.

Sara van Erp

Utrecht University

Fall meeting BMS-ANed
07-10-2022

Outline[1]

1. Setting the scene
2. Classical penalization
3. Why use Bayesian penalization?
4. Shrinkage priors: An application and simulation
5. Why not to use Bayesian penalization?
6. Extensions: Meta-analysis and SEM
7. Conclusion

[1] Slides are available at:
https://github.com/sara-vanerp/presentations

Setting the scene

Situations with many parameters relative to the sample size, e.g., many predictors in regression analysis.

Need to *select* predictors to avoid overfitting.

Example: Communities and crime data (Redmond & Baveja, 2002) [2]

- DV: number of violent crimes
- 172 predictors after creating dummies
- 319 observations after listwise deletion
- Split into a training ($n = 159$) and test ($n = 160$) set

[2] Available at: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`

Bias-variance tradeoff

Variance: extent to which the model changes if we use a different training set (more flexible model = higher variance).

Bias: the error introduced by using a model that simplifies reality (more flexible model = less bias).

Overfitting:

- Picking up noise in our data.
- Model with low bias but high variance, i.e., a flexible model (large number of predictors).
- As the number of predictors increases, the training MSE will decrease but the test MSE will start increasing again at some point.

Classical penalization

Adds a penalty to the estimation problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \{\text{RSS} + \text{penalty}\}$$

For example, the lasso ($\lambda \sum_{j=1}^{p} |\beta_j|$) or the ridge ($\lambda \sum_{j=1}^{p} |\beta_j^2|$) penalty.

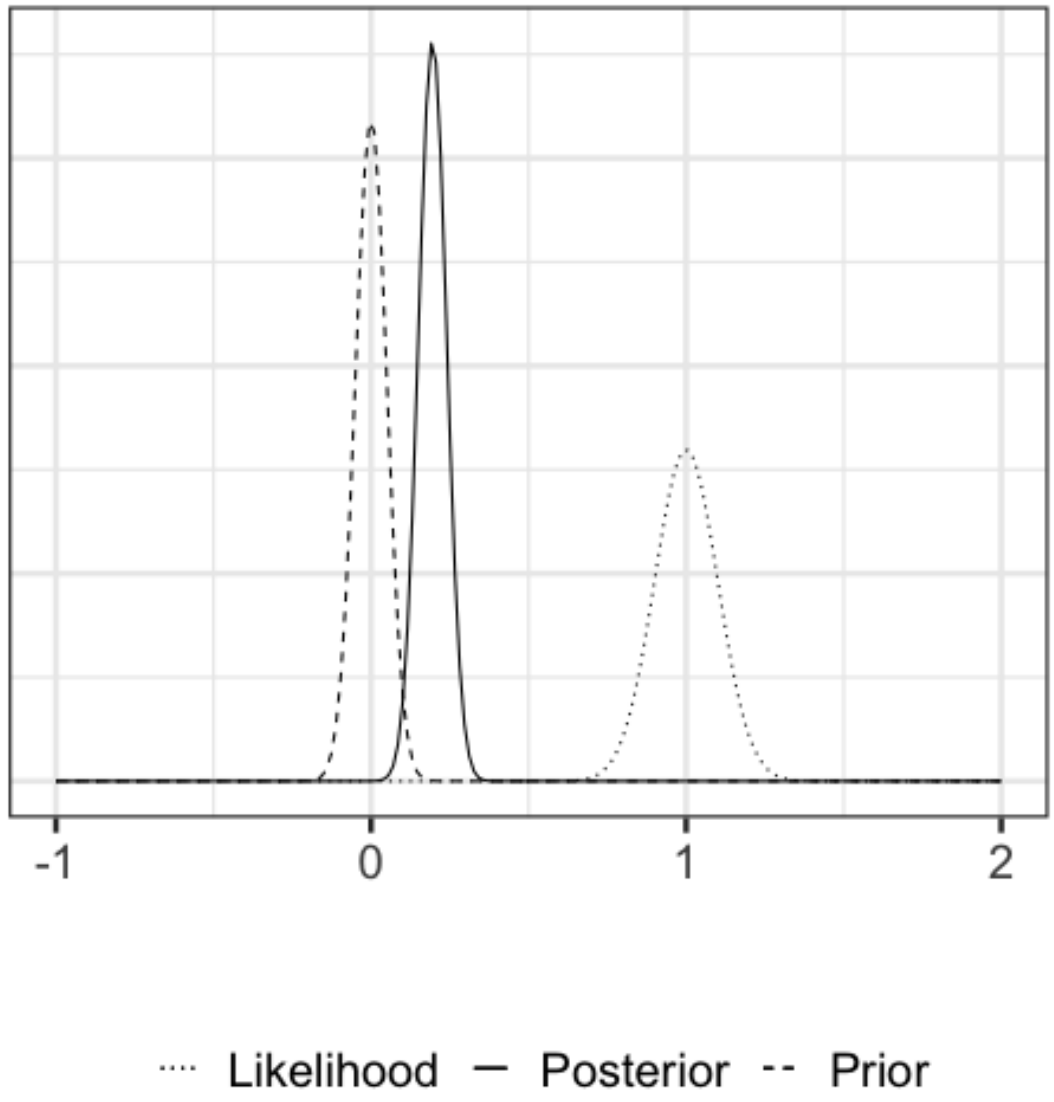
The tuning parameter λ is typically determined via cross-validation.

Communities and crime: Classical penalization

Method	PMSE	# of predictors
Ridge	0.258	160
Elastic net	0.460	26
Lasso	0.508	33

Bayesian analysis

Posterior $\propto$ Prior $\times$ Likelihood



Why use Bayes?

Shrinkage prior = penalty

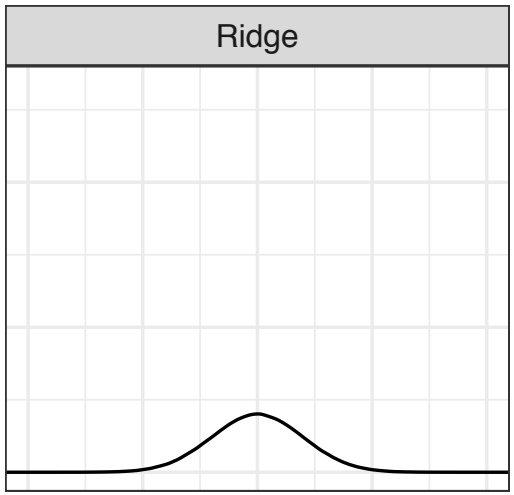
Advantages:

- Automatic uncertainty estimates
- Intuitive interpretation
- Incorporate prior info

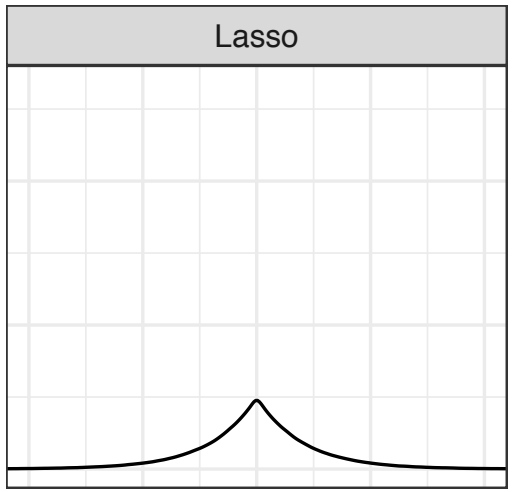
- Shrinkage natural through the prior
- Many different (non-convex) shrinkage priors exist (free lunch?)
- "Tuning" of λ via hyperprior specification

Shrinkage priors 1: Classical counterparts

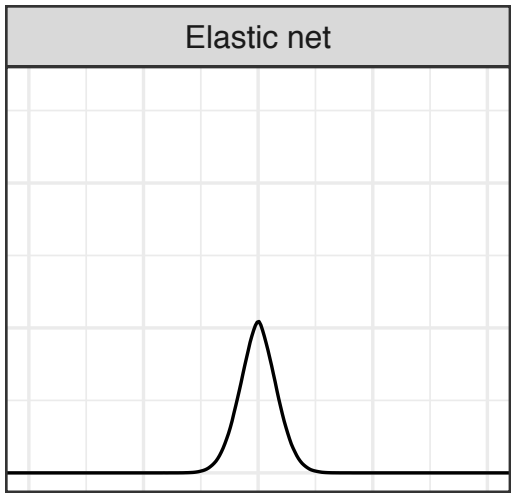
Ideally, shrinkage priors have a peak at zero and heavy tails.



Hsiang (1975)



Park & Casella (2008)



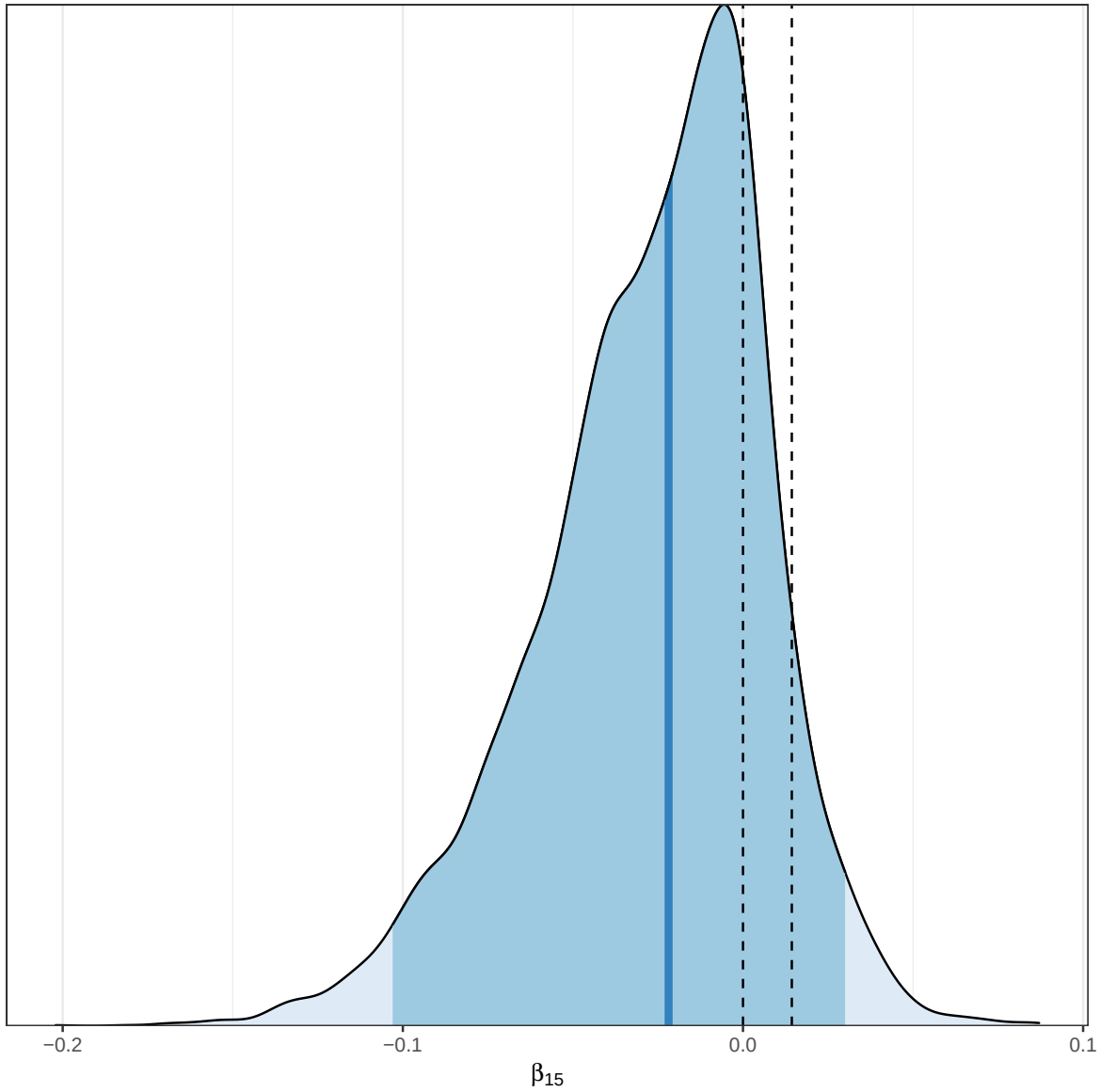
Li & Lin (2010)

Communities and crime: Bayesian penalization (1)

Framework	Method	PMSE	# of predictors
Classical	Ridge	0.258	160
	Elastic net	0.460	26
	Lasso	0.508	33
Full Bayes	Ridge	0.217	61
	Elastic net	0.216	62
	Lasso	0.216	46

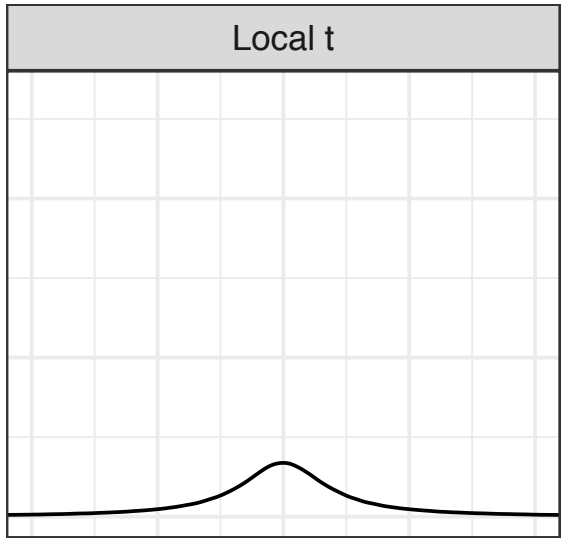
Note: selection of predictors is not automatic in the Bayesian framework.

Communities and crime: Bayesian vs. classical CIs

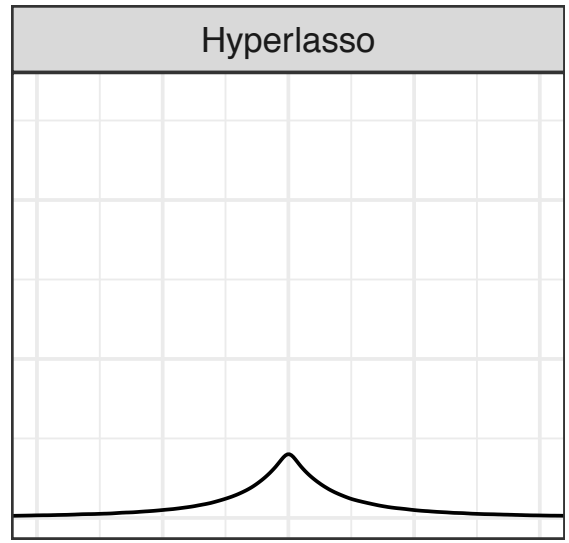


Shrinkage priors 2: “Bayesian“ shrinkage priors

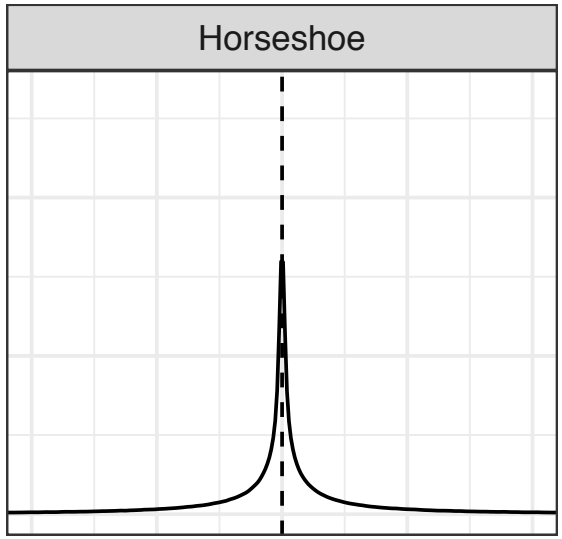
Ideally, shrinkage priors have a peak at zero and heavy tails.



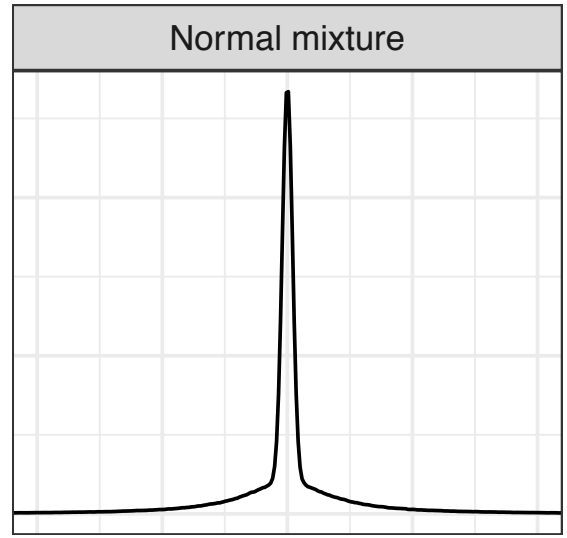
Meuwissen, Hayes, & Goddard (2001)



Griffin & Brown (2011)



Carvalho et al. (2010)



Mitchell & Beauchamp (1988); George & McCulloch (1993)

Communities and crime: Bayesian penalization (2)

Framework	Method	PMSE	# of predictors
Classical	Ridge	0.258	160
	Elastic net	0.460	26
	Lasso	0.508	33
Full Bayes	Ridge	0.217	61
	Elastic net	0.216	62
	Lasso	0.216	46
	Local t	0.216	60
	Hyperlasso	0.215	46
	Regularized horseshoe	0.226	31
	Normal mixture	1.683	54

Note: selection of predictors is not automatic in the Bayesian framework.

Shrinkage priors: Simulation study

Comparison of these priors in a simulation study[3] showed that:

- Different penalization methods generally perform very similarly in terms of prediction accuracy when $p < n$
- Differences become more pronounced as $p > n$ (only 1 condition)
- Not one method outperforms the others in terms of correct and false inclusion rates
- Selection accuracy (MCC) is low for all methods in $p > n$ condition
- Methods vary greatly in terms of computational efficiency

[3] van Erp, Oberski, & Mulder (2019)

Why not to use Bayes?

- Many shrinkage priors make it difficult to choose
- Tuning of priors
- Computationally inefficient
- No automatic variable selection

First two issues can be (partly) solved through a prior sensitivity analysis.

Do the advantages weigh up against the disadvantages?

Tuning of priors

Some priors have more hyperparameters to "tune" than others.

Ridge:

$$\beta_j|\lambda, \sigma^2 \sim \text{Normal}(0, \frac{\sigma^2}{\lambda})$$

Regularized horseshoe:

$$\begin{aligned}
\beta_j|\tilde{\tau}_j^2, \lambda &\sim N(0, \tilde{\tau}_j^2\lambda), \text{ with } \tilde{\tau}_j^2 = \frac{c^2\tau_j^2}{c^2 + \lambda^2\tau_j^2} \\
\lambda|\lambda_0^2 &\sim \text{half-Cauchy}(0, \lambda_0^2), \text{ with } \lambda_0 = \frac{p_0}{p - p_0}\frac{\sigma}{\sqrt{N}} \\
\tau_j &\sim \text{half-Cauchy}(0, 1) \\
c^2|\nu, s^2 &\sim \text{inverse Gamma}(\nu/2, \nu s^2/2)
\end{aligned}$$

Extension 1: Meta-analysis

Idea: use shrinkage priors to select relevant moderators in meta-analysis[4].

Lasso and regularized horseshoe priors implemented in the R package `pema` with selection based on CIs.

BRMA outperformed meta-regression in terms of predictive accuracy and specificity, especially for small n.

Regression coefficients are biased towards zero, but residual heterogeneity estimate is not.

[4] van Lissa, van Erp, & Clapper (under review)

Extension 2: SEM

Idea: use shrinkage priors to select non-zero parameter in SEMs.

Possible applications:

- Cross-loadings and residual covariances in CFA
- Loadings in EFA
- Covariates in MIMIC models
- Mediators in mediation analysis
- Violations of measurement invariance
- ...

Shrinkage priors can (in theory) be applied to any model with (too) many parameters where some are assumed to be zero.

Conclusion

In the current age of big data and complex models, shrinkage methods are more important than ever.

Bayesian shrinkage priors offer a natural way of penalization, with certain advantages.

More work is needed to apply Bayesian shrinkage methods in a user-friendly way, solving issues such as prior choice, variable selection, and computational efficiency.

Thank you!

s.j.vanerp@uu.nl
Twitter: @SaravanErp
Github: https://github.com/sara-vanerp
https://saravanerp.com

References

Redmond, M., & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. European Journal of Operational Research, 141(3), 660–678.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. The Statistician, 24(4), 267.

Park, T., & Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482), 681–686.

Li, Q., & Lin, N. (2010). The Bayesian elastic net. Bayesian Analysis, 5(1), 151–170.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819–1829.

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. Australian & New Zealand Journal of Statistics, 53(4), 423–442.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. Biometrika, 97(2), 465–480.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. Journal of the American Statistical Association, 83(404), 1023–1032.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. Journal of the American Statistical Association, 88(423), 881.

Erp, S. van, Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. Journal of Mathematical Psychology, 89, 31–50.

van Lissa, van Erp, & Clapper (under review). Select relevant moderators using Bayesian regularized meta-regression.