Machine Learning for Economists: Part 4 – Shrinkage and Sparsity

International Monetary Fund

Washington, D.C.,
November, 2019

Disclaimer #1:

The views expressed herein are those of the authors and should not be attributed to the International Monetary Fund, its Executive Board, or its management.

Regularization – A Refresher

Model with high relative representational capacity may overfit...

When they overfit and learn more about exceptions that 'true' pattern, they **generalize poorly** to new datasets

Regularization is "any modification we make to a learning algorithm that is intended to reduce its generalization error" (Goodfelow et al. 2017)

Often, prior belief about a simpler sub-model is put to test with the data...

Regularization

A common form of regularization in **parametric** models is penalizing coefficients deviation towards zero...

$$\min_{\beta} \sum_{i=1}^{N} (y - (\alpha_0 + x'\beta))^2 + \lambda \times \text{Penalty}(\beta - 0)$$

Three frequent specifications are:

- **Ridge Regression:** Penalty = $\sum_i \beta_i^2$
- **Lasso:** Penalty = $\sum_i |\beta_i|$
- **Elastic Net:** Penalty = $(1 - \alpha) \sum_i \beta_i^2 + \alpha \sum_i |\beta_i|$

!! Variables in **x** must be NORMALIZED !!

Shrinkage

Ridge due to Hoerl and Kennard (1970)

Ridge/Weight Decay/Tikhonov regularization: $\sum_i \beta_i^2$

- Shrinks coefficients towards the prior (zero)
- Coefficients rarely set to hard zero, the penalty is *smooth*
- Numerically stabilizes ill-conditioned models and those where we have more features than data points, $N_{obs} \leq p$
- $\widehat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{Y}$
- If only one λ, vairables must be normalized, so β_k are comparable…

Sparsity

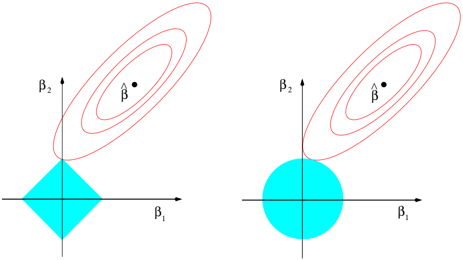
LASSO due to Robert Tibshirani (1996).

Lasso
(Least abs. shrinkage and selection operator): $\sum_i |\beta_i|$

- Can shrink some coefficents to **hard zero**
- Performs a form of 'continous variable selection', promotes **sparsity**
- If only one λ, vairables must be normalized, so β_k are comparable...

LASSO vs. Ridge

With **lasso** the combination of coefficients consistent with a constant penalty, e.g. $|\beta_1| + |\beta_2| = \text{const}$, has **corners**, allowing for corner solutions, combined with elliptical contours of the loss function...



$|\beta_1| + |\beta_2| = \text{const}$ $\beta_1^2 + \beta_2^2 = \text{const}$

Hastie, Tibshirani, and Freedman (2005)

With many variables, $p > 2$ the relevant penalty space has many corners, flat edges, and faces – many opportunities for params to be zero!

Orthogonal Regressors Case – Intuition

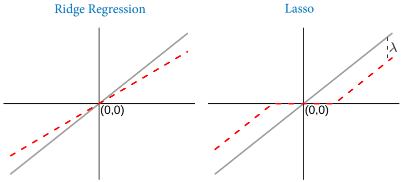
In the case of orthogonal components in **X** ridge and lasso elastic net have explicit solution that helps with intuition.

Ridge: – proportional shrinkage

$$\widehat{\beta}_j = \frac{\beta_{ols,j}}{(1+\lambda)} \tag{1}$$

Lasso: – soft thresholding

$$\widehat{\beta}_j = \text{sign}(\beta_{ols,j})(|\beta_{ols,j}| - \lambda)_+ \tag{2}$$



Ridge, Lasso, and Elastic Net

Ridge Regression:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - (\beta_0 + x_i'\beta))^2 + \lambda \frac{1}{2} ||\beta||_2 \right\} \tag{3}$$

Lasso:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - (\beta_0 + x_i'\beta))^2 + \lambda ||\beta||_1 \right\} \tag{4}$$

Elastic Net:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - (\beta_0 + x_i'\beta))^2 + \lambda \left[\frac{1}{2} (1-\alpha) ||\beta||_2 + \alpha ||\beta||_1 \right] \right\} \tag{5}$$

Bayesian View – Intuition

In Bayesian view, the prior information about the model parameters, $p(\beta)$, is getting **updated** by observing the data, $D = (Y, X)$, via its likelihood, $p(D|\beta)$:

$$
\begin{aligned}
p(\beta|D) &= \frac{P(D|\beta) \times p(\beta)}{p(D)} \\
&\propto P(D|\beta) \times p(\beta)
\end{aligned}
$$

$$
\log p(\beta|D) \propto \log P(D|\beta) + \log p(\beta)
$$

Intuitively, for point 'maximum a-posterior' (MAP) estimate, it is a '**penalized optimization**'

Bayesian View – Intuition

Thus, a **ridge** regression

$$\arg\max_{\beta,\beta_0} \left\{ \sum_{i=1}^{N} (y_i - (\beta_0 + x_i'\beta))^2 + \lambda \sum_{k=1}^{p} (\beta_k - 0)^2 \right\}$$

corresponds to a model with **Gaussian prior** belief:

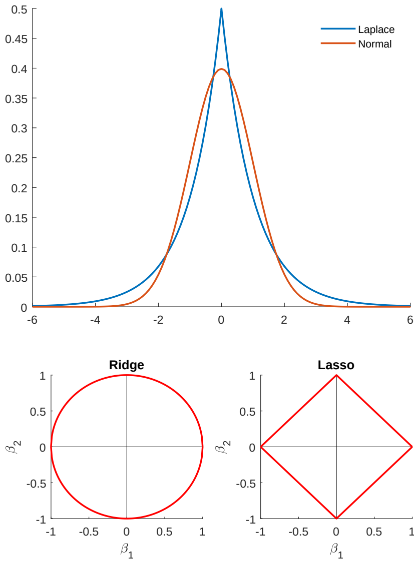
$$\beta_k \sim N(0, \sigma_k), \quad N_{pdf}(\beta_k, 0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\beta_k - 0)^2}{2\sigma^2}},$$

and thus

$$\text{argmax}_\beta \sum_{i=1}^{N} \log N_{pdf}(y_i; (\beta_0 + x_i'\beta), \sigma_e) + \sum_{k=1}^{p} \log N_{pdf}(\beta_k; 0, \sigma)$$

LASSO corresponds to a **Laplace prior**, $\beta \sim \frac{\lambda}{2\sigma} e^{-\frac{\lambda}{\sigma}|\beta_k|}$.

Ridge vs. Lasso – Priors and Equidistant Contours



Elastic Net (1)

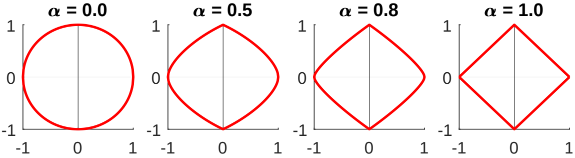
Elastic Net is a combination of Ridge and Lasso

"like a stretchable fishing net that retains 'all the big fish' "

Zou and Hastie (2005)

$$\min_{\beta_0,\beta}\left\{\sum_{i=1}^{N}(y_i-(\beta_0+x_i'\beta))^2+\lambda\left[\frac{1}{2}(1-\alpha)||\beta||_2+\alpha||\beta||_1\right]\right\}$$

ElasticNet introduces two hyperparameters, λ and α.



Elastic Net (2)

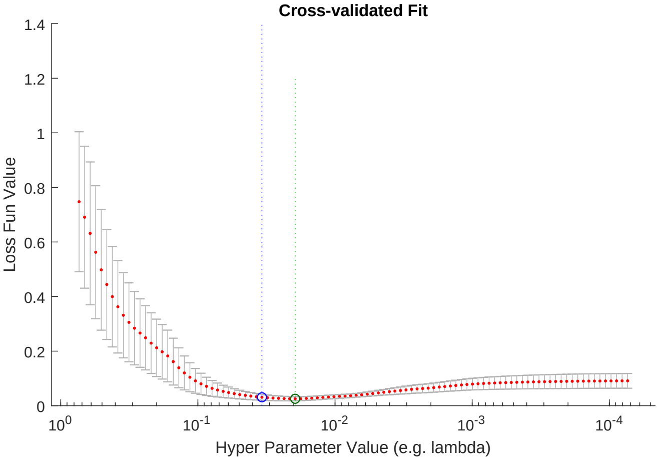
ElasticNet attempts to take the best $L1$ and $L2$ worlds.

Issues it solves:

- For cases where $p \geq N_{obs}$, ridge works but lasso saturates at N_{obs}
- Lasso handles poorly very correlated variables, picks arbitrarily one and eliminates the others, while ridge attributes the same weight to all, ElasticNet 'groups' the correlated variables
- For common situations with $N_{obs} >> p$, and highly correlated predictors, ridge dominates pure lasso...
- For $\lambda > 0$ and $\alpha < 1$ ElasticNet is strictly convex..., with a unique solution

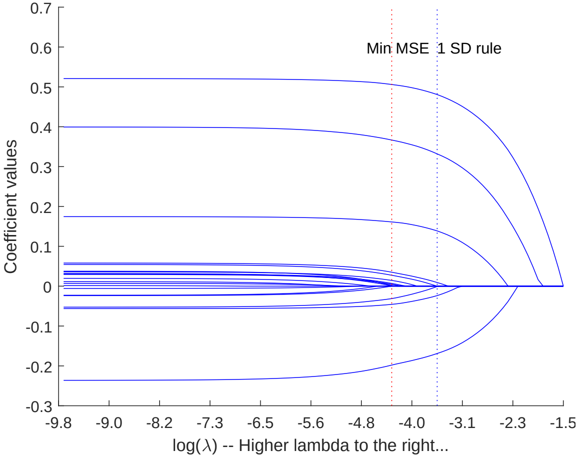
What Value for λ?

The **hyperparameter** λ can be estimated using a **hold-out** set (validation or cross-validation)



Regularization Path

It's worth looking at evolution of β as λ changes...



Why log λ? It is common and useful to create hyper-parameter grids in logs...

Prior Restriction on Coefficients

It is important to understand the principles of prior information about coefficients.

Lasso and Ridge should not be applied mindlessly...

In economics, the priors may be about shrinking to other values than **zero** and **economic theory** should be the guide

Example: **Bayesian VARs**

- Coefs shrunk to 0 or 1 (unit roots)
- For coefficients on higher lags, λ increases
- ...

Extensions

Group Penalties/Priors

- $L(\beta) = MSE(\beta) + \sum_{g=1}^{G} \lambda_g \left\{ \sum_{j \in g} \text{Penalty}(\beta_j) \right\}$
- Bayesian VARs, ...
- Regression with dummy-coded categorical inputs
- ...

Fused Penalties

- For problems with features having natural order, sometimes we prefer neighboring coefficients to be similar...
- Penalty $= \lambda_1 \sum_{k=1}^{p} ||\beta_i|| + \lambda_2 \sum_{k=1}^{p-1} ||\beta_{i+1} - \beta_i||$
- DNA, time series, ...

Many other extensions: hierarchical adaptive lasso, spike-and-slab lasso, ...

More on LASSO...

post-LASSO. . .

After Lasso, the estimated coefficient reflect the bias due to the "tresholding"

Post-LASSO:

1. Estimate some version of LASSO
2. Apply OLS to the selected model to remove the bias

Sometimes, people forget to do post-Lasso.

Don't be that person ;)

“Tune-free” Lasso. . . ?

Under certain conditions (Bickel, Ritov, Tsybakov, Ann. of Stat. 200) the rate-optimal choice of penalty level is

$$\lambda = \sigma 2\sqrt{2\log(pn)/n}. \tag{6}$$

Now. . . σ, variance of the error, is of course not known. . .

If need be, must be estimating iteratively, not a problem

The $\sqrt{\text{LASSO}}$

Belloni, Chernozhukov, Wang, Biometrika 2010

With a clever modification of the Lasso,

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i - x_i'\beta]^2} + \lambda||\beta||_1 \tag{7}$$

they show that the rate-optimal penalty level is **independent** of σ.

$$\lambda = \sqrt{2\log(pn)/n}$$

The solution method is different from "standard" Lasso approaches but this is as "tuning-free" Lasso as it gets...

Wonkish: More on Ridge Regression…

The problem is, for given λ

$$RSS(\lambda) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta'\beta \tag{8}$$

with the solution

$$\widehat{\beta}_r = (\mathbf{X}'\mathbf{X} - \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}. \tag{9}$$

The regularization by the diagonal matrix $\lambda\mathbf{I}$ ameliorates the collinearity and invertibility of the least-square problem…

Wonkish: Computing the LASSO parameters…

How can you solve LASSO? Many ways...

Coordinate Descent very simple to implement & intuitive

For $f(x) = g(x) + \sum_{i=1}^{n} h_i(x_i)$ with $g(x)$ convex and differentiable and each $h_i(.)$ convex, coordinate descent can find a global minimizer...
Start with $x^{(0)}$ and for $k = 1, 2, \ldots$ repeat

$$\begin{aligned}
x_1^{(k)} &= \arg\min_{x_1} f(x_1, x_2^{(k-1)}, x_3^{(k-1)}, \ldots, x_n^{(k-1)}) && (10)\\
x_2^{(k)} &= \arg\min_{x_2} f(x_1^{(k)}, x_2, x_3^{(k-1)}, \ldots, x_n^{(k-1)}) && (11)\\
&\ldots && (12)\\
x_n^{(k)} &= \arg\min_{x_n} f(x_1^{(k)}, x_2^{(k)}, x_3^{(k-1)}, \ldots, x_n) && (13)
\end{aligned}$$

And, crucially, there is a simple closed-form solution for each coordinate optimization problem for the LASSO…

Wonkish: Computing the LASSO parameters...

Let's have the problem of LASSO:

$$\min_{\beta} \frac{1}{2N} \sum_{i}^{N} (y_i - \sum_{j=1}^{p} x_{i,j}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{14}$$

1. Compute 'partial residuals', $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$
2. Compute the LS coefficient $\beta^* = \frac{1}{N}\sum_{i=1}^{N} x_{ij} r_{ij}$
3. Use soft-thresholding to update β_j

   $\beta_j = S(\beta_j^*, \lambda) = (\beta_j^*)(|\beta_j^*| - \lambda)_+$

Post-Selection Inference – SEE NEW SLIDES ON INFERENCE!!

(Machine learning pratictioners rarely care about inferences. . .)

After the model search and selection (e.g. choosing) you
CAN NOT
just use the p-values and such. . .

The whole model search process needs to be always, always, always taken into account.

For **explicit** and admitted model search the literature is now finding ways to do inference

One of the ways to account for model selection is to **boostratp** the whole selection & estimation process. . . (Efron, 2013, Estimation and Accuracy after Model Selection). Or sample splitting, double-selection lasso, etc.

Reading: Chernozhukov, Hansen, Spindler (2015), Tibshirani et al. (2016,JASA), . . .

ADDITIONAL SLIDES

Spike and Slab Model

Originally proposed by Mitchell and Beuchamp, 1988

In Bayesian variable selection, the requirement for sparsity is to set the loading coef as $\gamma_j = 1$ if 'relevant/useful' and $\gamma_j = 0$ otherwise

For small problems, the posterior prob. of inclusion can be computed in an exhaustive ways... but there are 2^p models!

Spike-and-slab is based on a hierarchical prior for coefficients, β:

$$p(\beta_j; \sigma, \gamma_j) = \begin{cases} 0 & \text{for } \gamma_j = 0 \\ N(\beta_j; 0, \sigma^2 \sigma_\beta^2) & \text{for } \gamma_j = 1 \end{cases}$$

and

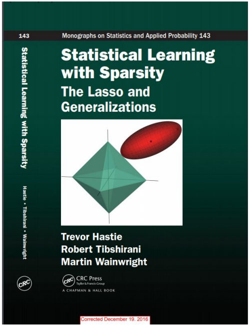
$$p(\gamma) = \prod_{k=1}^{p} \pi_0^{\gamma_k} (1 - \pi_0)^{1-\gamma_k} = \pi_0^{\sum_k^p \gamma_k} (1 - \pi_0)^{p - \sum_{k=1}^p} \tag{15}$$

so that the prior 'penalty' is

$$\log p(\gamma|\pi_0) = -\lambda \times \sum_{k=1} \gamma_k + \text{const}, \gamma \in \{0, 1\}$$

and π_0 is prior expected fraction of large β_js and $\lambda \equiv \log \frac{1-\pi_0}{\pi_0}$.

For Enthusiasts. . .



https://web.stanford.edu/ hastie/StatLearnSparsity_files/SLS.pdf

Thank you for your patience. . .