Variable selection for high-dimensional data: with applications in molecular biology

Bühlmann, Peter
ETH Zurich, Seminar für Statistik
Leonhardstrasse 27
CH-8092 Zürich, Switzerland
E-mail: buhlmann@stat.math.ethz.ch

In many application areas, the number of covariates is very large, e.g. in the thousands, while the sample size is quite small, e.g. in the dozens. In such high-dimensional settings, standard exhaustive search methods for variable selection are computationally infeasible and forward selection methods are typically very unstable yielding poor results.

We will show that ℓ^1-penalty methods, i.e. the Lasso (Tibshirani, 1996), can be very useful as a first stage: with high probability, the (mathematically) true model is a subset of the estimated model. Moreover, some adaptations correct Lasso's overestimation behavior yielding consistent variable selection schemes, and their exhaustive computation can be done very efficiently. In addition, when having multiple datasets, e.g. over different time points, we propose the new Smoothed Lasso (Meier and Bühlmann, 2007) which can lead to markedly improved variable selection and prediction in such data-structures. Finally, Boosting algorithms for complex, large-scale problems share some similarities with ℓ^1-penalization. The insights for adapted ℓ^1-penalty methods motivate a new Twin Boosting algorithm (Bühlmann, 2006b) for improved feature selection in much broader contexts than with the Lasso and its modifications.

Introduction

Our framework for high-dimensional data is as follows. We have observations

$$(1) \quad (X_1, Y_1), \ldots, (X_n, Y_n)$$

with p-dimensional covariates $X_i \in \mathcal{X} \subset \mathbb{R}^p$ and univariate response variables $Y_i \in \mathcal{Y} \subset \mathbb{R}$. Typically, we assume that the pairs (X_i, Y_i) are independent, identically distributed (i.i.d.) but the generalization to stationary processes (assuming reasonable decay of dependencies, e.g. rates of mixing) poses no essential problem. We say that the problem is *high-dimensional* if $p \gg n$.

A simple yet very useful model for high-dimensional data is a linear model

$$(2) \quad Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i \ (i = 1, \ldots, n),$$

where $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d., independent of $\{X_i;\ i = 1, \ldots, n\}$ and with $\mathbb{E}[\varepsilon_i] = 0$. For simplicity and without loss of generality, we assume that the intercept is zero and that all covariates are measured on the same scale, e.g. $\mathrm{Var}(X_i^{(j)}) \equiv$ const. for all j; both of these assumptions can be approximately achieved by mean centering and scaling with the standard deviation. The only unusual aspect of the linear model in (2) is the fact that $p \gg n$ and thus, the ordinary least squares estimator is not unique and will heavily overfit the data. There are numerous approaches for regularization in high-dimensional estimation: we focus on using the ℓ^1-penalty or some algorithmic approach within the framework of boosting algorithms where the latter originates from the machine learning community. Note that the model in (2) also arises when using an overcomplete dictionary of basis functions, e.g. in signal processing.

The Lasso for high-dimensional linear models

Estimation of the parameters in model (2) can be done with the Lasso (Tibshirani, 1996):

$$(3) \quad \hat{\beta}(\lambda) = \mathrm{argmin}_\beta(n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)})^2 + \lambda\|\beta\|_1),$$

where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$ and $\lambda \geq 0$ is a penalty parameter. The name Lasso stands for Least Absolute Shrinkage and Selection Operator: indeed, the estimator has the property that it does variable selection in the sense that $\hat{\beta}_j(\lambda) = 0$ for some j's (depending on the choice of λ) and $\hat{\beta}_j(\lambda)$ can be thought as a shrunken least squares estimator (where the shrinkage factor is zero if selection takes place).

It is well known that the optimization in (3) is convex, enabling efficient computation of the estimator. In fact, the LARS algorithm (Efron et al., 2004) computes all solutions $\hat{\beta}(\lambda)$, when varying λ over all possible submodels (see formula (4)), with $O(np\min(n,p))$ operation counts. Such algorithms which compute all solutions are known as path-following algorithms. Thus, for selecting a tuning parameter λ from data, we can efficiently compute all solutions and choose a good λ via e.g. a cross-validation score.

Prediction. We use the term prediction whenever the target is the estimation of the regression function $\mathbb{E}[Y|X = x] = \sum_{j=1}^{p}\beta_j x^{(j)}$ which is also the relevant quantity for predicting a new observation. Greenshtein and Ritov (2004) have proved that the Lasso consistently estimates the regression function as sample size $n \to \infty$ in the high-dimensional setting: roughly speaking, they assume that the number of covariates $p = p_n$ is allowed to grow like an arbitrary polynomial function in n, i.e. $p_n = Cn^\alpha$ for $0 \leq \alpha < \infty$ (and thus $p \gg n$), but the true underlying regression function has to be sparse in the sense that $\|\beta_n\|_1 = o((n/\log(n))^{1/4})$ (the true underlying parameter vector $\beta = \beta_n$ becomes a function of n as the dimension $p = p_n$ changes with n).

Besides this nice theoretical consistency result, it has been reported in various empirical applications that the Lasso has good or even excellent prediction capacity in high-dimensional problems. For example, the cross-validated misclassification error in a difficult (noisy) binary classification problem about lymph node status in breast cancer ($n = 49$) based on the expressions of $p = 7129$ genes is:

Lasso	L_2Boosting	FPLR	DLDA	SVM
21.1%	17.7%	35.25%	36.12%	36.88%

The following abbreviations have been used: L_2Boosting for boosting with the squared error loss, FPLR for forward ℓ^2-norm penalized logistic regression, DLDA for diagonal linear discriminant analysis (using the best 200 genes with respect to the 2-sample Wilcoxon test statistic), SVM for support vector machine with radial basis function (using the best 200 genes as in DLDA).

Variable selection. The problem of variable selection for a high-dimensional linear model in (2) is important. In many areas of application, the primary interest is about the relevance of covariates. As the number of sub-models is 2^p, computational feasibility is an important concern. The usual variable selection procedure using least squares and a penalty which involves the number of parameters in the candidate sub-model (e.g. AIC, BIC, MDL) is infeasible to compute exhaustively. Forward selection strategies are computationally fast but they can be very instable (Breiman, 1996). The requirement of computational feasibility and statistical accuracy can be met by the Lasso from (3); as described below, we will argue to use the Lasso not just once but in two (or more) stages.

We will first build up the methodology and theory by using the Lasso only once. Since the estimator is doing some variable selection, i.e. some of the coefficients are exactly zero ($\hat{\beta}_j(\lambda) = 0$ from some j's, depending on λ), we use it as follows:

$$\widehat{\mathcal{M}}(\lambda) = \{1 \leq j \leq p;\ \hat{\beta}_j(\lambda) \neq 0\}.$$

Note that no significance testing involved. The Lasso variable selection is computationally very efficient, requiring only $O(np\min(n,p))$ operation counts to compute all possible Lasso sub-models

$$(4) \quad \widehat{\mathcal{SUB}} = \{\widehat{\mathcal{M}}(\lambda);\ \text{all } \lambda\}.$$

It is worth pointing out that the number of sub-models is $|\widehat{\mathcal{SUB}}| = O(n)$.

The following asymptotic result shows that the Lasso can be very useful (in a first stage) for variable selection. To capture high-dimensionality of the model (2) in an asymptotic sense, we allow that the dimension $p = p_n$ and the coefficients $\beta_j = \beta_{j,n}$ depend on the sample size n. As the coefficients can change with n, the true model $\mathcal{M}_{true} = \mathcal{M}_{true,n} = \{1 \le j \le p_n;\ \beta_{j,n} \neq 0\}$ depends in general on n as well.

Theorem 1 (Meinshausen and Bühlmann, 2006)
Consider the linear model in (2) and assume that: (i) $p = p_n = O(n^\alpha)$ for $0 \le \alpha < \infty$ as $n \to \infty$, i.e. high-dimensionality; (ii) $\|\beta_n\|_0 = \sum_{j=1}^{p_n} I(\beta_{j,n} \neq 0) = O(n^\kappa)$ for $0 \le \kappa < 1$, i.e. sparsity; (iii) if $\beta_{j,n} \neq 0$, then $|\beta_{j,n}| \gg n^{-1/2+\delta}$ $(0 < \delta \le 1/2)$, i.e. the coefficients are outside the $1/\sqrt{n}$ range; (iv) the LfV (Lasso for variable selection) condition, as discussed briefly in Comment 2 below. Then, for $\lambda \sim \text{const.} n^{-1/2+\xi}$ with some suitable $\xi > 0$,

$$\mathbb{P}[\widehat{\mathcal{M}}_n(\lambda) = \mathcal{M}_{true,n}] = 1 - O(\exp(-\text{const.} n^{\xi/2}) \ \to 1 \ (n \to \infty).$$

More details are given in Meinshausen and Bühlmann (2006). Some complementing comments to Theorem 1 are instructive.

Comment 1: Consider the prediction optimal parameter supplied by an oracle:
$\lambda^* = \lambda_n^* = \ \text{argmin}_\lambda \mathbb{E}[(Y_{new} - \sum_{j=1}^{p_n} \hat{\beta}_{j,n}(\lambda) X_{new}^{(j)})^2]$, where (X_{new}, Y_{new}) are independent copies of (X_i, Y_i). Then, under the conditions of the Theorem:

$$\begin{aligned} &\mathbb{P}[\widehat{\mathcal{M}}_n(\lambda_n^*) \supset \mathcal{M}_{true,n}] \ \to 1 \ (n \to \infty), \\ (5) \qquad &\mathbb{P}[\widehat{\mathcal{M}}_n(\lambda_n^*) = \mathcal{M}_{true,n}] \ \to 0 \ (n \to \infty), \end{aligned}$$

see Meinshausen and Bühlmann (2006). Thus, the prediction optimal Lasso model is too large, containing with high probability the true model.

Comment 2: The LfV condition mentioned in assumption (iv) of Theorem 1 is sufficient and necessary for consistent model selection. In addition, it is restrictive and it fails to hold if the design matrix $\mathbf{X} = [X_i^{(j)}]_{i=1,\ldots,n;j=1,\ldots,p}$ exhibits a too strong degree of linear dependence (of course, if $p > n$, the matrix has linearly dependent columns). Details about this condition can be found in Zhao and Yu (2006) who gave a more accessible assumption, the irrepresentable condition, which is equivalent to the LfV condition.

Variable filtering. Comment 2 indicates a kind of negative result on the Lasso for variable selection. There is no way around the LfV (or irrepresentable) condition as it is sufficient and necessary. But if we require less than variable selection, the situation is quite different. Under much weaker assumptions than the LfV condition, it has been shown that for some suitable range of λ (including the prediction optimal λ^* under some assumptions)

$$(6) \quad \|\hat{\beta}_n(\lambda) - \beta_n\|_q \ \to 0 \text{ in probability } (n \to \infty),$$

where $q \in \{1, 2\}$ and $\|\beta\|_q = (\sum_j |\beta_j|^q)^{1/q}$; see van de Geer (2006), Bunea et al. (2006), Meinshausen and Yu (2006), Zhang and Huang (2007). A result of this type has been established first by Candes and Tao (2007) for the case of the Dantzig selector instead of the Lasso.

The result in (6) has fairly trivial but interesting implications in terms of variable importance and also variable selection. Most often, the main interest in practice is to find the covariates having substantial absolute coefficients $|\beta_{j,n}|$. More formally, for some $C > 0$, define the substantial covariates as

$$\mathcal{M}_{subst(C),n} = \{1 \le j \le p_n;\ \inf_{n\in\mathbb{N}} |\beta_{j,n}| \ge C\}.$$

Clearly, using the result in (6) which holds under much weaker assumptions than the LfV condition:

(7) for any fixed $0 < C < \infty$: $\mathbb{P}[\widehat{\mathcal{M}}_n(\lambda_n^*) \supset \mathcal{M}_{subst(C),n}] \to 1$ $(n \to \infty)$,

where λ_n^* is for prediction optimality. This result can be generalized to

(8) for $C_n \gg (\log(p_n)\|\beta_n\|_0/n)^{1/4}$: $\mathbb{P}[\widehat{\mathcal{M}}_n(\lambda_n^*) \supset \mathcal{M}_{subst(C_n),n}] \to 1$ $(n \to \infty)$,

where $\|\beta_n\|_0 = \sum_{j=1}^{p_n} I(\beta_{j,n} \neq 0)$, see Meinshausen and Yu (2006), and similar results are given by Zhang and Huang (2006). For the case where all variables are either ineffective (with corresponding coefficients equal to zero) or substantial with fixed $0 < C < \infty$, i.e. $\mathcal{M}_{true,n} = \mathcal{M}_{subst(C),n}$, formula (7) coincides wit the first line in (5) (but assuming much weaker conditions for (7) than for the derivation of (5) via Theorem 1; the latter has been the first mathematically rigorous argument).

We refer to the property in (7) or in (8) as *variable filtering*: the Lasso estimated model finds with high probability the substantial covariates.

The adaptive Lasso: from variable filtering to variable selection in a second stage

An interesting approach to correct Lasso's overestimation behavior, see formulae (5), (7) and (8), is given by the adaptive Lasso (Zou, 2006) which replaces the ℓ^1-norm penalty by a re-weighted version. For a linear model as in (2), it is defined as a two-stage procedure:

$$(9)\quad \hat{\beta}_{adapt}(\lambda) = \mathrm{argmin}_\beta(n^{-1}\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j X_i^{(j)})^2 + \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{|\hat{\beta}_{init,j}|}),$$

where $\hat{\beta}_{init}$ is an initial estimator. In the high-dimensional context, we propose to use the Lasso from a first stage as the initial estimator, tuned in a prediction optimal way, i.e. $\hat{\beta}_{init} = \hat{\beta}(\lambda^*)$ as in (3), and the prediction optimal λ^* can be estimated from cross-validation. The adaptive Lasso has the following obvious property:

$$\hat{\beta}_{init,j} = 0 \ \Rightarrow\ \hat{\beta}_{adapt,j}(\lambda) = 0 \text{ for all } \lambda > 0.$$

Moreover, if $|\hat{\beta}_{init,j}|$ is large, the adaptive Lasso employs a small penalty (i.e. little shrinkage) which implies less bias. In fact, it is the problem of bias which causes Lasso's overestimation behavior. Figure 1 illustrates clear superiority of the adaptive Lasso over the Lasso: the adaptive Lasso is much more insensitive to many additional noise covariates.

We state here informally, as a nice result, that the prediction optimally tuned adaptive Lasso is consistent for variable selection: for the low-dimensional case with fixed dimension p, see Zou (2006); for the high-dimensional case, consistency is meant in the sense of finding the substantial covariates, see Huang, Ma and Zhang (2006).

The smoothed Lasso for many high-dimensional linear models: with an application to finding transcription factor binding sites

In quite a few applications in molecular biology (a concrete problem will be described below), we have multiple data-sets over time:

$$\{\mathbf{X}(t), \mathbf{Y}(t);\ t = 1, \ldots, N\},$$

(10) $\mathbf{X}(t)_{n(t)\times p}$ the design matrix at t, $\mathbf{Y}(t)_{n(t)\times 1}$ the response vector at t,

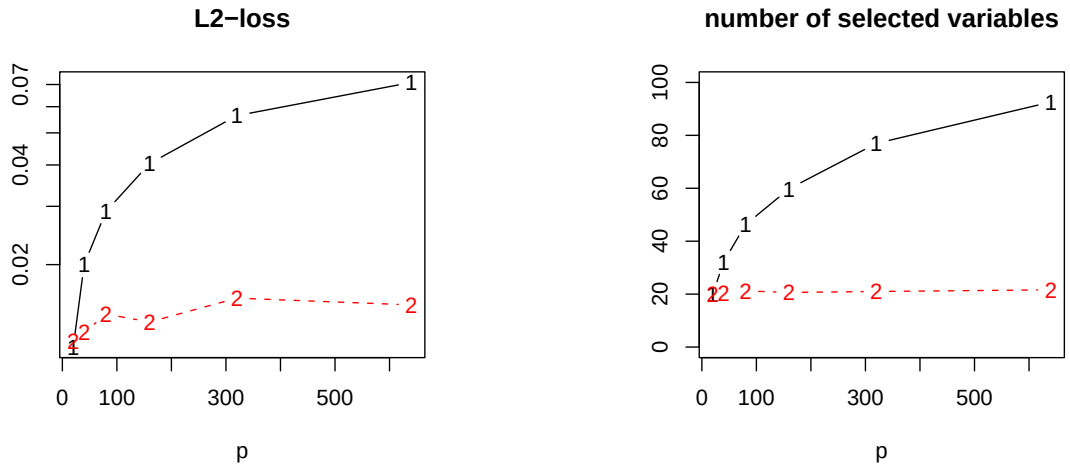

Figure 1: Left panel: $\mathbb{E}[(\sum_{j=1}^{p}(\hat{\beta}_j - \beta_j)X^{(j)})^2]$. Right panel: number of selected variables $\|\hat{\beta}\|_0$. Black (1): prediction optimal Lasso; Red (2): prediction optimal adaptive Lasso using prediction optimal Lasso as initial estimator. underlying linear model as in (2) with $\|\beta\|_0 = 20$ effective variables and $p - 20$ ineffective noise covariates (for various p), sample size $n = 300$.

that is, for every time point t we have data as in (1) with sample size $n(t)$.

As a simple model, we consider linear models with slowly changing high-dimensional $p \times 1$ parameter vector $\beta(t)$:

$$\mathbf{Y}(t) = \mathbf{X}(t)\beta(t) + \varepsilon(t),\ t = 1, \ldots, N, \tag{11}$$

where the $\varepsilon(t)$'s are independent with $\mathbb{E}[\varepsilon(t)] = 0$, $\mathrm{Cov}(\varepsilon(t)) = \sigma^2 I_{n(t)\times n(t)}$. We have in mind the situation where $p \gg n(t)$ and where the high-dimensional vector $\beta(\cdot)$ is slowly varying.

We then propose to use the smoothed Lasso (Meier and Bühlmann, 2007). For simplicity, consider the case where $n(t) \equiv n$ is constant for all t:

$$\hat{\beta}(\tau) = \ \mathrm{argmin}_\beta \sum_{t=1}^{N} K(\frac{\tau - t}{h})(n^{-1}\|\mathbf{Y}(t) - \mathbf{X}(t)\beta\|_2^2 + \lambda\|\beta\|_1),$$

where $K(\cdot)$ is a symmetric kernel function, e.g. $K(u) = I(u \in [-1/2, 1/2])$, and $h > 0$ is a bandwidth parameter. The smoothed adaptive Lasso is defined in analogy to the non-smoothed version.

It isn't a big surprise that we can do better with the smoothed (adaptive) Lasso than working individually at every time point. However, an asymptotic calculation helps to understand how quickly we will gain to improve the convergence rate.

Theorem 2 (Meier and Bühlmann, 2007)
Consider the time-varying linear model in (11) and assume that the individual coefficient functions $\beta_j(\cdot)$ are twice continuously differentiable. Suppose that the (optimally tuned) non-smoothed Lasso (for a single time point t) has convergence rate $\|\hat{\beta}(t) - \beta(t)\|_2^2 \asymp a_n$ in probability. Then, for a single time point t, the (optimally tuned) smoothed Lasso has faster convergence rate than the non-smoothed Lasso if $N \gg (a_n^{-1}\|\beta(t)\|_0)^{1/4}$.

The detailed assumptions and a proof are given in Meier and Bühlmann (2007). It is worth emphasizing that $(a_n^{-1}\|\beta(t)\|_0)^{1/4}$ is often a fairly small number: $a_n^{-1/4}$ is at most of order $n^{1/4}$ (and substantially smaller for high-dimensional problems). Thus, $N \geq 10$ often seems sufficient (if e.g. $n \leq 100$ and $\|\beta(t)\|_0 \leq 50$) to mimic the asymptotic requirement $N \gg (a_n^{-1}\|\beta(t)\|_0)^{1/4}$. Also note that a better convergence rate for $\|\hat{\beta}(t)) - \beta(t)\|_2^2$ implies improved ability for variable filtering, see

formulae (7) and (8). Meier and Bühlmann (2007) report substantial gains in prediction (squared error prediction for future response Y) of the smoothed (adaptive) Lasso over the univariate (adaptive) Lasso: e.g. for an underlying sparse model with $N = 18,\ n = 100$ and $p = 5000$.

Application to searching for transcription factor binding sites. We apply the smoothed Lasso to a problem of motif regression (Conlon et al., 2003) for detecting unknown transcription factor binding sites in DNA of yeast. The data is as in (10) with $N = 12$ time points. The design matrices $X(t)$ contain the scores of $p = 666$ candidate motifs, upstream of $n(t) \equiv 500$ genes (a motif is a short subsequence of DNA bases, e.g. the "word" TTACCGGTCG; the score essentially describes the frequency of occurrence of the motif over a larger upstream region of a gene in the DNA sequence). The response vectors $Y(t)$ are expressions of the $n(t)$ genes under consideration. The goal is to identify the relevant covariates in the model (11) in the hope to find unknown true motifs (true transcription factor binding sites). Note that this problem is not really high-dimensional in the sense that $p \gg n(t)$: however, the dimensionality is still fairly large and in addition, we expect that there is positive correlation within the $n(t)$ samples, implying a lower "effective sample size" (i.e. in comparison to the case with independence, the variance is inflated). A snapshot of the results are displayed in Table 1. The smoothed adaptive Lasso yields substantially sparser estimates and hence fewer expected

time point	sel. var. Ada Lasso	sel. var. Smoothed Ada Lasso	prediction gain
3	9	4	-3.0%
6	91	33	8.3%
7	45	28	- 0.7%
8	20	15	0.6%
11	56	35	5.6%
12	41	30	1.5%

Table 1: Motif search in yeast. Time points with largest difference between (univariate) adaptive Lasso and smoothed adaptive Lasso. Number of selected variables (2nd and 3rd column) and prediction gain (squared error for future response Y) of the smoothed adaptive Lasso over the (univariate) adaptive Lasso (last column).

false positives; there is common consensus that the main problem in motif finding are too many false positives and thus, the smoothed adaptive Lasso is very useful. In addition, the squared error prediction performance of the smoothed adaptive Lasso is slightly better than for the adaptive Lasso: it should be added here that these kind of data are very noisy and the squared prediction error is dominated by the variance of the noise term, implying fairly similar prediction performance among different methods.

High-dimensional generalized linear models and other extensions

From a conceptual and methodological point of view, the Lasso and the adaptive Lasso carry over to generalized linear models

$$Y_i|X_i \text{ has a distribution from the exponential family,}$$
$$\mathbb{E}[Y_i|X_i] = \mu_i,\ g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_i^{(j)}, \tag{12}$$

where $g(\cdot)$ is a known link function. The Lasso is defined as penalized maximum likelihood estimator:

$$\hat{\beta}(\lambda) = \ \text{argmin}_\beta(-\ell(\beta; (X_1, Y_1), \ldots, (X_n, Y_n)) + \lambda \sum_{j=1}^{p} |\beta_j|), \tag{13}$$

where $\ell(\beta;\cdot)$ is the log-likelihood function. If the likelihood depends not only on β but also on some additional nuisance parameter vector ϕ, we would consider optimization over β and ϕ but the penalty term would only apply for $\beta_1,\ldots,\beta_p$, i.e. the intercept β_0 and the nuisance parameters would not be penalized. The adaptive Lasso is defined analogously with a re-weighted penalty as in (9).

There is a substantial body of empirical evidence that also in case of high-dimensional generalized linear models: (i) the Lasso is very useful for prediction and variable filtering; (ii) the adaptive Lasso corrects Lasso's overestimation behavior and is an excellent method for variable selection.

Despite the convexity of the optimization problem in (9), the computation of the Lasso for all submodels when varying λ, see formula (4), is more difficult than in the linear model case. The solutions can be computed approximately (the approximation is meant with respect to all submodels): approximate path-following algorithms (Park and Hastie, 2007) or efficient sequential methods on a grid of λ's can be used (Meier et al., 2007).

The mathematical theory has not been worked out in such details as for linear models, with the exception of van de Geer (2006) who establishes a nice oracle result for $\|\hat{\beta}_n(\lambda) - \beta_n\|_1$ in a high-dimensional generalized linear model.

The Group Lasso. In presence of factor variables (categorical variables), the parameterization in (2) or (12) matters for the estimator in (3) or (13), respectively, due to the nature of the penalization. An interesting way to deal with this problem is to group the parameter vector β. Denote by $\{\mathcal{G}_j;\ j = 1,\ldots,k\}$ a partition of the index set $\{1,\ldots,p\}$ with $k \le p$: it corresponds to the parameterization of the factor variables, e.g. a main effect of a factor with say 4 levels corresponds to a group of size 3 (because of using the sum-to-zero constraints for main effects). Yuan and Lin (2006) propose the Group Lasso: instead of (13), the penalty employs the ℓ^2-norm of the groups:

$$\hat{\beta}(\lambda) = \ \mathrm{argmin}_\beta(-\ell(\beta;(X_1,Y_1),\ldots,(X_n,Y_n)) + \lambda\sum_{j=1}^{k}\|\beta_{\mathcal{G}_j}\|_2),$$

where $\beta_{\mathcal{G}_j} = \{\beta_k;\ k \in \mathcal{G}_j\}$. Note that the ℓ^2-norm yields invariance under orthogonal transformations for the parameterization of a group. In addition, if the group-size is one, i.e. $|\mathcal{G}_j| = 1$, then $\|\beta_{\mathcal{G}_j}\|_2 = \|\beta_{\mathcal{G}_j}\|_1 = |\beta_{\mathcal{G}_j}|$.

As another example, the concept of the Group Lasso can also be used for high-dimensional additive modeling, where every additive function would be approximated with a basis expansion (e.g. B-splines) and its associated parameters would build a group.

Some asymptotic theory as well as efficient computational algorithms for the Group Lasso are given in Meier et al. (2006).

Boosting: an algorithmic approach

Boosting has been recognized to be an excellent method for high-dimensional data analysis. The most famous proposal has been the AdaBoost algorithm for classification (Freund and Schapire, 1996) which has attracted much attention in the machine learning and statistics community. Through the interpretation of boosting as a gradient descent method in function space (Breiman, 1998), boosting can be seen as an interesting regularization scheme for estimating a high-dimensional model, including model (2) as discussed in Bühlmann (2006a), and extensions to model (12) or also more general high-dimensional nonparametric models as described in Bühlmann and Hothorn (2006).

Connections to the Lasso. Efron et al. (2004) pointed out an intriguing connection between L_2Boosting, based on the squared error loss, for linear models and the Lasso in (3). They consider a version of L_2Boosting and prove approximate equivalence to the Lasso via different modifications of their computationally efficient least angle regression (LARS) algorithm. Despite the fact that L_2Boosting and

Lasso are not equivalent methods in general, it may be useful to interpret boosting as being "related" to ℓ^1-penalty methods.

The two-stage approach of the adaptive Lasso has its analogy (along the lines of the analogy of boosting with the Lasso) in the algorithmic boosting world, called Twin Boosting (Bühlmann, 2006b). The advantage of the boosting approach, in comparison to the Lasso, is its flexibility for general nonparametric models and mixed data-types. Regarding the latter, if we have high-dimensional continuous, ordinal and categorical variables, boosting or twin boosting with regression trees is computationally feasible and an excellent method for prediction and feature or variable selection: see Lutz (2006) who has won the WCCI 2006 (World Congress of Computational Intelligence) performance prediction challenge by using a boosting algorithm.

REFERENCES

Breiman, L. (1996). Bagging predictors. Machine Learning **24**, 123–140.

Breiman, L. (1998). Arcing classifiers. Ann. Statist. **26**, 801–849 (with discussion).

Bühlmann, P. (2006a). Boosting for high-dimensional linear models. Ann. Statist. **34**, 559–583.

Bühlmann, P. (2006b). Twin Boosting: improved feature selection and prediction. Preprint.

Bühlmann, P. and Hothorn, T. (2006). Boosting algorithms: regularization, prediction and model fitting. Preprint.

Bunea, F., Tsybakov, A. and Wegkamp, M. (2006). Sparsity oracle inequalities for the Lasso. Preprint.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. To appear in Ann. Statist.

Conlon, E.M., Liu, X.S., Lieb, J.D. and Liu, J.S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. Proc. Natl. Acad. Sci. USA **100**, 3339–3344.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. Ann. Statist. **32**, 407–499 (with discussion).

Freund, Y. and Schapire, R.E. (1996). Experiments with a new boosting algorithm. In Machine Learning: Proc. Thirteenth International Conference, pp. 148–156. Morgan Kauffman, San Francisco.

Greenshtein, E. and Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. Bernoulli **10**, 971–988.

Huang, J., Ma, S. and Zhang, C.-H. (2006). Adaptive LASSO for sparse high-dimensional regression models. Preprint.

Lutz, R.W. (2006). LogitBoost with trees applied to the WCCI 2006 performance prediction challenge datasets. Proc. of the IJCNN 2006.

Meier, L. and Bühlmann, P. (2007). The smoothed Lasso. In preparation.

Meier, L., van de Geer, S. and Bühlmann, P. (2006). The Group Lasso for logistic regression. Preprint.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. Annals of Statistics **34**, 1436–1462.

Meinshausen, N. and Yu, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Preprint.

Park, M.-Y. and Hastie, T. (2007). An L1 regularization-path algorithm for generalized linear models. To appear in J. Roy. Statist. Soc., Ser. B.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc., Ser. B, **58**, 267–288.

van de Geer, S. (2006). High-dimensional generalized linear models and the Lasso. Preprint.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. J. Roy. Statist. Soc., Ser. B., **68**, 49–67.

Zhang, C.-H. and Huang (2007). The sparsity and bias of the Lasso selection in high-dimensional linear regression. Preprint.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. J. Machine Learning Res. **7**, 2541–2563.

Zou, H. (2006). The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc., **101**, 1418–1429.