ECOM168: Hierarchical shrinkage priors

Daniele Bianchi

School of Economics and Finance,
Queen Mary University of London

Outline

1. Introduction
2. Hierarchical shrinkage priors
3. Global-local shrinkage priors
4. Improving sampling efficiency

Introduction

Introduction

Simple priors such as the Zellner's g-prior are inadequate for learning interesting features and/or quantifying uncertainty in large-scale regression models.

$\hookrightarrow$ not explicitly designed for variables selection.

In a frequentist setting, penalised regression methods have been increasingly popular to guard against over-fitting and to select variables relevant for forecasting.

Penalised regressions

Recall that least squares implies that regression coefficients should minimize

$$\arg\min_{\beta} n^{-1}\|y - X\beta\|_2^2$$

The central idea of penalised regressions is to add a penalty to the main loss function. Thus, a general formulation of penalised regression implies

$$\arg\min_{\beta} n^{-1}\|y - X\beta\|_2^2 + \lambda\|\beta\|_q \qquad \text{with} \qquad \|\beta\|_q = \left(\sum_{j=1}^{p} |\beta_j|^q\right)^{\frac{1}{q}},$$

Goal: shrinking small coefficients toward zero while leaving large coefficients "intact".

Penalised regressions

The parameter λ reflects the strength of the penalty,

$\hookrightarrow$ large values implies more aggressive shrinkage towards zero, while $\lambda = 0$ corresponds to simple least squares estimates.

The choice of q determines the type of penalty induced;

$\hookrightarrow$ E.g., $q = 1$, least absolute shrinkage and selection operator (lasso; Tibshirani, 1996).

$\hookrightarrow$ E.g., $q = 2$, ridge regression (see Hoerl and Kennard, 1970).

See Hastie et al. (2015) for a comprehensive review of various penalised regression methods in a frequentist setting.

Penalised regressions

Thus, a ridge regression can be defined as:

$$\arg\min_{\beta} n^{-1}\|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{1}$$

For every λ, we will get a different estimate of β;

↪ For e.g., for $\lambda \to 0$ we converge to the least squares estimator.

How do we choose λ in a frequentist setting?

↪ Cross-validation, or any other out-of-sample criterion.

Note: the λ treats all the β_j the same, so we have to carefully consider the units of β_js.

↪ Solution: *standardise* the covariates before the estimates.

From penalised regressions to Bayesian inference

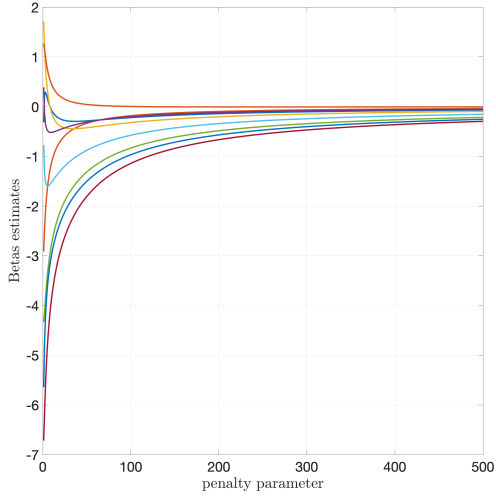
But, what is exactly the estimator $\widehat{\beta}_\lambda$?

Let's take the first order condition of the loss function in Eq.(1),

$$\lambda\beta = X^\top (y - X\beta),$$

such that

$$\widehat{\beta}_\lambda = \left(X^\top X + \lambda I_p\right)^{-1} X^\top y$$



The estimates of βs for different levels of penalty terms λ.

From penalised regressions to Bayesian inference

The estimator $\widehat{\beta}_\lambda$ recalls a standard Jeffrey's prior

$$\beta|\tau^2 \sim N\left(0, \sigma^2\tau I_p\right), \qquad \text{and} \qquad \sigma^2 \sim \frac{1}{\sigma^2},$$

which gives the posterior

$$\beta|\tau^2, \sigma^2, y \sim N\left(A^{-1}X^\top y, \sigma^2 A^{-1}\right),$$

with $A^{-1} = \left(X^\top X + \tau^{-2} I_p\right)^{-1}$. This implies that

$$E\left[\beta|\tau^2, y\right] = \left(X^\top X + \tau^{-2} I_p\right)^{-1} X^\top y,$$

that is equivalent to $\widehat{\beta}_\lambda$ with $\lambda = \tau^{-2}$.

Shrinkage estimation is embedded in Bayesian inference!!

Hierarchical priors

Why Bayesian?

Adopting a Bayesian perspective offers several advantages:

- Shrinkage via conditionally conjugate, hierarchical, priors.
- Parameter uncertainty and hypothesis testing as a by-product of posterior distributions.
- Penalty and model parameters can be simultaneously estimated as part of the same joint conditional distribution.
  - ↪ This is key when penalty parameters are multiple, such as elastic net. Sequential cross-validation often induce too much shrinkage.

Hierarchical priors

Hierarchical priors involve specifying prior distributions for the hyper-parameters, i.e., a multi-level prior structure. For e.g.,

$$y|\beta, \sigma^2 \sim N\left(X\beta, \sigma^2\right) \qquad \beta|\tau^2 \sim N\left(0, \tau^2\right) \qquad \tau^2 \sim \mathcal{D} \tag{2}$$

where $\mathcal{D}$ denotes a general continuous distribution.

Key idea: the *distribution* of τ^2 should be learned from the data;

$\hookrightarrow$ the choice of τ^2 is crucial for the posterior estimate of β

$\hookrightarrow$ hierarchical priors are often referred to as *full-Bayes* priors.

Hierarchical priors

Hierarchical models such as Eq.(2) offer several advantages:

- Easily generalisable to non-linear, non-Gaussian, multivariate models (more on lecture 5).
- Can define *adaptive* shrinkage: $\beta_j|\tau_j^2 \sim N\left(0, \tau_j^2\right)$
- Further layers of the hierarchy can be considered; for instance if the hyper-parameters of $\mathcal{D}$ cannot be easily interpreted, one can postule a prior on those parameters.
- While $\beta_j|\tau^2 \sim N\left(0, \tau^2\right)$, the marginal prior for β_j is not Gaussian

$$p\left(\beta_j\right) = \int p\left(\beta_j|\tau^2\right) p\left(\tau^2\right) d\tau^2,$$

that is, a *scale mixture of normals* representation that allows for complex shapes of β_j. This is crucial for hypothesis testing and decision making.

Hierarchical priors

Assumption: one can draw from a conditionally conjugate posterior distribution by using a hierarchical representation.

Let $D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$, depending on whether we also condition on the regression variance parameter σ^2, we can distinguish between:

- $\beta|\sigma^2, \tau^2 \sim N\left(0, \sigma^2 D\right)$, i.e., natural conjugate prior.
- $\beta|\tau^2 \sim N\left(0, D\right)$, i.e., independent prior.

Hierarchical priors

Posterior conditionals are convenient to sample from. For e.g., consider the case of an independent prior, such that the joint posterior is of the form,

$$p\left(\beta, \sigma^2, \tau^2 | y\right) \propto p\left(y|\beta, \sigma^2\right) p\left(\beta|\tau^2\right) p\left(\tau^2\right) p\left(\sigma^2\right),$$

Since $p\left(\tau^2\right)$ and $p\left(\sigma^2\right)$ are constants give τ^2, the conditional posterior of β takes the form

$$p\left(\beta|\sigma^2, \tau^2, y\right) \propto p\left(y|\beta, \sigma^2\right) p\left(\beta|\tau^2\right),$$

Similar arguments hold for:

- $\hookrightarrow$ $p\left(\sigma^2|\beta, \tau^2, y\right) \propto p\left(y|\beta, \sigma^2\right) p\left(\sigma^2\right)$, i.e., identical to the non-shrinkage case.
- $\hookrightarrow$ $p\left(\tau^2|\beta, \sigma^2, y\right) \propto p\left(\beta|\tau^2\right) p\left(\tau^2\right)$, i.e., the density $p\left(y|\beta, \sigma^2\right)$ does not contain information on τ^2, and is affected only through the density $p\left(\beta|\tau^2\right)$.

Hierarchical priors

The conditional structure of the hierarchical priors implies:

$\hookrightarrow$ common formulations for the conditional posterior of β and σ^2.

$\hookrightarrow$ the conditional posterior of τ^2 will depend on how complicate is $p\left(\tau^2\right)$.

A Gibbs sampler will cycle through these conditional posteriors, obtaining a sample of draws for each of the parameters of interest.

We are going to see how the prior $p\left(\tau^2\right)$ is key for modeling shrinkage.

Hierarchical shrinkage priors

Overview of shrinkage priors

Given the extensive number of shrinkage priors proposed in the literature (see, e.g., Korobilis et al., 2022 for an extensive review), we will limit our focus to:

- ↪ Priors related to frequentist penalised regressions.
- ↪ Priors popular in the statistics/econometrics literature.

For each prior we discuss the hierarchical structure and the main characteristics.

- ↪ Some further computational details will be discussed at the end of the lecture.

Diffuse hierarchical prior

A natural choice for the variance parameter τ^2 in Eq.(2) is a prior distribution that is *diffuse*, or non-informative.

Similar to a g-prior, the choice $p\left(\tau^2\right) \propto \tau^{-2}$ (or equivalently $U\left(0, \infty\right)$), reflect the lack of information about the nature of sparsity patterns in the data.

For instance, one may might want each β_j to be shrunk adaptively, so that the hierarchical shrinkage prior takes the form

$$\beta_j | \tau_j^2 \sim N\left(0, \tau_j^2\right), \qquad p\left(\tau_j^2\right) \propto \frac{1}{\tau_j^2} \qquad p\left(\sigma^2\right) \propto \frac{1}{\sigma^2},$$

Diffuse hierarchical prior

Some comments:

- ↪ While a diffuse prior on τ_j^2 is a first natural attempt towards hierarchical modeling, it will lead to an "improper" posterior (see Morris, 1983).
- ↪ A uniform prior $U(-\infty, \infty)$ on $\log(\tau^2)$ would also not work (see Gelman, 2006).
- ↪ A similar "improper" density can be obtained assuming $\tau^2 \sim \text{IG}(a, b)$ with $a, b \to 0$.
- ↪ Tipping (2001) propose an inverse-Gamma prior on τ^2 and adopts the limit case of $a = b = 10^{-4}$ as the default hyper-parameter choice.
  - ↪ Gelman (2006) argues that the $\text{IG}(a, b)$ prior does not have any proper limiting posterior distribution; simply setting a, b is not a solution.

Bottom line: diffuse priors should NOT be the first choice in high-dimensional or ultra-high-dimensional settings.

Student-t shrinkage prior

While we argued that IG priors are not ideal as a way to impose a diffuse prior on τ^2, informative IG priors provide flexible parametric shrinkage.

For instance, Armagan and Zaretzki (2010) propose a hierarchical prior of the form,

$$\beta|\sigma^2, D \sim N\left(0, \sigma^2 D\right), \qquad \text{with} \qquad D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right) \tag{3}$$

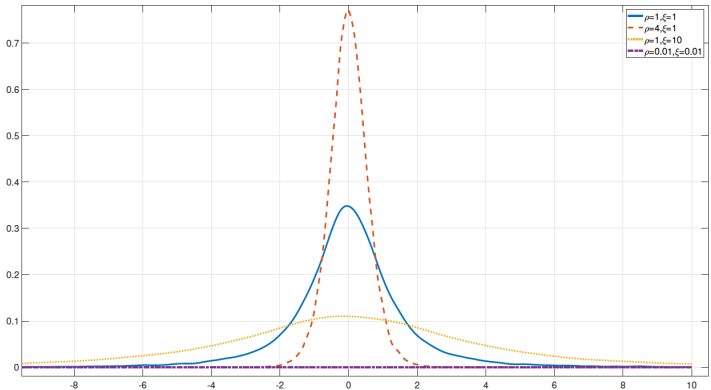
and

$$\tau_j^2 \sim IG\left(a, b\right), \qquad j = 1, \ldots, p, \qquad \sigma^2 \sim \frac{1}{\sigma^2}$$

This is a scale mixture of normals representation of the fat-tailed Student-t distribution.

Key advantages: (1) parameters are shrunk at zero more aggressively compared to the simple normal distribution, and (2) the fat tails accommodate the possibility of very large values of β_j.

Student-t shrinkage prior



The shape of the marginal distribution of β_j for various values of the hyper-parameters of the inverse-Gamma. Source: Korobilis et al. (2022).

Student-t shrinkage posterior

The prior conjugancy offer convenient numerical alternatives for the estimation. The conditional posteriors are of the form,

$$\beta|D,\sigma,y \sim N\left(A^{-1}X^{\top}y, \sigma^2 A^{-1}\right) \qquad \tau_j^2|\beta_j, y \sim IG\left(a+0.5, b+\frac{\beta_j^2}{2\sigma^2}\right)$$

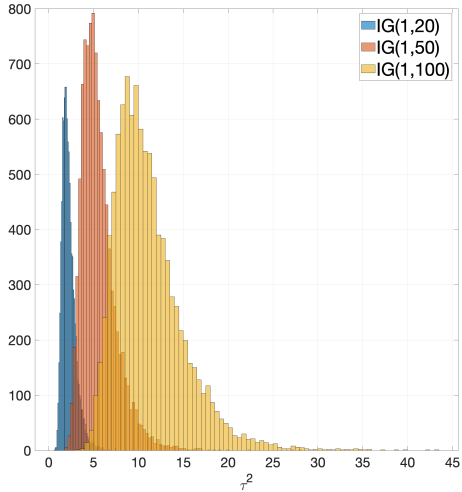
and

$$\sigma^2|\beta, y \sim IG\left(\frac{s_0+p}{2}, \frac{s+\beta^{\top}D^{-1}\beta}{2}\right),$$

where $A = \left(X^{\top}X + D^{-1}\right)^{-1}$, $s = (y-X\beta)^{\top}(y-X\beta)$, and $D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$.

A standard Gibbs sampler can be used to sequentially sample from these conditional posterior distributions.

Shrinkage mechanism



A larger β_j^2 implies a larger τ^2, which in turns implies less shrinkage. That is, shrinkage is inversely proportional to the magnitude of the regression coefficients (and their uncertainty).

Bayesian lasso

The least absolute shrinkage and selection operator (lasso) of Tibshirani (1996) has been established as a key workhorse when working with high-dimensional regression models.

In its frequentist formulation, the estimator takes the form

$$\arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \qquad \text{where} \qquad \|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|,$$

λ is a tuning parameter that controls how shrinkage is exercised.

$\hookrightarrow$ for $\lambda \to 0$ the penalty vanishes and the lasso converges to a least squares estimator.

Bayesian lasso

Tibshirani (1996) itself was the first noting that the lasso estimate can be interpreted as the posterior mode under a specific Laplace prior distribution of the form

$$p(\beta) = \prod_{j=1}^{p} \frac{\lambda}{2} \exp(\lambda|\beta_j|), \tag{4}$$

Caveat: the posterior distribution under the Laplace prior in Eq.(4) makes the uncertainty quantification using the Bayesian lasso unreliable.

↪ The coefficient λ needs to be large enough to penalise β_j to zero, but not too large such that nonzero coefficients can still be modelled.

↪ A hierarchical prior structure can certainly help here.

Bayesian lasso

The Laplace prior admits a hierarchical representation in the form of a double exponential mixture. This implies that a hierarchical representation of this prior is of the form

$$\beta_j|\tau_j \sim N\left(0, \tau_j^2\right), \qquad \tau_j^2|\lambda^2 \sim Exp\left(\frac{\lambda^2}{2}\right) = \left(\frac{\lambda^2}{2}\right) \exp\left(\frac{\lambda^2}{2}\tau_j^2\right)$$

Conditional on λ, the marginal distribution of β_j takes the desired form

$$p\left(\beta_j|\lambda^2\right) = \frac{\sqrt{\lambda}^2}{2} \exp\left(\sqrt{\lambda^2}|\beta_j|\right) \approx \frac{\lambda}{2} \exp\left(\lambda|\beta_j|\right) \tag{5}$$

Bayesian lasso

A formal Bayesian treatment of the Bayesian lasso can be found in Park and Casella (2008). They express the Bayesian lasso as a normal-exponential mixture, but conditional on the regression variance σ^2.

$\hookrightarrow$ A hierarchical prior of β_j *independent* of σ^2 would result in a multimodal posterior for β_j.

The Park and Casella (2008) Laplace prior takes the form

$$\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim N\left(0, \sigma^2 D\right), \qquad \tau_j^2|\lambda^2 \sim Exp\left(\frac{\lambda^2}{2}\right), \qquad \lambda^2 \sim G\left(a, b\right),$$

with $p\left(\sigma^2\right) \propto 1/\sigma^2$ and $D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$.

Bayesian lasso

Given this hierarchical prior structure, Park and Casella (2008) shown that the conditional posteriors can be derived as

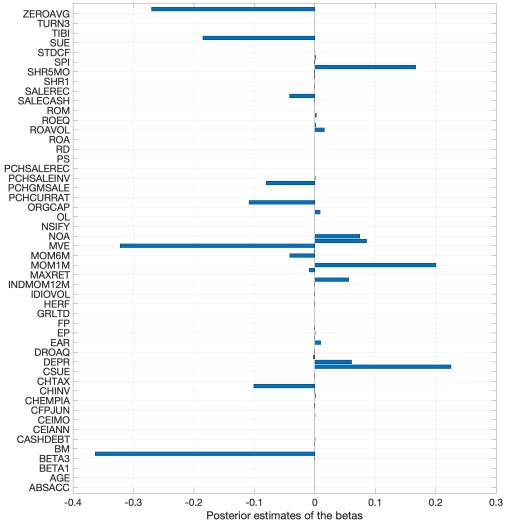
$$\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2, y \sim N\left(A^{-1}X^\top y, \sigma^2 A^{-1}\right), \qquad \tau_j^2|\beta_j, y \sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right),$$

and

$$\lambda^2|\tau_1^2, \ldots, \tau_p^2 \sim G\left(a+p, b+\frac{\sum_{j=1}^p \tau_j^2}{2}\right), \quad \sigma^2|\beta, y \sim IG\left(\frac{s_0+p}{2}, \frac{s+\beta^\top D^{-1}\beta}{2}\right), \quad (6)$$

where $A = \left(X^\top X + D^{-1}\right)^{-1}$, $s = (y - X\beta)^\top (y - X\beta)$, and $D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$.

Example: Anomalies and the expected market returns



Posterior estimates from the Hierarchical Bayesian lasso prior. The estimates are based on the full sample of observations from 1970 to 2017, monthly.

Extensions of the Bayesian lasso

Bayesian group lasso

If there is a group of covariates among which the pairwise correlation is high, the lasso tends to select only individual variables from the group.

Yuan and Lin (2006) introduced the *group* lasso which takes such group structure into account,

$$\widehat{\beta} = \arg\min_{\beta} (y - X\beta)^{\top} (y - X\beta) + \lambda \sum_{j=1}^{K} (\beta' G_k \beta)^{1/2}$$

with G_k a positive semi-definite matrix (typically $G_k = I_{m_k}$ with m_k is the size of the coefficient vector in group k).

Such penalty allows to do variable selection at the group level and is invariant under orthogonal transformations of the covariates.

Extensions of the Bayesian lasso

Kyung et al. (2010) proposed a Bayesian group lasso with the following conditional prior

$$p\left(\beta|\sigma^2\right) \propto \exp\left(-\frac{\lambda}{\sigma}\sum_{k=1}^{K}\left(\beta_{G_k}^{\top}\beta_{G_k}\right)^{1/2}\right)$$

This is equivalent to the following gamma mixture of Normals,

$$\beta_{G_k}|\tau_k^2,\sigma^2 \sim N\left(0,\sigma^2\tau_k^2 I_{m_k}\right) \qquad \text{and} \qquad \tau_k^2 \sim G\left(\frac{m_k+1}{2},\frac{\lambda^2}{2}\right)$$

with β_{G_k} the vector of β_js in group $k = 1,\ldots,K$.

Extensions of the Bayesian lasso

The conditional posteriors are of the form,

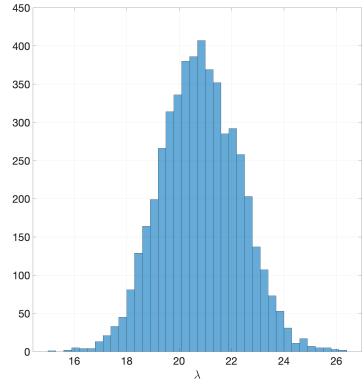
$$\beta_{G_k}|\beta_{-G_k},\sigma^2,\tau_1^2,\ldots,\tau_K^2,\lambda,y \sim N\left(A_k^{-1}X_k\left(y-\frac{1}{2}\sum_{k'\neq k}X_{k'}\beta_{G_k'}\right),\sigma^2A_k^{-1}\right),$$

where $\beta_{-G_k}=\left(\beta_{G_1},\ldots,\beta_{G_{k-1}},\beta_{G_{k+1}},\ldots,\beta_{G_K}\right)$ and $A_k=X_k'X_k+\tau_k^{-2}I_{m_k}$.

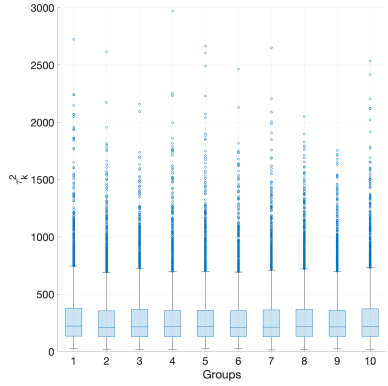
$$1/\tau_k^2|\ldots\sim IG\left(\sqrt{\frac{\lambda^2\sigma^2}{\|\beta_{G_k}\|^2}},\lambda^2\right)\mathbb{I}\left(\tau_k^2>0\right)\qquad \lambda^2|\ldots\sim G\left(\frac{p+K}{2}+a,\frac{1}{2}\sum_{k=1}^{K}\tau_k^2+b\right),$$

$$\sigma^2|\ldots\sim IG\left(\frac{n-1+p}{2},\frac{1}{2}\|y-X\beta\|^2+\frac{1}{2}\sum_{k=1}^{K}\frac{1}{\tau_k^2}\|\beta_{G_k}\|^2\right),$$

Example: Anomalies and the expected market returns



(a) λ



(b) τ_k^2

The groups have been randomly created. So it makes sense there is not much heterogeneity in the group-specific shrinkage. Overall, there is lots of sparsity as implied by λ.

Extensions of the Bayesian lasso

Bayesian adaptive lasso

One drawback of the lasso is that it can performs automatic variable selection but it produces biased estimates for the larger coefficients.

$\hookrightarrow$ the oracle property in the sense of Fan and Li (2001) does not hold.

To obtain the oracle property, Zou (2006) introduced the Bayesian adaptive lasso estimator as

$$\widehat{\beta} = \arg\min_{\beta} (y - X\beta)^{\top} (y - X\beta) + \sum_{j=1}^{p} \lambda_j |\beta_j|,$$

with $\lambda_j = \lambda |\widehat{\beta}_j|^{-r}$ for $j = 1, \ldots, p$ for $\widehat{\beta}_j$ is a $\sqrt{n}$ consistent estimator such as least squares.

Extensions of the Bayesian lasso

Leng et al. (2014) proposed a Bayesian version of the adaptive lasso.

$\hookrightarrow$ key idea: Laplace distribution as a scale mixture of Normals with exponential mixing.

Proposal: replace the conditional prior for $\tau_j^2|\lambda$ as in Eq.(6) with

$$\tau_j|\lambda_j \sim Exp\left(\frac{\lambda_j^2}{2}\right),$$

We allow different λ_j, one for each coefficient. Small penalties are applied to importance covariates while large penalties are applied to those which are unimportant.

One can show that the prior of β conditional on σ^2 is

$$p\left(\beta|\sigma^2\right) = \prod_{j=1}^{p} \frac{\lambda_j}{2\sqrt{\sigma^2}} \exp\left(-\lambda_j|\beta_j|/\sqrt{\sigma^2}\right),$$

Global-local shrinkage priors

Global-local shrinkage priors

Consider a standard linear regression model

$$y = X\beta + \epsilon$$

Broadly we can assume a scale mixture of normals formulation for the prior distribution of the regression parameters,

$$\beta \sim N(0, D), \qquad \text{with} \qquad D = \text{diag}\left(\lambda_1^2, \ldots, \lambda_p^2\right), \tag{7}$$

and then $\lambda_i^2 \sim \mathcal{G}$ for some distribution $\mathcal{G}$.

We have seen so far how the properties of the posterior distribution are determined by the choice of the hyper-prior on $\mathcal{G}$.

Global-local shrinkage priors

Notably, $\mathcal{G}$ can involve both local and global hyper-parameters.

$\hookrightarrow$ Each β_j has its own idiosyncratic scale λ_j^2.

$\hookrightarrow$ Also a global (across all β_js) scale parameter τ.

Thus, a conjugate global-local prior for the regression coefficient $\beta_j,\ j = 1, \ldots, p$ can be written as

$$\beta_j \sim N\left(0, \sigma^2 \lambda_j^2 \tau^2\right) \qquad \lambda_j \sim \mathcal{D} \qquad \tau \sim \mathcal{F}$$

where $\mathcal{F}$ and $\mathcal{D}$ are general probability densities with positive support, i.e., $\lambda_j \geq 0, j = 1, \ldots, p$ and $\tau \geq 0$.

Horseshoe prior

Carvalho et al. (2010) has been one of the first to introduce sparsity-inducing "global-local" priors as scale mixture of Normals in linear regression models.

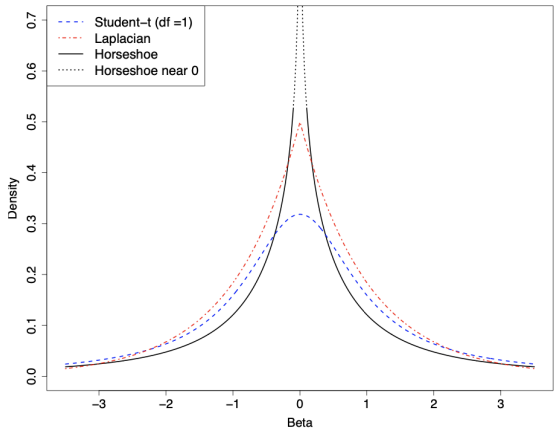
They introduced the so-called *horseshoe* prior, which can be represented as a scale mixture of Normals with half-Cauchy mixing distributions,

$$\beta|\lambda_1,\ldots,\lambda_p,\tau \sim N\left(0,\sigma^2\tau^2\Lambda\right), \quad \tau \sim C^+\left(0,1\right), \quad \lambda_j|\tau \sim C^+\left(0,1\right), \quad j=1,\ldots,p$$

where $\Lambda = \text{diag}\left(\lambda_1^2,\ldots,\lambda_p^2\right)$ and $C^+\left(0,\alpha\right)$ is the half-Cauchy distribution on the positive real axis with scale parameter $\alpha = 1$, such that,

$$\lambda_j|\tau = \frac{2}{\pi\tau\left(1+\left(\lambda_j/\tau\right)^2\right)}$$

Horseshoe prior



The shape of the horseshoe prior vs Student-t and Laplacian.

Horseshoe prior

Under this hierarchical specification, the marginal prior for each β_j is unbounded at the origin and has tails that decay polynomially, not exponentially.

Two features which makes it one of the "gold standards" among hierarchical shrinkage priors:

- ↪ Its flat, Cauchy-like tails allow strong signals to remain large *a posteriori*.
- ↪ The large spike at the origin aggressively shrink to zero "weak signals", that is elements of β which are non-significant.
- ↪ N.B., the horseshoe could "over-shrink" in the context of low signal-to-noise ratios.

Horseshoe prior

There are various computational approaches to the horseshoe prior, but the most straightforward (in my opinion) is the one proposed by Makalic and Schmidt (2015)

Key idea: the half-Cauchy distribution can be written as a mixture of inverse-Gamma distributions. In particular they observed that, if

$$x^2|z \sim (1/2, 1/z)\,, \qquad z \sim IG\left(1/2, 1/\alpha^2\right)$$

then the marginal $x \sim C^+(0, \alpha)$.

Horseshoe prior

Makalic and Schmidt (2015) proposed a hierarchical specification of the horseshoe prior as

$$\beta|\lambda_1,\ldots,\lambda_p,\tau,\sigma^2 \sim N\left(0,\sigma^2\tau^2\Lambda\right), \qquad \lambda_j^2|\nu_j \sim IG\left(1/2,1/\nu_j\right), \qquad \nu_j \sim IG\left(1/2,1\right)$$

$$\tau^2|\xi \sim IG\left(1/2,1/\xi\right), \qquad \xi \sim IG\left(1/2,1\right)$$

where $\Lambda = \text{diag}\left(\lambda_1^2,\ldots,\lambda_p^2\right)$ and $p\left(\sigma^2\right) \propto \sigma^{-2}$.

Horseshoe posterior

So that the conditional posterior take the form,

$$\beta|\ldots \sim N\left(A^{-1}X^{\top}y, \sigma^2 A^{-1}\right),$$

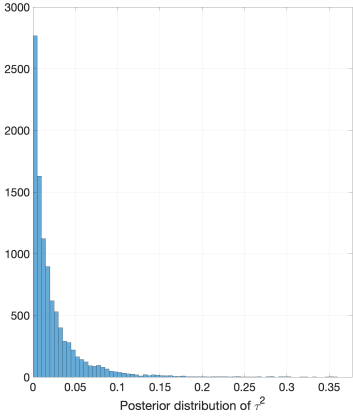
$$\lambda_j^2|\ldots \sim IG\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \qquad \nu_j|\ldots \sim IG\left(1, 1 + \frac{1}{\lambda_j^2}\right), \qquad j = 1, \ldots, p$$

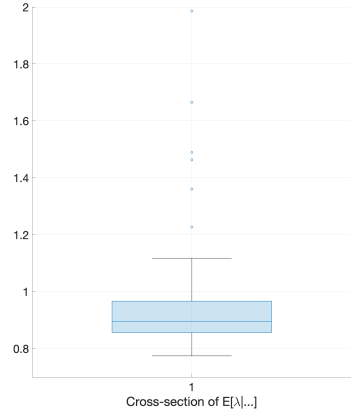
$$\tau^2|\ldots \sim IG\left(\frac{1+p}{2}, \frac{1}{\xi} + \frac{1}{2\sigma^2}\sum_{j=1}^{p}\frac{\beta_j^2}{\lambda_j^2}\right),$$

$$\xi|\ldots \sim IG\left(1, 1 + \frac{1}{\tau^2}\right) \qquad \sigma^2|\ldots \sim IG\left(\frac{n+p}{2}, \frac{s + \beta^{\top}D_{\lambda}^{-1}\beta}{2}\right),$$

where $A = \left(X^{\top}X + D_{\lambda}^{-1}\right)^{-1}$, $s = (y - X\beta)^{\top}(y - X\beta)$, and $D_{\lambda} = \text{diag}\left(\tau^2\lambda_1^2, \ldots, \tau^2\lambda_p^2\right)$.

Example: Anomalies and the expected market returns





The posterior distribution of the global shrinkage parameter λ (left panel), and the cross-sectional distribution of the posterior mean of the local shrinkage parameters τ_i^2 (right panel).

Normal-gamma prior

Brown and Griffin (2010) propose a prior where the distribution $\mathcal{G}$ in Eq.(7) has a gamma density of the form

$$\lambda_i^2 \sim G(\lambda, 1), \qquad \text{with} \qquad G(x|, \lambda, b) \propto x^{\lambda-1} \exp(-b)$$

It forms a natural extension of the lasso which implicitly applies an exponential prior.

Few key characteristics:

- ↪ Similar to horseshoe, it is suitable for long-tailed distributions (finance data).
- ↪ Differently from horseshoe, tails are exponential and not polynomial. That is, tails are heavier in the horseshoe prior, which makes it suitable for more extreme signals.

Normal-gamma prior

The normal-gamma prior of Brown and Griffin (2010) takes the form,

$$\beta|\tau_1,\ldots,\tau_p \sim N\left(0, D\right), \qquad \tau|\lambda,\gamma^2 \sim G\left(\lambda, \frac{1}{2\gamma^2}\right), \qquad \sigma^2 \sim \frac{1}{\sigma^2}$$

where $D = \text{diag}\left(\tau_1^2,\ldots,\tau_p^2\right)$.

As usual, we consider a Jeffrey's prior for $\sigma^2 \propto \sigma^{-2}$.

$\hookrightarrow$ The error variance can also be incorporated in the form of a conjugate prior (see, e.g., Van Der Pas et al., 2016).

Normal-gamma prior

The conditional posteriors for β and σ^2 are of the usual form

$$\beta|\tau_1,\ldots,\tau_p,y \sim N\left(\left(\sigma^2 A^{-1}\right)X^\top y, A^{-1}\right),$$
$$\sigma^2|y \sim IG\left(\frac{n}{2},\frac{1}{2}\left(y-X\beta\right)^\top\left(y-X\beta\right)\right)$$

where $A^{-1}=\left(X^\top X/\sigma^2+D^{-1}\right)^{-1}$.

The parameters $\tau_j, j=1,\ldots,p$ can be updated in a block since the full conditional are independent,

$$\tau_j|\beta_j \sim GIG\left(\lambda-\frac{1}{2},\frac{1}{\gamma^2},\beta_j^2\right), \qquad j=1,\ldots,p$$

with $GIG(\cdot)$ a generalised inverse Gaussian distribution.

Dirichlet-Laplace prior

Bhattacharya et al. (2015) expanded the existing set of hierarchical shrinkage methods by proposing a new class of Dirichlet-Laplace (DL) priors,

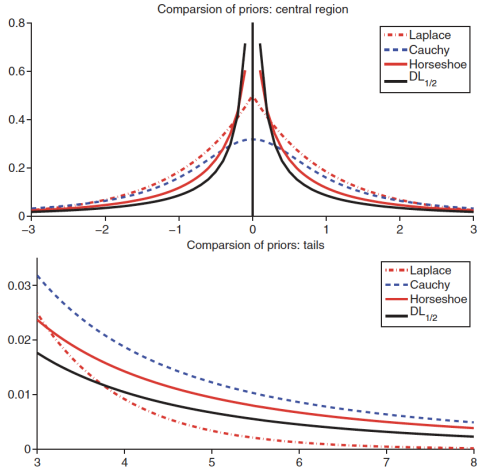
$\hookrightarrow$ optimal posterior concentration + efficient posterior computation.

The underlying idea of the DL prior is to modify the Bayesian lasso prior by chopping up the regression parameter into random chunks as determined by a Dirichlet distribution

$$\beta_j|\phi,\lambda \sim DE\left(\phi,\lambda\right), \qquad \phi \sim Dir\left(\alpha,\ldots,\alpha\right), \qquad \lambda \sim G\left(n\alpha,1/2\right)$$

where $DE\left(\phi_j,\tau\right)$ is a double exponential (i.e., Laplace) distribution, $Dir\left(\alpha,\ldots,\alpha\right)$ is a Dirichlet distribution over the simplex with equal (pre-specified) parameter values.

Dirichlet-Laplace prior



Marginal density of the DL prior with $\alpha = 1/2$ in comparison to other shrinkage priors.

Dirichlet-Laplace prior

In its simplest form, the DL hierarchical prior is a generalisation of the Laplace prior,

$$\beta|\tau_1^2,\ldots,\tau_p^2,\phi,\lambda,\sigma^2 \sim N\left(0, D_{\lambda,\tau,\phi}\right), \qquad \tau_j^2 \sim Exp\left(1/2\right),$$
$$\phi \sim Dir\left(\alpha,\ldots,\alpha\right), \qquad \lambda \sim G\left(n\alpha, 1/2\right), \qquad \sigma^2 \sim \frac{1}{\sigma^2},$$

where $D_{\lambda,\tau,\phi} = \text{diag}\left(\lambda^2\tau_1^2\phi_1^2,\ldots,\lambda^2\tau_p^2\phi_p^2\right)$.

Dirichlet-Laplace posterior

The conditional posteriors can be sampled from via Gibbs sampling and take the form,

$$\beta|\ldots \sim N(\left(A^{-1}X^{\top}y, \sigma^2 A^{-1}\right), \qquad \lambda \sim GIG\left(2\frac{\sum_{j=1}^{2}|\beta_j|}{\phi_j \sigma}, 1, p\left(\alpha-1\right)\right),$$

$$\phi_j|\ldots = \frac{T_j}{\sum_{j=1}^{p} T_j}, \qquad \text{with} \qquad T_j \sim GIG\left(2\sqrt{\frac{\beta_j^2}{\sigma^2}}, 1, \alpha-1\right),$$

$$\frac{1}{\sigma^2}|\ldots \sim G\left(a, b\right), \qquad \frac{1}{\tau_j^2}|\ldots \sim IG\left(c, 1\right),$$

where $a = \left(n+p\right)/2$, $b = \left(s + \beta^{\top} D_{\lambda,\tau,\phi}\beta\right)/2$, $c = \sqrt{\lambda^2\phi_j^2\sigma^2/\beta_j^2}$ the parameters if the Inverse-Gaussian of the scale and the local-shrinkage parameters, and $A^{-1} = \left(X^{\top}X + D_{\lambda,\tau,\phi}^{-1}\right)^{-1}$.

Improving sampling efficiency

Some computational trick

For p large, the most computational intensive step is to draw $\beta| \ldots \sim N(\mu, \Sigma)$,

For instance, the posterior mean from the horseshoe prior is

$$\beta|\sigma^2, \tau_1^2, \ldots, \tau_p^2, y \sim N\left(A^{-1}X^\top y, \sigma^2 A^{-1}\right) \qquad \text{where} \qquad A = \left(X^\top X + D^{-1}\right)^{-1},$$

and $D = \text{diag}\left(\tau_1^2, \ldots, \tau_p^2\right)$. This step involves inverting a potentially really large-dimensional covariance matrix A^{-1}.

As the dimension p becomes large, matrix inversion and Cholesky factorizations can get quite complicated/numerically cumbersome.

Some computational trick

Rue (2001) provides a fast, and accurate, sampling method for the covariance matrix A^{-1}. The procedure takes the following steps:

1. Factorize $A^{-1} = LL^{\top}$, with L a lower-triangular matrix (Cholesky factorization).
2. Generate $Z \sim N(0, I)$.
3. Set $\nu = L^{-1}X^{\top}y$, $\mu = \left(L^{\top}\right)^{-1}\nu$, and $u = \left(L^{\top}\right)^{-1}Z$.
4. Generate $\beta = \mu + u$.

It follows that $E(\beta) = E(\mu + u) = E(\mu) = \left(L^{\top}\right)^{-1}L^{-1}X'y = \left(L^{\top}L\right)^{-1}X'y = A^{-1}X^{\top}y$, and $Var(\beta) = Var(\mu + u) = \left(L^{\top}\right)^{-1}IL^{-1} = A^{-1}$.

The main advantage is that we invert the Cholesky factor L instead of A^{-1}.

Some computational trick

Scalable Gibbs sampling:

In order to speed up the posterior sampling, Rajaratnam et al. (2019) propose to collapse the sampling of $\beta|\ldots$, and $\sigma^2|\ldots$ in a single block.

$\hookrightarrow$ main advantage: less correlation between draws = higher efficiency.

For a standard hierarchical prior formulation like the ones we have seen above, a *scalable* Gibbs takes the form:

$$\sigma^{-2}|\tau^2, y \sim G\left(\frac{n-1}{2}, y^\top \left(I_n - XA^{-1}X^\top\right) y/2\right),$$
$$\beta|\sigma^2, \tau^2, y \sim N\left(A^{-1}X^\top y, A^{-1}\right)$$

with $p\left(\tau^2|\beta, \sigma^2, y\right)$ depending on the type of hierarchical shrinkage prior considered.

References

Armagan, A. and Zaretzki, R. L. (2010). Model selection via adaptive shrinkage with t priors. *Computational Statistics*, 25:441–461.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Brown, P. J. and Griffin, J. E. (2010). Inference with normal-gamma prior distributions in regression problems.

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

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143:143.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Korobilis, D., Shimizu, K., et al. (2022). Bayesian approaches to shrinkage and sparse estimation. *Foundations and Trends® in Econometrics*, 11(4):230–354.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

Leng, C., Tran, M.-N., and Nott, D. (2014). Bayesian adaptive lasso. *Annals of the Institute of Statistical Mathematics*, 66:221–244.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Morris, C. N. (1983). Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381):47–55.

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

Rajaratnam, B., Sparks, D., Khare, K., and Zhang, L. (2019). Uncertainty quantification for modern high-dimensional regression via scalable bayesian methods. *Journal of Computational and Graphical Statistics*, 28(1):174–184.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

Van Der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). Conditions for posterior contraction in the sparse normal means problem.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.