07b - Regularization & Sparsity

Bayesian Statistics

Spring 2022-2023

Josep Fortiana

Matemàtiques - Informàtica UB

Monday, April 24, 2023

07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

Bias-variance tradeoff

A general principle when several models can describe the same data.

If the model is enlarged (more parameters, more complexity) to fit better the observed data *(less bias)*, then it becomes unstable *(more variance)*.

A model with large variance will be a worse fit to different data sets from the same population; predicions will be unreliable.

Example: polynomial regression

Data: Pairs (y_i, x_i), y_i: response,
x_i: predictors.

Least squares adjustment:

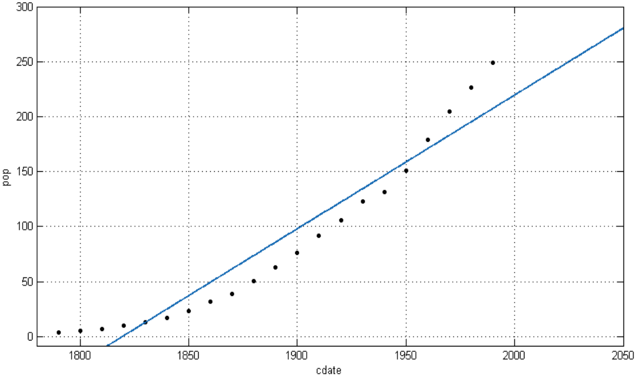
- Linear regression $y = a + b\,x$. Dim 2.
- Quadratic regression $y = a + b_1\,x + b_2\,x^2$. Dim 3.

  $\vdots$
- Polynomial, deg. k $y = a + b_1\,x + b_2\,x^2 + \cdots + b_k\,x^k$. Dim k+1.

Larger degree, more instability.

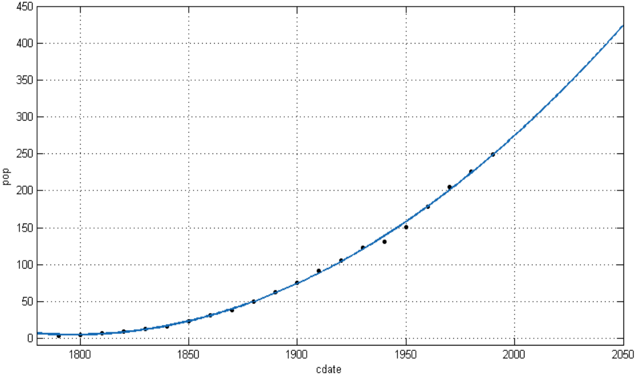
US population 1790 – 1990. Prediction for 2050

Linear regression



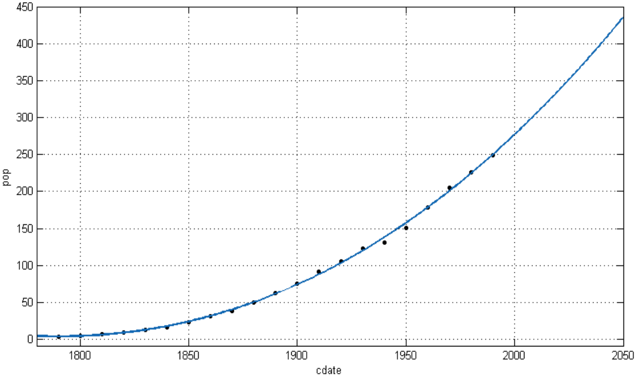
US population 1790 – 1990. Prediction for 2050

Quadratic



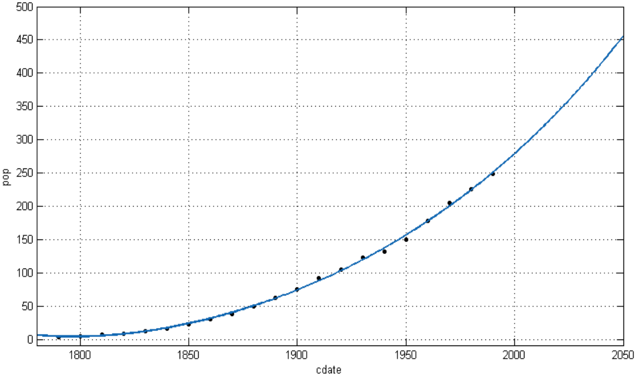
US population 1790 – 1990. Prediction for 2050

Degree = 3



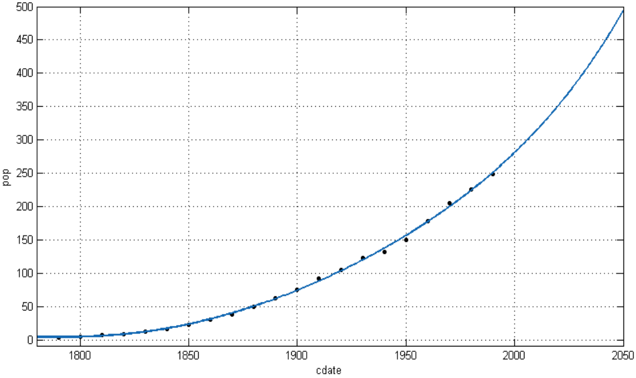
US population 1790 – 1990. Prediction for 2050

Degree = 4



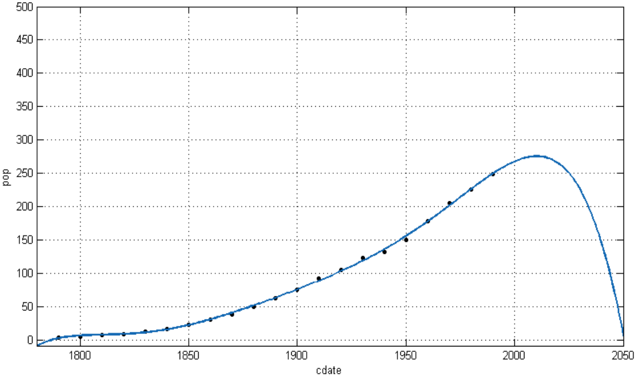
US population 1790 – 1990. Prediction for 2050

Degree = 5



US population 1790 – 1990. Prediction for 2050

Degree = 6



07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

Linear model instability

A linear model, with independent observations with equal variance (Gauss-Markov condition),

$$\boldsymbol{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

is a random vector with *n* observations of a *response variable.*

Linear model instability

The *model matrix* $$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

contains the constant, known values, of the p *predictors*.

The vector: $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$ contains the *random errors.*

Linear model instability

The $p \times 1$ vector of parameters,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

contains the regression coefficients.

Setting $\mathrm{E}(y) = \boldsymbol{X} \cdot \boldsymbol{\beta}$, the ϵ_i are i.i.d. $\sim (0, \sigma^2)$.

Linear model instability

The classical *(OLS, Ordinary Least Squares)* estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is a solution of the optimization problem, of minimizing:

$$F(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2.$$

When $p < n$ and $\operatorname{rank}(\boldsymbol{X}) = p$, there exists a unique solution:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y}.$$

Linear model instability

In this case, the *fitted values vector* is:

$$\hat{y} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}} = \boldsymbol{H} \cdot \boldsymbol{y},$$

and the *residuals vector*:

$$\tilde{y} = \boldsymbol{y} - \hat{y},$$

where $\boldsymbol{H}$, the *hat matrix*, is the orthogonal projector on the linear subspace $\langle \boldsymbol{X} \rangle \subset \mathbb{R}^n$, is given by:

$$\boldsymbol{H} = \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}'.$$

Linear model instability

Even when there is not a unique solution, and:

$$\boldsymbol{Q} = \boldsymbol{X}' \cdot \boldsymbol{X}$$

is singular, the subspace $\langle \boldsymbol{X} \rangle \subset \mathbb{R}^n$ is well defined and so is $\boldsymbol{H}$, its uniquely defined orthogonal projector.

If $\mathrm{Var}(\boldsymbol{y}) = \sigma^2\, \boldsymbol{I}$ (Gauss-Markov condition), then:

$$\mathrm{Var}(\hat{\boldsymbol{y}}) = \sigma^2\, \boldsymbol{H},$$

as $\boldsymbol{H}$ is an idempotent matrix.

Linear model instability

When $\boldsymbol{Q} = \boldsymbol{X}' \cdot \boldsymbol{X}$ is nonsingular,

$$\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \, \boldsymbol{Q}^{-1}.$$

What happens when $\boldsymbol{Q}$ is close to being singular?

More generally, when the *condition number* of $\boldsymbol{Q}$ (or $\boldsymbol{X}$) is too large?

Ridge regression

Ridge regression is a method of finding an intently biased estimator $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$, having a smaller variance, i.e., a more stable estimator.

Solution of the minimization problem:

$$F_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda\, \|\boldsymbol{\beta}\|^2,$$

where $\lambda > 0$ is *the regularization parameter*, to be chosen.

This is a *penalized least squares* problem,

a *Tikhonov regularization* of an *ill-posed problem*.

Ridge regression

After computations:

$$\hat{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}' \cdot \boldsymbol{X} + \lambda\, \boldsymbol{I})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y}.$$

Choosing a sufficiently large λ, we can get a non-singular:

$$\boldsymbol{Q}_\lambda = \boldsymbol{X}' \cdot \boldsymbol{X} + \lambda\, \boldsymbol{I}$$

so that the variance of $\hat{\boldsymbol{\beta}}_\lambda$ is acceptable, at the cost of adding bias.

The *Ridge* hat-matrix

$$\hat{y} = \boldsymbol{X} \cdot \hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X} + \lambda\, \boldsymbol{I})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y} = \boldsymbol{H}_\lambda \cdot \boldsymbol{y}.$$

By analogy with the OLS model,

$$\boldsymbol{H}_\lambda = \boldsymbol{X} \cdot (\boldsymbol{X}' \cdot \boldsymbol{X} + \lambda\, \boldsymbol{I})^{-1} \cdot \boldsymbol{X}' \quad \text{is called the } \textit{Ridge} \text{ hat-matrix.}$$

It is *not* an idempotent matrix (i.e., not an orthogonal projector).

Anyhow,

$$\mathrm{df}(\lambda) = \mathrm{tr}(\boldsymbol{H}_\lambda),$$

is the *equivalent number of degrees of freedom* of the model.

The *LASSO*

LASSO is the acronym of *Least Absolute Shrinkage and Selection Operator.*

Statisticians are not above word playing - A close antecedent of this method, by Leo Breiman (1995), is called “garrote".

Like ridge regression Lasso gives an $\underline{\text{intently biased}}$ estimator $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$, having a smaller variance, i.e., a more stable estimator.

Optimization

We want to minimize the sum of squares:

$$\|y - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2,$$

subject to a constraint on the l^1 norm of the regression coefficients, instead of the l^2 norm in ridge regression:

$$\|\boldsymbol{\beta}\| = t,$$

for some fixed $t > 0$.

Lagrange multiplier optimization

As in the ridge case, this is equivalent to solving the *penalized minimization* problem:

$$F_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda \, \|\boldsymbol{\beta}\|, \qquad (\star)$$

$\lambda > 0$ is *the regularization parameter*, to be chosen.

Unintended (?) consequences

Substituting l^1 for l^2 in the constraint might seem a purely formal generalization.

Nothing further from the truth.

The Lasso has a *variable selection* functionality, which did not appear at all in ridge regression.

Sparsity: Shrink redundant parameters to zero

Usual *shrinkage* feature:

When the regularization parameter λ increases,
the norm $\|\boldsymbol{\beta}\|$ of the regression coefficients decreases.

New here:

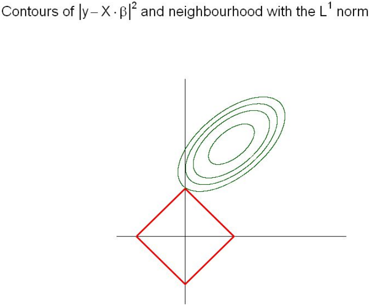
Some β_j, corresponding to irrelevant predictor variables,
actually shrink to 0, yielding an optimal predictor subset.

When does this Lasso variable selection work?

Precisely when it is most useful:

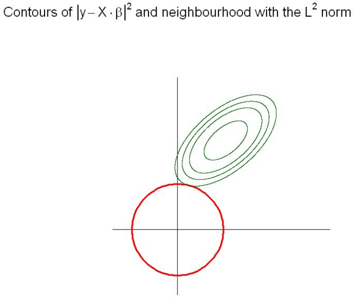
- Large number of predictors (big data)
- *Sparsity,* just a fraction of them are good predictors.

Why does this Lasso variable selection work?


Contours of $|y - X \cdot \beta|^2$ and neighbourhood with the L^1 norm

With the L^1 norm neighbourhoods of zero in the β space have extremal points on the axes (one coordinate is zero).

Why does this Lasso variable selection work?



Contours of $|y - X \cdot \beta|^2$ and neighbourhood with the L^2 norm

With the L^2 norm neighbourhoods of zero in the β space are circular. The optimal point will have a small value in a given coordinate, not zero.

When does the Lasso fail?

Gabriel Vasconcelos - R-bloggers - June 14, 2017.

Generalizations

Elastic net (*GLMnet*). Minimize:

$$\| \boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta} \|^2 + \lambda \left[(1-\alpha)||\boldsymbol{\beta}||_2^2/2 + \alpha||\boldsymbol{\beta}||_1 \right], \quad \alpha \in (0,1).$$

Bridge regression. Minimize:

$$\|\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^\gamma, \quad \gamma > 0.$$

07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

Model

Normal linear (Gauss-Markov) model,

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$\boldsymbol{X} : n \times (p+1)$, with a first column of ones;

$\boldsymbol{\beta} : (p+1) \times 1; \quad \boldsymbol{y}, \boldsymbol{\epsilon}, \boldsymbol{\mu} = \boldsymbol{X} \cdot \boldsymbol{\beta}$, are $n \times 1$.

$$(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) \sim \textbf{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \qquad \boldsymbol{\Sigma} = \sigma^2 \, \boldsymbol{I}_n.$$

Likelihood

A multivariate Gaussian pdf:

$$f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{2\,\pi\,\sigma^2} \right)^{n/2} \cdot \exp \left\{ -\frac{1}{2\,\sigma^2}\, (\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta})' \cdot (\boldsymbol{y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) \right\}.$$

The Normal-IG conjugate prior family

Joint prior pdf:

$$\begin{array}{rcl} h(\boldsymbol{\beta}, \sigma^2) & = & h(\boldsymbol{\beta} \,|\, \sigma^2) \cdot h(\sigma^2) \\ (\boldsymbol{\beta} \,|\, \sigma^2) & \sim & \textbf{Normal}(\boldsymbol{b}, \sigma^2\, \boldsymbol{B}), \\ \sigma^2 & \sim & \mathrm{IG}(\alpha, \beta), \qquad \alpha, \beta > 0. \end{array}$$

$\boldsymbol{B}$: $p \times p$ symmetric, positive definite, $\boldsymbol{b}$: $p \times 1$.

Usually $\boldsymbol{B} = (1/\lambda)\, \boldsymbol{I}$ and $\boldsymbol{b} = \mathbf{0}$,

Joint $(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2)$ pdf

Taking $-2\ \log$, the exponent is proportional to:

$$\frac{1}{\sigma^2}\,\|\boldsymbol{y}-\boldsymbol{X}\cdot\boldsymbol{\beta}\|^2-\lambda\,\|\boldsymbol{\beta}\|^2-2\ \log h(\sigma^2\mid\boldsymbol{y}).$$

Given σ^2, the target function in the ridge optimization.

The posterior pdf is proportional to this function.

The MAP estimator is just the Ridge solution.

Joint posterior pdf

$$h(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = h(\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y}) \cdot h(\sigma^2 \mid \boldsymbol{y})$$

where:

$$\begin{aligned} (\boldsymbol{\beta} \quad \mid \sigma^2, \boldsymbol{y}) &\sim \textbf{Normal}(\widetilde{\boldsymbol{b}}, \sigma^2\, \widetilde{\boldsymbol{B}}), \\ (\sigma^2 \;\mid \boldsymbol{y}) &\sim \mathrm{IG}(\widetilde{\alpha}, \widetilde{\beta}). \end{aligned}$$

$\widetilde{\boldsymbol{b}}$, $\widetilde{\boldsymbol{B}}$, $\widetilde{\alpha}$, $\widetilde{\beta}$ are the updated parameters.

Formulas for updating parameters

$$
\begin{aligned}
\widetilde{\boldsymbol{b}} &= (\boldsymbol{B}^{-1} + \boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot (\boldsymbol{B}^{-1} \cdot \boldsymbol{b} + \boldsymbol{X}' \cdot \boldsymbol{y}),\\
\widetilde{\boldsymbol{B}} &= (\boldsymbol{B}^{-1} + \boldsymbol{X}' \cdot \boldsymbol{X})^{-1},\\
\widetilde{\alpha} &= \alpha + \frac{n}{2},\\
\widetilde{\beta} &= \beta + \frac{1}{2}\left[\boldsymbol{b}' \cdot \boldsymbol{B}^{-1} \cdot \boldsymbol{b} + \boldsymbol{y}' \cdot \boldsymbol{y} - \widetilde{\boldsymbol{b}}' \cdot \widetilde{\boldsymbol{B}}^{-1} \cdot \widetilde{\boldsymbol{b}}\right].
\end{aligned}
$$

Recovering the classical *Ridge* regression

In particular, when the prior parameters are:

$$\begin{aligned} \boldsymbol{b} &= \mathbf{0}, \\ \boldsymbol{B} &= (1/\lambda)\,\boldsymbol{I}, \qquad \lambda > 0, \end{aligned}$$

Updating for $\boldsymbol{b} = \mathbf{0}$, $\boldsymbol{B} = (1/\lambda)\,\boldsymbol{I}$, $\lambda > 0$

$$\begin{aligned}
\widetilde{\boldsymbol{b}} &= (\lambda\,\boldsymbol{I} + \boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot (\boldsymbol{X}' \cdot \boldsymbol{y}),\\
\widetilde{\boldsymbol{B}} &= (\lambda\,\boldsymbol{I} + \boldsymbol{X}' \cdot \boldsymbol{X})^{-1},\\
\widetilde{\alpha} &= \alpha + \frac{n}{2},\\
\widetilde{\beta} &= \beta + \frac{1}{2}\left[\boldsymbol{y}' \cdot \boldsymbol{y} - \boldsymbol{y}' \cdot \boldsymbol{X} \cdot (\lambda\,\boldsymbol{I} + \boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X} \cdot \boldsymbol{y}\right].
\end{aligned}$$

Bayesian *Ridge* regression

The *Ridge regression* coefficients are the posterior expected values.

λ can be interpreted as the size of virtual prior sample with mean **0** (redefine $1/\lambda \to \sigma^2/\lambda$), thus shrinking the posterior pdf of the regression coefficients towards **0**.

07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

Sticking to the success story

Can we repeat this reasoning with the Lasso?

Replace the Gaussian prior for each β_j with a Laplace (double exponential) pdf:

$$f(\beta_j) = \frac{1}{2\,\sigma}\,\exp\left(-\frac{|\beta_j - \mu|}{\sigma}\right),$$

with $\mu = 0, \sigma = 1/\lambda$ (or, better, σ^2/λ).

Joint $(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2)$ pdf

Taking $-2\ \log$, the exponent has a first summand

$$\propto\ \frac{1}{\sigma^2}\,\|\boldsymbol{y}-\boldsymbol{X}\cdot\boldsymbol{\beta}\|^2,\qquad \text{the sum of residual squares,}$$

and a second one $\propto$ the l^1 norm of $\boldsymbol{\beta}$, $\lambda\ \sum_{j=1}^{p}|\beta_j|$.

Given σ^2, the target in the Lasso optimization.

Why condition on σ^2?

Conditioning on σ^2 is important because it guarantees a unimodal full posterior. For σ^2 prior we can choose:

$$\sigma^2 \sim \mathrm{IG}(a, b),$$

or the limit improper noninformative pdf,

$$h(\sigma^2) = \frac{1}{\sigma^2}.$$

07b - Reg. & Sparsity

Regularization: Bias-variance tradeoff

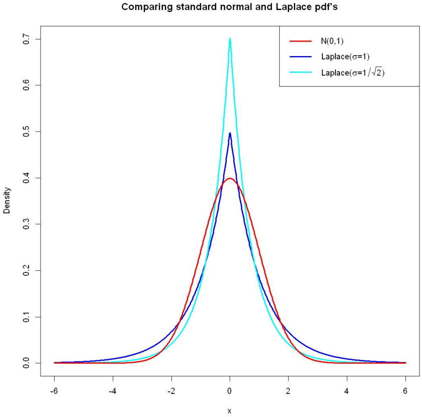
Ridge regression & The *LASSO*

Bayesian Ridge regression

The Bayesian LASSO

Horseshoe and shrinkage priors

Comparing Lasso and Ridge priors



The Scale Mixture of Normals (SMN) trick

The $|\,\cdot\,|$ function is non differentiable.

This is trouble for simulation.

Following Park and Casella (2008), the identity:

$$\frac{a}{2}\, e^{-a\,|z|} = \int_0^\infty \frac{1}{\sqrt{2\,\pi\,s}}\, e^{-z^2/(2\,s)} \cdot \frac{a^2}{2} \cdot e^{-a^2\,s/2}\, ds.$$

shows the Laplace pdf is an SMN.

The SMN allows a Bayesian description

$$
\begin{aligned}
z &\sim \mathrm{DExp}(0, a) \ \text{ is equivalent to:}\\
z &\sim \mathrm{Normal}(0, s), \quad \text{and}\\
s &\sim \mathrm{Exp}\left(\frac{a^2}{2}\right),
\end{aligned}
$$

(thus an MCMC sampling is possible)

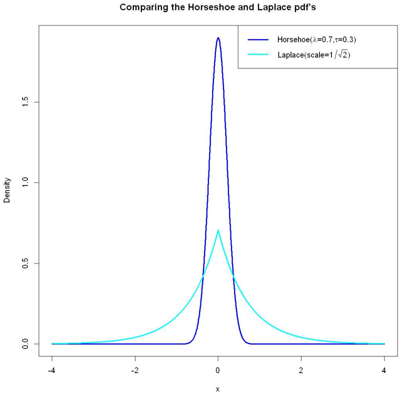
Possible generalizations

Try to obtain priors with a sharper peak.

Substitute other mixing pdf's for the Exp( ).

E.g. Half-Cauchy$(0, \ ) \Rightarrow$ The horseshoe.

The horseshoe prior


Comparing the Horseshoe and Laplace pdf's
Horsehoe(λ=0.7,τ=0.3)
Laplace(scale=$1/\sqrt{2}$)
Density
x