Least Absolute Shrinkage is Equivalent to Quadratic Penalization

Yves Grandvalet
Heudiasyc, UMR CNRS 6599, Université de Technologie de Compiègne,
BP 20.529, 60205 Compiègne Cedex, France
Yves.Grandvalet@hds.utc.fr

Abstract

Adaptive ridge is a special form of ridge regression, balancing the quadratic penalization on each parameter of the model. This paper shows the equivalence between adaptive ridge and lasso (least absolute shrinkage and selection operator). This equivalence states that both procedures produce the same estimate. Least absolute shrinkage can thus be viewed as a particular quadratic penalization.

From this observation, we derive an EM algorithm to compute the lasso solution. We finally present a series of applications of this type of algorithm in regression problems: kernel regression, additive modeling and neural net training.

1 Introduction

In supervised learning, we have a set of explicative variables $\boldsymbol{x}$ from which we wish to predict a response variable y. To solve this problem, a learning algorithm is used to produce a predictor $\widehat{f}(\boldsymbol{x})$ from a learning set $s_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{\ell}$ of examples. The goal of prediction may be:

- to provide an accurate prediction of future responses, accuracy being measured by a user-defined loss function;
- to quantify the effect of each explicative variable in the response;
- to better understand the underlying phenomenon.

Penalization is extensively used in learning algorithms. It is used to decrease the predictor variability to improve the prediction accuracy. It is also expected to produce models with few non-zero coefficients if interpretation is planned.

Ridge regression and subset selection are the two main penalization procedures. The former is stable, but does not shrink parameters to zero, the latter gives simple models, but it is unstable [1]. These observations motivated the search for new penalization techniques such as garrotte, non-negative garrotte [1], and lasso (least absolute shrinkage and selection operator) [2].

Adaptative noise injection [3] was proposed as a means to automatically balance penalization on different coefficients. Adaptive ridge regression is its deterministic version. Section 2 presents adaptive ridge. The equivalence of this special form of ridge regression with the lasso is briefly shown in section 3. This result connects least absolute shrinkage to quadratic penalization. Thanks to this link, the EM algorithm of

section 4 computes the lasso solution. A series of its possible applications is given in section 5.

2 Adaptative ridge regression

For clarity of exposure, the formulae are given here for linear regression with quadratic loss. The predictor is defined as $\widehat{f}(\boldsymbol{x}) = \boldsymbol{\beta}^T \boldsymbol{x}$, with $\boldsymbol{\beta}^T = (\beta_1, \ldots, \beta_d)$. Adaptive ridge is a modification of the ridge estimate, which is defined by the quadratic constraint $\sum_{j=1}^{d} \beta_j^2 \leq C$ applied to the parameters. It is usually computed by minimizing the Lagrangian

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}} \sum_{i=1}^{\ell} \big(\sum_{j=1}^{d} \beta_j x_{ij} - y_i \big)^2 + \lambda \sum_{j=1}^{d} \beta_j^2 \quad , \tag{1}$$

where λ is the Lagrange multiplier varying with the bound C on the norm of the parameters. When the ordinary least squares (OLS) estimate maximizes likelihood[1], the ridge estimate may be seen as a maximum a posteriori estimate. The Bayes prior distribution is a centered normal distribution, with variance proportional to $1/\lambda$.

This prior distribution treats all covariates similarly. It is not appropriate when we know that all covariates are not equally relevant.

The garrotte estimate [1] is based on the OLS estimate $\widehat{\boldsymbol{\beta}}^{\circ}$. The standard quadratic constraint is replaced by $\sum_{j=1}^{d} \beta_j^2 / \widehat{\beta}_j^{\circ 2} \leq C$. The coefficients with smaller OLS estimate are thus more heavily penalized.

Other modifications are better explained with the prior distribution viewpoint. Mixtures of Gaussians may be used to cluster different set of covariates. Several models have been proposed, with data dependent clusters [4], or classes defined a priori [5]. The automatic relevance determination model [6] ranks in the latter type. Following [3], we propose to use such a mixture, in the form

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}} \sum_{i=1}^{\ell} \big(\sum_{j=1}^{d} \beta_j x_{ij} - y_i \big)^2 + \sum_{j=1}^{d} \lambda_j \beta_j^2 \quad . \tag{2}$$

Here, each coefficient has its own prior distribution. The priors are centered normal distributions with variances proportional to $1/\lambda_j$. To avoid the simultaneous estimation of these d hyper-parameters by trial, the constraint

$$\frac{1}{d} \sum_{j=1}^{d} \frac{1}{\lambda_j} = \frac{1}{\lambda} \qquad , \ \lambda_j > 0 \tag{3}$$

is applied on $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^T$, where λ is a predefined value. This constraint is a link between the d prior distributions. Their mean variance is proportional to $1/\lambda$. The

[1] If the pairs $(\boldsymbol{x}_i, y_i)$ of the sample are independently and identically drawn from some distribution, and that some $\boldsymbol{\beta}^*$ exists, such that $Y_i = \boldsymbol{\beta}^{*T} \boldsymbol{x}_i + \varepsilon$, where ε is a centered normal random variable, then the empirical cost based on the quadratic loss is proportional to the log-likelihood of the sample. The OLS estimate $\widehat{\boldsymbol{\beta}}^{\circ}$ is thus the maximum likelihood estimate of $\boldsymbol{\beta}^*$.

values of λ_j are automatically[2] induced from the sample, hence the qualifier adaptative. Adaptativity refers here to the penalization balance on each coefficient, not to the tuning of the hyper-parameter λ.

3 The adaptative ridge estimate is the lasso estimate

We show below that adaptive ridge and least absolute value shrinkage are equivalent, in the sense that they yield the same estimate. First, the adaptive ridge estimate should be defined by another parameterization. The equations (2) and (3) in their present form may lead to divergent solutions for $\boldsymbol{\lambda}$ ($\lambda_j \to \infty$). Thus, we define new variables

$$\gamma_j = \sqrt{\frac{\lambda_j}{\lambda}}\beta_j \quad , \quad \text{and} \qquad c_j = \sqrt{\frac{\lambda}{\lambda_j}} \quad \text{for } j = 1, \ldots, d \ . \tag{4}$$

The optimization problem (2) with constraint (3) is then written

$$\begin{cases} (\widehat{c}, \widehat{\boldsymbol{\gamma}}) = \underset{(\boldsymbol{c},\boldsymbol{\gamma})}{\text{Argmin}} \sum_{i=1}^{\ell} \big(\sum_{j=1}^{d} c_j \gamma_j x_{ij} - y_i \big)^2 + \lambda \sum_{j=1}^{d} \gamma_j^2 \\ \text{subject to } \sum_{j=1}^{d} c_j^2 = d \ , \ c_j \geq 0 \ . \end{cases} \tag{5}$$

The definition of $\widehat{\boldsymbol{\gamma}}$ shows that adaptive ridge is related to garrotte. If $\widehat{c}$ is replaced by the OLS estimate in (5), we obtain the garrotte estimate.

After some algebra, we obtain the optimality conditions

$$\text{for } k = 1, \ldots, d \quad \begin{cases} \sum_{i=1}^{\ell} x_{ik} \big(\sum_{j=1}^{d} \widehat{\beta}_j x_{ij} - y_i \big) + \dfrac{\lambda}{d} \text{sign}(\widehat{\beta}_k) \sum_{j=1}^{d} |\widehat{\beta}_j| = 0 \\ \text{or } \widehat{\beta}_k = 0 \ . \end{cases} \tag{6}$$

These optimality conditions are the normal equations of the problem

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}} \sum_{i=1}^{\ell} \big(\sum_{j=1}^{d} \beta_j x_{ij} - y_i \big)^2 + \frac{\lambda}{d} \big(\sum_{j=1}^{d} |\beta_j| \big)^2 \ . \tag{7}$$

The estimate $\widehat{\boldsymbol{\beta}}$ (7) is equivalent to the lasso estimate, defined similarly to the ridge estimate by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmin}} \sum_{i=1}^{\ell} \big(\sum_{j=1}^{d} \beta_j x_{ij} - y_i \big)^2 \quad \text{subject to} \qquad \sum_{j=1}^{d} |\beta_j| \leq K \ . \tag{8}$$

The adaptive ridge estimate is thus the lasso estimate. The only difference in their definition is that the adaptive ridge estimate uses the constraint $(\sum_{j=1}^{d} |\beta_j|)^2 / d \leq K'$ instead of $\sum_{j=1}^{d} |\beta_j| \leq K$.

[2] Adaptive ridge, as ridge or lasso, is not scale invariant, so that the covariates should be normalized to produce sensible estimates.

4 A new algorithm for finding the lasso solution

Tibshirani [2] proposes to use quadratic programming to find the lasso solution, with $2d$ variables (positive and negative parts of β_j) and $2d+1$ constraints (signs of positive and negative parts of β_j plus constraint (8)). The writing (2) suggests to use the EM algorithm. At each step s, the EM algorithm estimates the optimal parameters $\lambda_j^{(s)}$ of the Bayes prior based on the estimate $\beta_j^{(s-1)}$, and then maximizes the posterior to compute the current estimate $\beta_j^{(s)}$. Again, we use the parameterization $(\boldsymbol{\gamma}, \boldsymbol{c})$ (4) instead of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ to avoid divergence:

$$\left\{ \begin{array}{l} c_j^{(s+1)^2} = \dfrac{d\gamma_j^{(s)^2}}{\sum_{k=1}^{d} \gamma_k^{(s)^2}} \\ \boldsymbol{\gamma}^{(s+1)} = \left(\mathrm{diag}(\boldsymbol{c}^{(s+1)}) X^T X \mathrm{diag}(\boldsymbol{c}^{(s+1)}) + \lambda I\right)^{-1} \mathrm{diag}(\boldsymbol{c}^{(s+1)}) X^T \boldsymbol{y} \end{array} \right. , \tag{9}$$

where $X_{ij} = x_{ij}$, I is the identity matrix, and $\mathrm{diag}(\boldsymbol{c})$ is the square matrix with the vector $\boldsymbol{c}$ on its diagonal.

The algorithm can be initialized by the ridge or the OLS estimate. In the latter case, $\boldsymbol{\beta}^{(1)}$ is the garrotte estimate. This observation confirms the connection put forward in the previous section between garrotte and lasso.

Practically, if $\gamma_j^{(s)}$ is small compared to numerical accuracy, then $c_j^{(s+1)}$ is set to zero. In turn, $\gamma_j^{(s+1)}$ is zero, and the system to be solved in the M-step to determine $\boldsymbol{\gamma}$ can be reduced to the other variables. If c_j is set to zero at any time during the optimization process, the final estimate $\widehat{\beta}_j$ will be zero. The computations are simplified, but it is not clear whether global convergence can be obtained with this algorithm. It is easy to show the convergence towards a local minimum, but we did not find general conditions ensuring global convergence. If these conditions exist, they rely on initial conditions.

Finally, we stress that the optimality conditions for $\boldsymbol{c}$ (or in a less rigorous sense for $\boldsymbol{\lambda}$) do not depend on the first part of the cost minimized in (2). In consequence, *the equivalence between adaptive ridge and lasso holds for any model/loss function*. The EM algorithm can be applied to these other problems, without modifying the E-step.

5 Applications

Adaptive ridge regression may be applied to a variety of problems. They include kernel regression, additive modeling and neural net training.

Kernel smoothers [7] can benefit from the sparse representation given by soft-thresholding methods like lasso. For these regressors, $\widehat{f}(\boldsymbol{x}) = \sum_{i=1}^{\ell} \beta_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \beta_0$, there are as many covariates as pairs in the sample. The quadratic procedure of lasso with $2\ell + 1$ constraints becomes computationally expensive, but the EM algorithm of adaptive ridge is reasonably fast to converge. An example of result is shown on figure 1 for the motorcycle dataset [7]. On this example, the final fitting uses only a few kernels (17 out of 133).

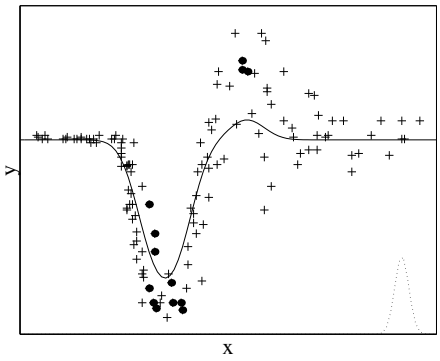

Figure 1: adaptive ridge applied to kernel smoothing on motorcycle data. The plus are data points, and dots are the prototypes corresponding to the kernels with non-zero coefficients. The Gaussian kernel used is represented dotted in the lower right-hand corner.

Girosi [8] showed an equivalence between a version of least absolute shrinkage applied to kernel smoothing, and Support Vector Machine (SVM) [9]. However, adaptive ridge, as applied here, is not equivalent to SVM fitting, because the response variables are noisy and the cost minimized is different. The prototypes shown on figure 1 are thus very different from the support vectors that would be obtained by SVM fitting.

Additive models [10] are sums of univariate functions, $\widehat{f}(\boldsymbol{x}) = \sum_{j=1}^{d} \widehat{f}_j(x_j)$, where the $\widehat{f}_j$ are smooth but unspecified functions. These models are easily represented and thus interpretable, but they require the choice of covariates to be included in the model, and of the smoothness of $\widehat{f}_j$.

In the form presented in the two previous sections, adaptive ridge regression penalizes differently each individual coefficient, but it is easily extended to the pooled penalization of coefficients. If the functions $\widehat{f}_j$ are defined as solution of a cost in the form "data misfit plus quadratic penalization", then adaptive ridge may be used to balance the penalization parameters on each $\widehat{f}_j$. A simulated example in dimension three is shown on figure 2. The fitted univariate functions are plotted for four values of λ. As λ increases, there is more penalization applied to the third (irrelevant) variable, then to the second variable, and finally to the first variable. The two first features are relevant, but the dependency on the first feature is smoother, hence easier to capture and more relevant for the predictor based on local averaging.

This generalization can also be applied to Multi-Layered Perceptrons to control the complexity of the fit. If weights are penalized individually, adaptive ridge is equivalent to the lasso. If weights are pooled by incoming and outcoming weights of a unit, node pruning is performed. Another interesting configuration gathers outcoming weights from each input unit, and incoming weights from each output unit (one set per input plus one per output). The goal here is to penalize/select the input variables according to their relevance, and each output variable according to the smoothness of the corresponding mapping [11].

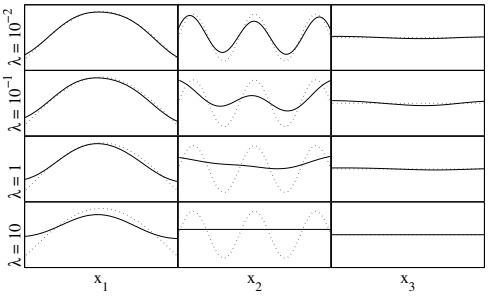

Figure 2: adaptive ridge applied to additive modeling on simulated data. The solid curves are the estimates of the univariate functions (Gaussian kernel smoothers) for different values of the overall tuning parameter λ. The true function is represented in dotted line.

References

[1] L. Breiman. Heuristics of instability and stabilization in model selection. The Annals of Statistics 1996, 24(6):2350–2383.

[2] R.J. Tibshirani. Regression shrinkage and selection via the lasso. Technical report, University of Toronto, June 1994.

[3] Y. Grandvalet and S. Canu. Adaptive noise injection for input variables relevance determination. In: *ICANN'97*, Springer-Verlag 1997, pp 463–468.

[4] S.J. Nowlan and G.E. Hinton. Simplifying neural networks by soft weight-sharing. Neural Computation 1992, 4(4):473–493.

[5] D.J.C. MacKay. A practical Bayesian framework for backprop networks. Neural Computation 1992, 4(3):448–472.

[6] R. M. Neal. Bayesian Learning for Neural Networks. Lecture Notes in Statistics. Springer-Verlag, New York, 1996.

[7] W. Härdle. Applied Nonparametric Regression, volume 19 of Economic Society Monographs. Cambridge University Press, New York, 1990.

[8] F. Girosi. An equivalence between sparse approximation and support vector machines. Technical Report 1606, M.I.T. AI Laboratory, Cambridge, MA., 1997.

[9] V.N. Vapnik. The Nature of Statistical Learning Theory. Springer Series in Statistics. Springer-Verlag, New York, 1995.

[10] T.J. Hastie and R.J. Tibshirani. Generalized Additive Models, volume 43 of Monographs on Statistics and Applied Probability. Chapman & Hall, New York, 1990.

[11] S. Canu, Y. Grandvalet, and M.-H. Masson. Black-box software sensor design for environmental monitoring. In: *ICANN'98*, Springer-Verlag 1998.