Les Cahiers du GERAD ISSN: 0711–2440

Generalized Elastic Net Regression

G. Mouret, J.-J. Brault,
V. Partovi Nia

G–2013–109

December 2013

Generalized Elastic Net Regression

Geoffroy Mouret

GERAD & Department of Electrical Engineering
Polytechnique Montréal
Montréal (Québec) Canada, H3C 3A7

geoffroy.mouret@polymtl.ca

Jean-Jules Brault

Department of Electrical Engineering
Polytechnique Montréal
Montréal (Québec) Canada, H3C 3A7

jean-jules.brault@polymtl.ca

Vahid Partovi Nia

GERAD & Department of Mathematics and Industrial Engineering
Polytechnique Montréal
Montréal (Québec) Canada, H3C 3A7

vahid.partovinia@polymtl.ca

December 2013

Les Cahiers du GERAD

G–2013–109

Abstract: This work presents a variation of the elastic net penalization method. We propose applying a combined ℓ_1 and ℓ_2 norm penalization on a linear combination of regression parameters. This approach is an alternative to the ℓ_1-penalization for variable selection, but takes care of the correlation between the linear combination of parameters. We devise a path algorithm fitting method similar to the one proposed for the least angle regression. Furthermore, a one-shot estimation technique of ℓ_2 regularization parameter is proposed as an alternative to cross-validation. A simulation study is conducted to check the validity of the new technique.

Key Words: Elastic net, fitted values, generalized lasso, least angle regression, variable selection.

1 Introduction

Regression aims at predicting the response variable $\mathbf{y}$, given p covariates $\mathbf{x}_1, \ldots, \mathbf{x}_p$

$$\mathbf{X} = (\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_p)\,, \boldsymbol{\beta} = \begin{pmatrix} [0.5]\beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \mathbf{1} = \begin{pmatrix} [0.5]1 \\ \vdots \\ 1 \end{pmatrix}, \mathbf{y} = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\varepsilon}$ is a vector of Gaussian noise. Regression has two aspects, parameter estimation and prediction. It is known that the famous ordinary least squares (OLS) estimator, obtained by minimizing the residual squared error, presents too much variance if explanatory variables are correlated or unimportant covariates are considered in the linear model. The OLS method is inefficient in both prediction and estimation of $\boldsymbol{\beta}$ for the case of large p and small n which is appearing in modern applications.

The use of penalized parameters as a regularization term in linear regressions has proven to be an effective approach to the problem of large variance estimates. Ridge regression is a good example of such technique. The ridge regression improves the prediction performance by reducing variance at the cost of a small bias by solving

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \operatorname{argmin}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right), \tag{2}$$

where $\|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \beta_j^2$ and λ is the penalization constant.

However, ℓ_2-penalization can only shrink the parameters towards 0 and does not provide the sparsity required for successful variable selection. Replacing this penalization term by the ℓ_1-norm of the parameters leads to the lasso (Tibshirani, 1996), a well-known technique for variable selection

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \operatorname{argmin}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right), \tag{3}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$.

Though the lasso presents good computational properties, it fails to select groups of correlated variables. The lasso also estimates only up to n coefficients, which may be inconvenient when working in high dimensional spaces ($p \gg n$).

Over the past twenty years, many regularization terms have been proposed. For instance Frank and Friedman (1993) proposed a ℓ_q-norm based penalization in the *Bridge* regression

$$\hat{\boldsymbol{\beta}}_{\text{bridge}} = \operatorname{argmin}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p} |\beta_j|^q\right). \tag{4}$$

Lasso and Ridge are both particular cases of this penalization for $q = 1$ and $q = 2$ respectively. Optimizing over λ and more specially q is computationally heavy. The bridge regression only offers variable selection for $q \leq 1$. Furthermore, $0 < q < 1$ gives a non-convex optimization problem, since $|\beta_j|^q$ is a non-convex ball. Hence, it is difficult to find the solution of the regression problem, even for fixed values of q.

More recent works suggest non-convex penalization, using a different approach such as *SCAD* of Fan and Li (2001). The penalty function is singular at the origin to produce sparsity and is bounded to get unbiased estimates for large coefficients.

However, ℓ_1 and ℓ_2 penalization are still used as improved versions of the OLS estimates. Zou and Hastie (2005) took advantage of both techniques to get a reliable way of handling selection of groups of correlated variables, by combining ℓ_1 and ℓ_2 regularization terms, called the elastic net penalty

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \operatorname{argmin}\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\sum_{j=1}^{p} |\beta_j| + \lambda_2\sum_{j=1}^{p} |\beta_j|^2\right). \tag{5}$$

The optimization of two different penalization constants through cross-validation can be highly time-consuming, but as shown in Zou and Hastie (2005), solving the elastic net for a given value of λ_2 is equivalent to solving the lasso problem. The least angle regression (Efron et al., 2004) is a path algorithm that finds the solution path to the lasso problem with computational complexity of a single least squares. Therefore, cross-validation is only required on the quadratic penalization constant λ_2. By taking a Bayesian perspective we propose using marginal likelihood maximization to estimate λ_2.

Our proposed model is inspired by a combination of the elastic net (Zou and Hastie, 2005) and the generalized lasso (Tibshirani and Taylor, 2011). Instead of penalizing the parameters, we propose penalizing a linear combination of the parameters

$$\hat{\boldsymbol{\beta}}_{\text{GEN}} = \operatorname{argmin} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\mathbf{D}_1\boldsymbol{\beta}\|_1 + \lambda_2 \|\mathbf{D}_2\boldsymbol{\beta}\|_2^2 \right). \tag{6}$$

This model collects a wide set of regression problems for different choices of $\mathbf{D}_1$ and $\mathbf{D}_2$. An ℓ_1-penalization to structure the sparsity and ℓ_2-penalization to control the correlation among the linear combination of parameters.

In some applications, sparsity over a linear combination of parameters is required, such as the fused-lasso (Tibshirani et al., 2005) in which the penalization is applied on the difference between consecutive parameters. The ℓ_2 penalization is not always an appropriate regularization. We suggest to generalize the ℓ_2 penalty by implementing the Mahalanobis regularization which gives ℓ_2 penalization as a special case. The choice of the regularization design matrices $\mathbf{D}_1$ and $\mathbf{D}_2$ depends on the context.

2 Generalized Elastic Net

We encourage the Bayesian perspective over the regularized regression problem. In other words, we propose to look at the squared residual terms as the log of the likelihood, and regularization over the parameters as the log of the prior distribution assumed for the regression parameters. This view gives a convenient method for estimation of the regularization constant. Moreover, motivated by the hierarchical Bayes modelling, if likelihood distribution is connected to the distribution of $\mathbf{y} \mid \mathbf{X}\boldsymbol{\beta}$, then the regularization term is connected to the distribution of $\boldsymbol{\beta} \mid \mathbf{X}$ and the distribution of $\mathbf{X}$. This view suggests using a penalty that is a function of the design matrix $\mathbf{X}$.

2.1 Tuning ℓ_2 penalization

We propose taking a Bayesian view to the linear regression problem. The least squares estimates coincide with the maximum likelihood estimator if we assume that response variable y is normally distributed around the fitted linear combination of the covariates $\mathbf{x}_j$. The model-based variant will be $\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, in which $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate normal distribution having the mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Therefore, the Mahalanobis penalty is equivalent to a Gaussian prior over the regression parameters $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \tau^2\boldsymbol{\Omega}), \tag{7}$$

where $\boldsymbol{\Omega}$ is the shape of the penalization prior and τ^2 is the scale parameter.

The maximum posterior estimation of $\boldsymbol{\beta}$ is equivalent to

$$\hat{\boldsymbol{\beta}}_{\text{MAP}} = \operatorname{argmin} \left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\sigma^2}{\tau^2} \|\boldsymbol{\Omega}^{-\frac{1}{2}}\boldsymbol{\beta}\|_2^2 \right). \tag{8}$$

The maximum a posteriori estimation of $\boldsymbol{\beta}$ coincides with the generalized ridge estimation if $\frac{\sigma^2}{\tau^2}$ is replaced by λ and $\boldsymbol{\Omega}^{-\frac{1}{2}}$ is replaced by $\mathbf{D}_2$ in (8). As both problems are analytically equivalent, we can use the Empirical

Table 1: Prior distribution shapes $\mathbf{\Omega}$ on the parameters and associated quadratic penalties.

$\mathbf{\Omega}$	Penalization
$\mathbf{I}_p$	$\|\boldsymbol{\beta}\|_2^2$
$(\mathbf{X}^\top\mathbf{X})^{-1}$	$\|\mathbf{X}\boldsymbol{\beta}\|_2^2$
$\left(\frac{\mathbf{X}^\top\mathbf{X}}{n}\right)^{-1}$	$\frac{1}{n}\|\mathbf{X}\boldsymbol{\beta}\|_2^2$
$(\alpha\mathbf{I}_p + (1-\alpha)\mathbf{X}^\top\mathbf{X})^{-1}$	$\alpha\|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\mathbf{X}\boldsymbol{\beta}\|_2^2$
$\left(\alpha\mathbf{I}_p + \frac{(1-\alpha)}{n}\mathbf{X}^\top\mathbf{X}\right)^{-1}$	$\alpha\|\boldsymbol{\beta}\|_2^2 + \frac{(1-\alpha)}{n}\|\mathbf{X}\boldsymbol{\beta}\|_2^2$

Bayes engine to estimate the prior parameters, here λ. The empirical Bayes principle maximizes the marginal likelihood as a function of λ

$$p_\lambda(\mathbf{y} \mid \mathbf{X}) = \int \dots \int_{\mathbb{R}^p} p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) p_\lambda(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\beta} \qquad \lambda = \frac{\sigma^2}{\tau^2}. \tag{9}$$

It is easy to see

$$\mathbf{y} \mid \mathbf{X}, \lambda \sim \mathcal{N}_n \left[\mathbf{0}, \sigma^2 \left\{ I_n - \mathbf{X} \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{\Omega}^{-1} \right)^{-1} \mathbf{X}^\top \right\}^{-1} \right]. \tag{10}$$

Since the choice of a penalization is equivalent to a prior distribution on the parameters, one can select a subjective regularization. For instance, if there is a knowledge that some variables are inter-correlated, it is more sensible to choose regularization that reflects this information. This is why we encourage choosing $\mathbf{\Omega} = (\mathbf{X}^\top\mathbf{X})^{-1}$. This choice of $\mathbf{\Omega}$ corresponds to penalization over the fitted values, see Table 1 for more details. Table 1 also presents different covariance matrices $\mathbf{\Omega}$ and the associated regularizations.

We optimize the likelihood as a function of λ and $\mathbf{\Omega}$. Hence, it is possible to make a more flexible model by introducing another parameter α. This parameter is reserved to discriminate between two possible prior distributions. Setting $\alpha = 0$ corresponds to a ℓ_2-penalization of the fitted values and setting $\alpha = 1$ corresponds to the ridge regression.

2.2 Tuning ℓ_1 penalization constant

It is feasible to reduce the elastic net problem to the lasso regression. Once we are brought back to the lasso, the path algorithm (Efron et al., 2004) provides the whole solution path.

We apply a similar analogy to reduce the generalized elastic net problem to a generalized lasso problem. The whole solution is provided in Tibshirani and Taylor (2011). It is not difficult to see the following change in the response vector and the design matrix reduces a generalized elastic net regression to a generalized lasso regression

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{\Omega}^{-\frac{1}{2}} \end{pmatrix} \text{ and } \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_{p\times 1} \end{pmatrix}, \tag{11}$$

where $\mathbf{\Omega}^{-\frac{1}{2}} = \mathbf{D}_2$ is the quadratic design matrix. In other words,

$$\|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{D}_1\boldsymbol{\beta}\|_1 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\mathbf{D}_1\boldsymbol{\beta}\|_1 + \lambda_2\|\mathbf{D}_2\boldsymbol{\beta}\|_2^2. \tag{12}$$

The generalized elastic net is therefore equivalent to the generalized lasso for the transformed $\mathbf{y}^*$ and $\mathbf{X}^*$. This allows us to use the computationally efficient path algorithm presented in Tibshirani and Taylor (2011) to solve the generalized elastic net problem .

3 Simulation

We set up two separate simulations to verify the behaviour of our proposed method. The first simulation explores the estimation quality of the ℓ_2 regularization constant. The second simulation aims at showing the behaviour of penalizing linear combinations. We choose a common choice for $\mathbf{D}_1 = \mathbf{D}_2 = \mathbf{X}$. This penalization regularizes the fitted values.

3.1 Estimating the ridge parameter

In the first simulation we simulate 30 explanatory variables, each variable independently sampled from $\mathcal{N}(0,1)$. We produce $n = 450$ of such data points and sample the regression coefficients from $\mathcal{N}(0, 5\boldsymbol{\Omega})$. We have chosen three different structures for $\boldsymbol{\Omega}$. The variance-covariance structure $\boldsymbol{\Omega}^{-1}$ is $\mathbf{I}_p$, $\mathbf{X}^\top\mathbf{X}$, or a convex combination of these two structures being $0.2\mathbf{I}_p + 0.8\mathbf{X}^\top\mathbf{X}$. Note that we did not consider an ℓ_1 penalization for this simulation since the aim of the study is to check the efficiency of the method for estimation of the ℓ_2 penalizing constant. The errors $\boldsymbol{\varepsilon}$ are taken independently from $\mathcal{N}(0, 0.01)$.

We used the following log reparametrization (θ_1, θ_2) to avoid constrained numerical optimization procedures. We used $\theta_1 = \log\tau^2$ and $\theta_2 = \log\left\{\frac{\alpha}{1-\alpha}\right\}$.

3.1.1 Penalizing parameters $\boldsymbol{\Omega} = \mathbf{I}_p$

If we choose the independent structure, maximum a posteriori analysis coincides with the ridge regression with $\lambda = \frac{\sigma^2}{\tau^2}$. Therefore, the estimation of λ coincides with the estimation of the inverse of the signal to noise ratio. For $\sigma^2 = 0.01$, $\hat{\lambda} = \frac{0.01}{\exp(\hat{\theta}_1)}$, and it is more intuitive to check the quality of estimation directly on θ_1, see Figure 1. Figure 1 (left panel) illustrates the marginal likelihood curves and the maximum marginal likelihood value for 100 simulations. The true value of θ_1 matches the marginal maximum likelihood estimator, confirming the estimation quality for $\boldsymbol{\Omega} = \mathbf{I}_p$.

We get an estimator of $\log(\tau^2)$ by averaging the maxima over a hundred simulations. This leads to the estimator $\hat{\tau}^2 = 4.98$ within the 95% confidence interval of $[4.91, 5.06]$, which includes the true value used for the simulation ($\tau^2 = 5$).

3.1.2 Penalizing fitted values: $\boldsymbol{\Omega} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$

The behaviour of the marginal likelihood is not much different with independent structure $\boldsymbol{\Omega} = \mathbf{I}_p$. In both cases the asymptotic normal confidence interval includes the true value of θ_1.

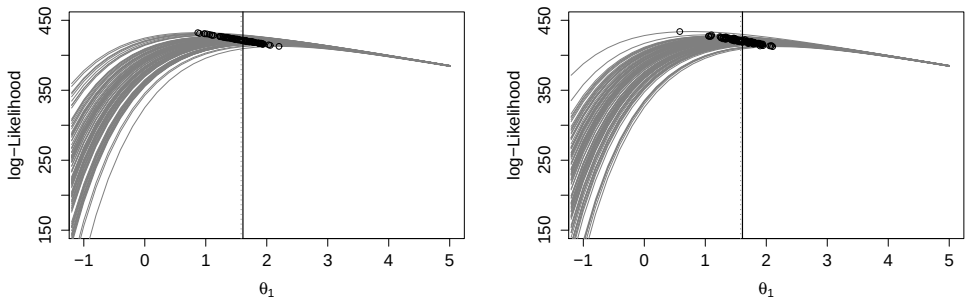

Figure 1: Marginal log likelihood curves, $\log p_\lambda(\mathbf{y} \mid \mathbf{X})$ for $\boldsymbol{\Omega} = \mathbf{I}_p$ (left panel) and $\boldsymbol{\Omega} = (\mathbf{X}^\top\mathbf{X})^{-1}$ (right panel). The gray curves demonstrate the marginal log likelihood and black circles show the maximum marginal log likelihood found by the numerical optimization routine. The vertical dashed line illustrates the mean of the maximum likelihood points while the solid vertical line shows the true value.

3.1.3 Mixture penalization $\boldsymbol{\Omega} = \left(\alpha \mathbf{I}_p + (1-\alpha)\left(\mathbf{X}^\top \mathbf{X}\right)\right)^{-1}$

The mixture structure $\boldsymbol{\Omega} = \left(\alpha \mathbf{I}_p + (1-\alpha)\left(\mathbf{X}^\top \mathbf{X}\right)\right)^{-1}$ does not have an intuitive interpretation. Though, in this setting λ estimates the volume of the ball, and α estimates the shape of the ball. Figure 2 shows marginal log likelihood contour. This figure suggests that estimation of the volume λ is easier than the shape α.

3.2 Tuning the lasso parameter

The path algorithm allows to find the whole solution path for different values of λ_1. Figure 3 shows the trajectories of the fitted values (left panel). We observe that the choosing $\boldsymbol{\Omega}$ cannot define much sparsity over the fitted values even while the ℓ_1 penalization is over the fitted values. A similar behaviour appears when only one regression parameter is in the model, see Figure 4.

4 Conclusion

We proposed a generalization of the elastic net regression. We applied a combination of ℓ_1 and ℓ_2 penalization on a linear combination of the parameters, defined by two design matrices $\mathbf{D}_1$ and $\mathbf{D}_2$. Both matrices have a separate and distinct impact, as $\mathbf{D}_2$ takes care of the correlation appearing between the linear combinations and $\mathbf{D}_1$ selects the linear combinations.

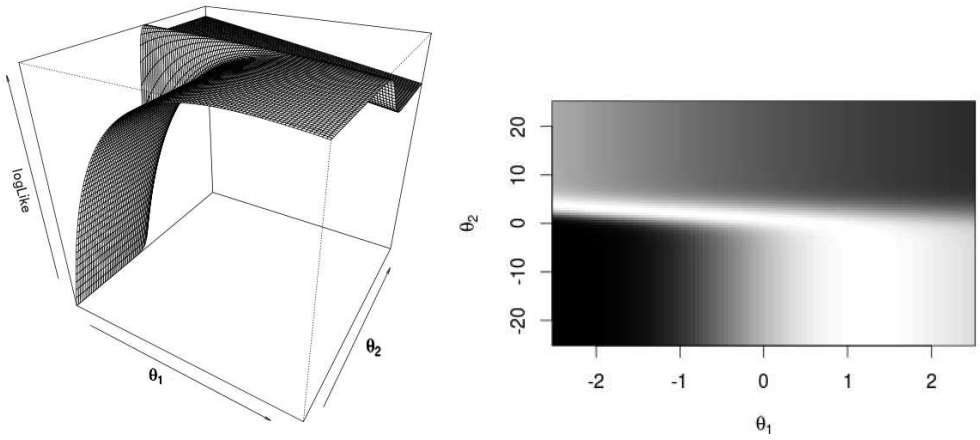

Figure 2: Left panel: the marginal log likelihood surface. Right panel: the marginal log likelihood contour.

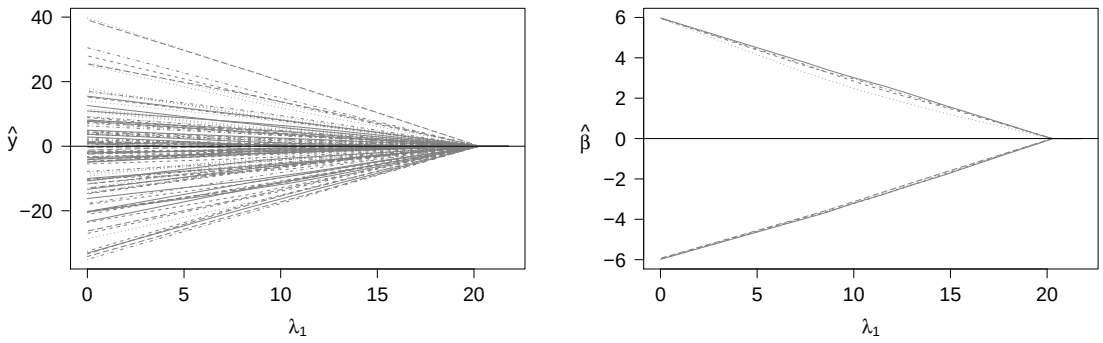

Figure 3: The fitted values (left panel) and parameter (right panel) trajectories while data are simulated with $\boldsymbol{\beta}^\top = (6, 6, 6, 6, -6, -6, -6, -6)$.

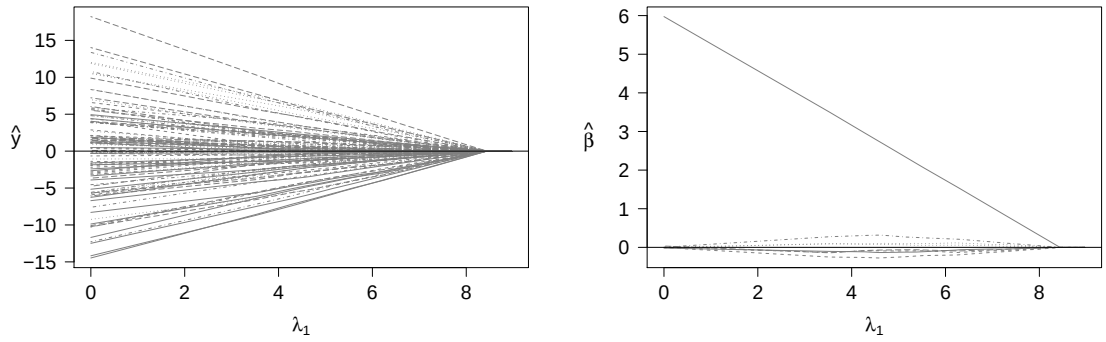

Figure 4: The fitted values (left panel) and parameter (right panel) trajectories while data are simulated with $\boldsymbol{\beta}^\top = (6, 0, 0, 0, 0, 0, 0, 0)$.

The design matrix for the quadratic penalization is directly linked to a prior distribution. This view provides a fast method of estimation for the quadratic penalization constant through the maximum marginal likelihood. The marginal likelihood can be successfully maximized and be used as an alternative to cross-validation.

The design matrix $\mathbf{D}_1$ defines sparsity constraints on the parameters. Solution to this problem is found by the generalized lasso path algorithm (Tibshirani and Taylor, 2011). For a given λ_1 and λ_2 the path algorithm for the generalized lasso can be used by little modification on the response vector and the regression design matrix to fit the generalized elastic net regression.

As a quick example, we applied the generalized elastic net regression to penalize the fitted values. We surprisingly observed that shrinkage appears over fitted values by increasing the ℓ_1 penalizing constant, but not sparsity. It looks that the sparsity behaviour of the lasso penalty over a linear combination of parameters depends on the design matrix $\mathbf{D}_1$ that produces the linear combinations.

References

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) Least angle regression. *The Annals of Statistics* **32**(2), 407–499.

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

Frank, L. E. and Friedman, J. H. (1993) A statistical view of some chemometrics regression tools. *Technometrics* **35**(2), 109–135.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **58**(1), 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **67**(1), 91–108.

Tibshirani, R. J. and Taylor, J. (2011) The solution path of the generalized lasso. *The Annals of Statistics* **39**(3), 1335–1371.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **67**, 301–320.