Sociedad de Estadística e Investigación Operativa
Test (2006) *Vol.* **15**, *No.* 2, pp. 271–344

Regularization in Statistics

Peter J. Bickel
Department of Statistics
University of California, Berkeley, USA
Bo Li
School of Economics and Management
Tsinghua University, China

Abstract

This paper is a selective review of the regularization methods scattered in statistics literature. We introduce a general conceptual approach to regularization and fit most existing methods into it. We have tried to focus on the importance of regularization when dealing with today's high-dimensional objects: data and models. A wide range of examples are discussed, including nonparametric regression, boosting, covariance matrix estimation, principal component estimation, subsampling.

Key Words: Regularization, linear regression, nonparametric regression, boosting, covariance matrix, principal component, bootstrap, subsampling, model selection.

AMS subject classification: Primary 62G08, 62H12; Secondary 62F12, 62G20, 62H25.

1 Introduction

The concept of regularization was first introduced in the context of solving integral equation numerically by Tikhonov (1943). As is well known if $f \in L_2(\mathbb{R})$ and $K(x,y)$ is a smooth kernel, the range of the operator A, $\mathcal{R}(A)$, $A : L_2(\mathbb{R}) \mapsto L_2(\mathbb{R})$ with $(Af)(y) \equiv \int K(x,y)f(x)\mathrm{d}x$ is dense in $L_2(\mathbb{R})$ but not onto. Thus, the inverse A^{-1} is ill-posed. The solution to the equation

$$Af = g \tag{1.1}$$

is hard to determine since approximations to g easily lie outside $\mathcal{R}(A)$. Tikhonov's solution was to replace (1.1) by the minimization of $\|Af - g\|^2 + \gamma W(f)$, where the *Tikhonov factor* $\gamma > 0$ is a regularization parameter and $W(f)$ is a smoothness penalty such as $\int [f'(x)]^2 \mathrm{d}x$. Numerical (finite

*Correspondence to: Peter J. Bickel. Department of Statistics, University of California, Berkeley, USA. E-mail: bickel@stat.Berkeley.EDU

dimensional) approximations to this problem are much stabler. Note that unless $\gamma = 0$, the solution will not satisfy (1.1).

There has been an enormous amount of work in statistics dealing with regularization in a wide spectrum of problems. An exhaustive survey is beyond our scope. We want to present a unifying view encompassing more recent developments. The main features of most current data are both size and complexity. The size may permit us to nonparametrically estimate quantities which are "unstable" and "discontinuous" functions of the underlying distribution of the data, with the density being a typical example. Complexity of the data, which usually corresponds to high dimensionality of observations, makes us attempt more and more complex models to fit the data. The fitting of models with a large number of parameters is also inherently unstable (Breiman, 1996). Both of these features, as we shall see in our examples, force us to regularize in order to get sensible procedures. For recent discussions of these issues from different points of view, see Donoho (2000) and Fan and Li (2006). We will consider only the asymptotics of regularization and only in the simplest context, i.i.d samples of size n of p dimensional vectors. The main issues are already quite clear in this context.

We will define regularization formally in Section 2. But, as we shall see, loosely, regularization is the class of methods needed to modify maximum likelihood to give reasonable answers in unstable situations. There are also a number of generic issues that will arise such as the reasons for choosing particular forms of regularization, how to determine the analogue of the Tikhonov factor γ which, as we shall see, is somewhat driven by our particular statistical goals, and last but not least, computational issues which are also critical nowadays. We shall discuss these questions in connection with examples as we proceed in this and Section 3, Section 4 and Section 5.

Variable selection and prediction. In statistics, the first instance of this type of problem arose in the context of multiple linear regression with continuous predictor variables, when the number of predictor variables is larger than the sample size. Suppose we observe an i.i.d sample $(Z_i, Y_i), i = 1, \cdots, n$, where $Z_i = (Z_i^{(1)}, \cdots, Z_i^{(p)})$. We model

$$Y_i = Z_i^T \beta + \epsilon_i \tag{1.2}$$

where $\epsilon_i, i = 1, \cdots, n$ are i.i.d $N(0, \sigma^2)$. In the case of $p > n$, the usual least squares equations "overfit". All observations are predicted perfectly, but

there are many solutions to the coefficients of the fit and new observations become not uniquely predictable. The classical solution to this problem was to try to reduce the number of variables by processes such as forward and backward regression with reduction in variables determined by hypothesis tests, see Draper and Smith (1998), for example. An alternative strategy that emerged (Hoerl and Kennard, 1970) was ridge regression, adding to the residual sum of squares $\sum_{i=1}^{n}(Y_i - Z_i^T\beta)^2$ a penalty, $\lambda\sum_{j=1}^{p}\beta_j^2$, which now yields a unique solution.

These methods, often actually have two aims,

(I) To construct a good predictor. The values of coefficients in the regression are then irrelevant.

(II) To give causal interpretations of the factors and determine which variables are "important".

Regularization is important for both aims. But, as well shall see, the appropriate magnitude of the regularization parameter may be governed by which aim is more important.

Goal (I) is the one which is primary in machine learning theory. The model postulated is nonparametric,

$$Y = m(Z) + \varepsilon \tag{1.3}$$

where $E(\varepsilon|Z) = 0$ and m is essentially unknown. A fundamental approach is to consider a family of basis functions $g_j(Z), j = 1, 2, \cdots$, such that m is arbitrarily well approximated in, for instance, the L_2 sense, $\inf_\beta E(m(Z) - \sum_{j=1}^{p}\beta_j g_j(Z))^2 \to 0$ as $p \to \infty$, where $\beta = (\beta_1, \cdots, \beta_p)^T$. A parametric model postulation with $g_j(Z) = Z^{(j)}, j = 1, \cdots, p$, corresponds to the linear model specification. Then, since, as we have seen, minimizing $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j g_j(Z_i))^2$ is unreasonable for $p >> n$, it is consistent with the penalty point of view to minimize

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p}\beta_j g_j(Z_i))^2 + \gamma Pen(\beta) \tag{1.4}$$

The ridge regression choice of $Pen(\beta) = \sum_{j=1}^{p}\beta_j^2$ is not nowadays the one attracting the greatest attention theoretically, but the "lasso", $Pen(\beta) =$

$\sum_{j=1}^{p} |\beta_j|$ (Tibshirani, 1996) is being studied extensively. This stems from the idea that, at least to a high degree of approximation, most $|\beta_j|$ in the best representation of $m(Z)$ as a linear combination of p basis elements $g_j(Z)$ in the L_2 sense are 0. That is, the representation is "sparse" in the sense of Donoho and Johnstone (1998). Then the "natural" penalty is

$$Pen(\beta) = \sum_{j=1}^{p} 1(|\beta_j| > 0) \tag{1.5}$$

an unpleasant function of β. Evidently, $\sum_{j=1}^{p} |\beta_j|$ is the closest convex member of the family of penalties $\sum_{j=1}^{p} |\beta_j|^{\alpha}, \alpha > 0$ to (1.5).

We shall discuss this approach, the critical choice of γ, and point to recent results as n and p tends to infinity in Section 3.

Minimizing subject to penalty (1.5) may also be seen as selecting a model including the variables with $\beta_j \neq 0$, following aim (II). This approach and its generalization to generalized linear and other models as well as related penalties has been developed by Fan and coworkers and others, see Fan and Li (2001), Fan and Peng (2004), Fan and Li (2006) and Zou and Hastie (2005). Note that, at least implicitly, this point of view implies that we believe a meaningful (sparse) representation in basis functions g_j.

$$m(Z) = \sum_{j=1}^{p^*} \beta_j g_j(Z) \tag{1.6}$$

is true for some $p^* << p$.

Penalization is far from the only form of regularization that has arisen in statistics. In the context of density estimation, binning in histograms is the oldest method, and kernel methods were proposed by Rosenblatt (1956) and Parzen (1962). In turn these methods led to Nadaraya-Watson estimation (Nadaraya, 1964; Watson, 1964) in nonparametric regression.

There are also methods which have appeared outside nonparametric regression contexts, where formulations such as semiparametric or generalized linear models do not capture the necessary structure. Here is the first.

Covariance and eigen structure estimation. Suppose P is the probability distribution of $(X_1, X_2, \cdots)$, a Gaussian process with mean 0 and

covariance matrix Σ_p for $(X_1, \cdots, X_p)$. Suppose that Σ_p has distinct eigenvalues given by $\lambda_p = (\lambda_{1p}, \cdots, \lambda_{pp})$, corresponding to orthonormal eigenvectors, $\nu_1, \cdots, \nu_p$. We can think of $\Sigma_p(P)$ as a parameter, and the vectors $\lambda_p(\Sigma_p)$ as functions of the main parameter. Then $\hat{\Sigma}_p, \hat{\lambda}_p, \hat{\nu}_p$, the empirical versions of these are consistent for fixed p. If $p \to \infty, \frac{p}{n} \to c > 0$, this is no longer true. Suppose $\Sigma_p = J_p$, the identity matrix which doesn't fall under our assumptions, but for which still, if p is fixed, $\hat{\lambda}_{jp} \xrightarrow{P} 1 = \lambda_{jp}$, for $1 \leq j \leq p$. Then it is well known (Wachter, 1978; Wigner, 1955) that the maximum eigenvalue $\hat{\lambda}_{pp} \xrightarrow{P} \lambda_{max} > 1$. Recently, Johnstone and Lu (2006) showed that if $\frac{p}{n} \to c > 0$ and $\Sigma_p = J_p + K_p$, where K_p is a degenerate matrix, all of whose eigenvalues except the top t are 0, then

$$\text{limsup}\big(E < \hat{\nu}_p, \nu_p > \big) < 1 \tag{1.7}$$

where $E < \hat{\nu}_p, \nu_p >$ is the expected inner product between the empirical and true eigenvectors corresponding to λ_{pp}. Regularization is needed and Johnstone and Lu (2006) suggest a method which yields consistency under their assumption. Bickel and Levina (2004) effectively show that in this case banding the matrix, replacing Σ_p by $\Sigma_{kp}(P)$, the matrix obtained by setting all entries with indices (i, j) such that $|i - j| > k$ equal to 0, yields consistency under much weaker conditions.

Subsampling and m out of n bootstrap. A final example where irregularity can occur in important situations is Efron's nonparametric bootstrap. Here we resample samples of size n from the empirical distribution (the sample) and then act as if these were samples from the unknown population. Breakdowns of this method have been noted by many authors, see Mammen (1992) and a more recent discussion in Bickel et al. (1997). We discuss a regularization method that has been proposed in this regard briefly.

In this paper, we propose to define what we mean by "regularization", a concept which encompasses all these situations. We proceed as follows. In Section 2, we introduce our general mathematical framework and define regularization in general, linking it to the examples we have cited, and pose what we view as the basic questions to be faced. In Section 3, we discuss nonparametric regression and classification in detail. In Section 4, we discuss estimation of high dimensional covariance matrices, their inverses and

eigenstructures and in Section 5, subsampling and the m out of n bootstrap. The discussion we give will be in terms of behavior asymptotic in the sample size, and sometimes dimension, though we will at least refer to confirmatory simulations. Thus when we talk of statistical procedures, we think of sequences of such procedures with the $n-$th one depending on the n observations available. This does not mean that conclusions only hold for $n = p = \infty$, rather, that we hope that, as seems to be the case in practice, the approximations are good for samples and dimensions of the size we expect.

2 What is regularization

Throughout we limit ourselves to the case where our observations $X_1, \cdots, X_n$ are i.i.d, taking values in a space $\mathcal{X}$, typically $\mathbb{R}^p$. We assume that their common distribution $P \in \mathcal{P}$, our model, which through most of our discussion, we assume is nonparametric, effectively all P, although we can and shall impose smoothness or other general properties on the members of $\mathcal{P}$. We let P_n denote the empirical distribution, placing mass n^{-1} at each observation.

For our treatment of covariance estimation it may be convenient to think of $\mathbf{X} = (X_1, X_2, \cdots, X_p, \cdots)^T$, as a stochastic process for which we have data of size n on the first p coordinates, and of the unknown P as living on $\mathbb{R}^\infty$. However, we will only be interested in estimating the covariance matrix of these first p coordinates.

Most statistical activities center around estimation or testing hypotheses or putting confidence regions on parameters, which we define as functions $\theta(P)$, mapping $\mathcal{P}$ into Θ. Θ is not necessarily just $\mathbb{R}$ or a Euclidean space. We shall limit ourselves almost exclusively to function valued parameters. For instance, suppose $P \in \mathcal{P}$ are characterized as having densities $f(\cdot)$, which are continuous. Then $\theta(P) = f(\cdot)$ is a parameter. If P is the joint distribution of (Z, Y), then $\theta(P) = E(Y|Z = \cdot)$, the regression function is a parameter. It will also be convenient for both Section 4 and Section 5 to think of parameters which themselves vary with n and p, $\theta^{(n,p)}(P)$. Thus, the covariance matrix Σ of $(X_1, \cdots, X_p)^T$, which we are interested in studying is $\theta^{(p)}(P)$ if we think of our observation as being $(X_1, X_2, \cdots)^T$. Similarly, the extreme percentile of the distribution of $X \in \mathbb{R}$, $F^{-1}(1)$ where F is the empirical distribution function of X,

typically equals ∞ and cannot be estimated, but $F^{-1}(1-\frac{1}{n})$, the quantile corresponding to the maximum of $X_1, \cdots, X_n$ can. We will usually suppress such dependence on p and n.

Any estimate $\hat{\theta}(X_1, \cdots, X_n)$ of $\theta(P)$ may, by sufficiency of the P_n, be thought of as a function $\theta_n(P_n)$, where the domain of θ_n is at least the possible empirical distributions and typically includes at least all finite discrete distributions on $\mathcal{X}$. The least we can require of an estimate (really a sequence of estimates) is *consistency*:

$$\rho(\hat{\theta}, \theta(P)) \xrightarrow{P} 0 \tag{2.1}$$

where ρ is Euclidean distance if Θ is Euclidean and ρ is a suitably defined metric, e.g., the L_2 distance, if Θ is a function space.

If $\mathcal{P}$ contains all discrete distribution, then the natural thing to use as an estimate of $\theta(P)$ is the "plug-in" estimate $\theta(P_n)$. For instance, if $\mathcal{X} = \mathbb{R}$, and $\theta(P)$ is the mean, which we represent as $\theta(P) = \int x \mathrm{d}P(x)$, then $\theta(P_n) = \int x \mathrm{d}P_n = \bar{X}$, the sample mean. If $\theta(P) = F(\cdot)$, where $F(x) = P(X \leq x)$, the cdf of X, then $\theta(P_n)$ is the empirical cdf, $\theta(P_n) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$. Consistency for plug-in estimates follows if

(a) θ is continuous in ϱ, for a given metric ϱ on $\mathcal{P}$.

(b) P_n is consistent with respect to ϱ. That is, $\varrho(P_n, P) \xrightarrow{P} 0$ if P is true.

In the usual situations, where Θ is Euclidean, $\theta \mapsto p(\cdot, \theta)$ is smoothly invertible, and $\theta(P_n)$ makes sense, consistency holds. But, consider the situation we have discussed, $\theta(P) = f(\cdot)$. Now the density. $\theta(P_n)$ doesn't make sense, since the discrete distributions do not belong to $\mathcal{P}$. What is done, in this case, and implicitly in all such situations we know about is *regularization*. We summarize a generic regularization process as,

(1) *A sequence of approximations.*

  (i) We construct a sequence θ_k defined on $\mathcal{P}$ and the discrete distributions, say on $\mathcal{M}$ such that $\rho(\theta_k(P), \theta(P)) \to 0$, that is, $\theta_k(P) \to \theta(P)$, or more generally $\varrho(\theta_k(P), \theta^{(n,p)}(P)) \to 0$ as $k, n, p \to \infty$, for each $P \in \mathcal{P}$.

  (ii) $\theta_k(P_n) \xrightarrow{P} \theta_k(P)$ for all k.

(2) *Selection of approximations.* We select a data determined value $\hat{k}_n(X_1, \cdots, X_n)$ and use as estimate, $\theta_{\hat{k}_n}(P_n)$.

That is, we approximate $\theta(P)$ by a "nice", call it regular, parameter θ_k which can be estimated by plug-in and then determine how fine an approximation we will use. Of course, k need not be an integer, but could be a continuous parameter such as the bandwidth. It is often useful to decompose the difference

$$\theta_k(P_n) - \theta(P) = [\theta_k(P_n) - \theta_k(P)] + [\theta_k(P) - \theta(P)] \tag{2.2}$$

The first term is naturally identified with variance, the second with bias, and the choice of k is the choice of best balance between the two. In this review, we necessarily mention only a small subset of the many ways the approximations have been chosen, but do stress the importance of the choice of k in many instances.

3 Nonparametric regression and classification (supervised learning)

3.1 Regression

Sequence Approximation. We return to model (1.3), which could equally well be written that we observe (Z, Y) with a completely unknown joint distribution (subject possibly to moment and smoothness conditions). Our goal is estimation in the $L_2(P)$ sense of the function valued parameter $\theta(P) = m(\cdot) = E(Y|Z = \cdot)$. This goal makes sense if we wish, knowing P, to predict a new Y given a new Z. If we use the predictor $\delta(Z)$, our loss is

$$\ell(P, \delta(Z)) = \int (y - \delta(z))^2 \mathrm{d}P(z, y) \tag{3.1}$$

The best choice of $\delta(Z)$ if, of course, $m(Z)$. Since we don't know P, we must use our "training sample" $(X_1, \cdots, X_n)$ to construct $\hat{\delta}(Z; X_1, \cdots, X_n)$. Since $m(Z)$ cannot be estimated by plug-in if Z is continuous, we need to apply regularization.

The first step is to select a sequence of approximation $\theta_k(P)$ which are meaningful if $P = P_n$. As we mentioned, there are many ways of selecting the sequence $\{\theta_k(P) = m_k(\cdot)\}$, penalization as in (1.4), see, for instance,

Zhang et al. (2004), or in a more structured way, sometimes referred to as the *method of sieves*, which we now explain.

We consider the models $\mathcal{P}_k = \{P : m(Z) = \sum_{j=1}^{k} \beta_j g_j(Z) \text{ for some } \beta\}$, and define an estimate appropriate to the parametric model $\mathcal{P}_k$. Least squares is the natural choice here: compute $\hat{\beta}_k$, the least squares estimate and $\hat{m}_k(z) = \hat{\beta}_k^T g(Z)$, where $g(Z) = (g_1(Z), \cdots, g_k(Z))^T$. The corresponding population $m_k(\cdot)$ is just $\sum_{j=1}^{k} \beta_j g_j(z)$, where $\beta = (\beta_1, \cdots, \beta_k)^T = \text{argmin}_\beta\{\int (y - \sum_{j=1}^{k} \beta_j g_j(z))^2 \mathrm{d}P(z, y)\}$.

Choice of regularization parameter in regression. We want to select $\hat{k} = k(P_n)$, which is "optimal" in terms of our loss function,

$$R(P, \delta) = E_P(Y - \delta(Z; X_1, \cdots, X_n))^2 \tag{3.2}$$

the expected squared error integrated out with respect to Z and $(X_1, \cdots, X_n)$. And so our first goal is consistency, $R(P, \hat{m}_{\hat{k}}(\cdot)) \to R(P, m(\cdot))$. It is easy to see that, by orthogonality, this is equivalent to $\int (\hat{m}_{\hat{k}}(z) - m(z))^2 \mathrm{d}P(z) \xrightarrow{P} 0$. This is equivalent to choose ρ to be $L_2(P)$ distance in the range of $\theta(P)(\cdot)$, which we identify as all square integrable functions of Z. Consistency corresponds to what we have called Goal (I).

As a concrete example, suppose that we believe that $\mathcal{P}_k$ is correct for some k, and our goal is to find the correct model or smallest correct model if the P_k are nested, as in our case, and then estimate β. The type (I) goal formulation leads, after construction of an unbiased estimator of the $MSE_k = E(\hat{m}_k(Z) - m(Z))^2$, where $m(\cdot)$ is the true population parameter, to a solution due to Akaike (1970), Mallows (1973) and others, "choose $\hat{k}$ to minimize $\sum_{i=1}^{n}(Y_i - \hat{m}_k(Z_i))^2 + 2k$". This choice comes from the representation

$$E(Y_i^0 - \hat{m}_k(Z_i))^2 = E(Y_i - \hat{m}_k(Z_i))^2 + 2\text{Cov}(\hat{m}_k(Z_i), Y_i) \tag{3.3}$$

where $Y_i^0 = m(Z_i) + \varepsilon_i^0$ is a new independent observation, and $2\sum_{i=1}^{n} \text{Cov}(\hat{m}_k(Z_i), Y_i) = k$ under the normality assumption on ε, see Efron (2004) for more details. On the other hand, pursuit of the type (II) goal puts great importance on identifying $k_0(P) = \min\{k : P \in \mathcal{P}_k\}$, the smallest model containing P first and then estimating β for purposes of interpretation. A Bayesian argument (Schwarz, 1978) to choose k by maximizing the posterior probability of $\mathcal{P}_k$ leads to the penalty $k \log n$ which evidently leads to much lower values of $\hat{k}$. The Akaike/Mallows criterion

does choose a model which is "correct" but not of smallest size. Readers are referred to Shao (1997) for more discussion on this issue. When p is allowed to increase with n, Bunea et al. (2006) show that consistent variable selection can also be achieved via multiple testing. Much more general choices of k involving types of cross validation are given later in this section.

Boosting and stagewise regression. There is another approach which does not specify the sieve in advance. In this case, we identify θ_k with the kth step of an algorithm, move from $\mathcal{P}_k$ to $\mathcal{P}_{k+1}$ on each step. Regularization here still means stopping the algorithm, i.e., choosing k in a data determined way. In stagewise regression, we fit one variable at a time, choosing one variable at step $k+1$ according to an optimization criterion based on the residuals of stage k. We discuss this type of method further in the section on classification. That regularization is necessary, can be seen by noting that the classical boosting method, recognized by Breiman as the Gauss-Southwell algorithm in numerical analysis, converges to the full regression on p variables in the context of the linear model. So, overfitting is still the main problem, see Hastie et al. (2001) for an excellent discussion of this.

Optimality. The penalized methods as (1.4) have been studied extensively by Birgé and Massart (2001) and many others, see Györfi et al. (2002) for an extensive overview. The criteria used in their analyses are worst case ones. They try to construct sieves and penalties which may be data determined so that, as we noted above,

a) $\theta_{\hat{k}}(P_n) \xrightarrow{P} \theta(P)$ as $n \to \infty$, P fixed, consistency in a more abstract formulation.

b) Further, for smoothness classes $\mathcal{P}$, the maximum regret, defined as the maximum difference between the risk of $\hat{m}^{(n)} = \hat{\theta}_{\hat{k}}(P_n)$ and that of $m \equiv \theta(P)$, the Bayes risk, converges to 0 at a rate which cannot be improved by any competitors for the given $\mathcal{P}$, that is

$$\sup_{\mathcal{P}}\{R(P,\hat{m}^{(n)}) - R(P,m)\} \asymp \inf_{\hat{\delta}}\sup_{\mathcal{P}}\{R(P,\hat{\delta}) - R(P,m)\} \quad (3.4)$$

where $\hat{\delta}$ depends on $X_1, \cdots, X_n$ only but not P. These rates are always of the form $n^{-2s/(2s+p)}\Omega(n)$, where $\Omega(n)$ is a slowly varying function, and p is the dimensionality of the data, s is a measure of the assumed smoothness of the members of $\mathcal{P}$.

Another approach to optimality which applies to both types of Goals (I) and (II) and is particularly favored by the machine learning community, following the work of Vapnik (1998), is to, from the beginning, restrict consideration to a fixed regularization class of possible procedures, as we do in our formulation, but then define $k(P) = k^*$ as the minimizer of $\rho(\theta_k(P_n), \theta(P))$, assuming P is known. This is the "oracle"'s choice. The goal then is to match the oracle for any P, i.e., choose $\hat{k}$ so that

$$\frac{\rho(\theta_{\hat{k}}(P_n), \theta(P))}{\rho(\theta_{k^*}(P_n), \theta(P))} \xrightarrow{P} 1 \tag{3.5}$$

To ensure uniformity over large classes, results are stated in terms of oracle inequalities of the form,

$$P[\rho(\theta_{\hat{k}}(P_n), \theta(P)) \leq C\rho(\theta_{k^*}(P_n), \theta(P)) + g(n, \gamma)] \geq 1 - f(P, n, \gamma) \tag{3.6}$$

for all n and P, where $C \geq 1$ is a constant, g goes to 0 as $\gamma \to 0$ and $f(P, n, \gamma) \to 0$ as $n \to \infty$ for fixed γ. Oracle inequalities can be used to prove possibly weaker results than (3.5), but suggest the construction of so called *adaptive procedures* (Donoho and Johnstone, 1998; Lugosi and Nobel, 1999) $\hat{m}_{\hat{k}}$ which get the correct rate over a whole scale of $\mathcal{P}$ of specified smoothness, $0 < s < \infty$.

3.2 Classification

The classification problem and boosting as an example. Boosting was first applied to the classification problem where $Y \in \{1, \cdots, N\}$. A classifier $\delta(Z) \in \{1, \cdots, N\}$ and the natural choice of loss is $\ell(P, \delta) = 1(\delta \neq Y)$. Arguing as before, if P is known, the δ minimizing $\int \ell(P, \delta(z))\mathrm{d}P(z)$ is $\theta(P) \equiv \delta_P(z) = j$ if $P[Y = j|Z = z] = \max\{P[Y = s|Z = z] : 1 \leq s \leq N\}$. Approximating $\theta(P)$ here can be done by estimating $m_j(Z) \equiv P[Y = j|Z]$ for $j = 1, \cdots, N$, where $m_N(Z) = 1 - \sum_{j=1}^{N-1} m_j(Z)$. If we treat each $m_j(Z)$ as a regression to be estimated, we are back in the regression formulation. This is what has implicitly been done in many current classification methods, with the exception of neural nets and perhaps support vector machines and the theoretically important methods of Mammen and Tsybakov (1999). We can think of the problem of classification into N categories as the same as the $\binom{N}{2}$ problems of classifying into pairs of categories i and j — though

this is not necessarily the best approach. Therefore, without loss of generality, we continue with the case $N = 2$. In this case, we relabel our 2 categories as -1 and 1 and we need only consider $m(Z) \equiv m_1(Z) = P[Y = 1|Z]$ since $P[Y = -1|Z] = 1 - m(Z)$. The Bayes rule is just

$$\delta(Z) = \text{sgn}(2m(Z) - 1) \tag{3.7}$$

For this situation, a large number of classes of procedures, such as neural nets, support vector machines, boosting have been studied, see Hastie et al. (2001) for an extensive coverage.

We will mainly discuss boosting, which, in this context, constructs estimates of a function $\hat{F}(Z)$, which estimates $q(2m(Z) - 1)$, where q is nondecreasing and $\text{sgn}(q(t)) = \text{sgn}(t)$. The classifier $\text{sgn}(\hat{F}(Z))$ is then an estimate of $\text{sgn}(2m(Z)-1)$, the Bayes rule. If q is strictly increasing, we obtain an estimate $q^{-1}(\hat{F}(Z))$ of $2m(Z)-1$. The type of estimates $\hat{F}(Z)$ proposed by boosting are of the additive type. Given a base space of classifiers (or more generally functions taking values in $[-1, 1]$), $\hat{F}(Z) = \sum_{j=1}^{k} c_j h_j(Z)$, where $h_j \in \mathcal{H}$, a predetermined base space of functions such that the linear span of $\mathcal{H}$ can approximate any $r(Z) \in L_2(P)$. Earlier approaches such as the sieves we have discussed were of this type also, but the structure of boosting is distinguished by constructing the sequence $\theta_k(P_n)$ as consecutive outputs of an algorithm with $\theta_k(P_n) \in \mathcal{P}_k$. Boosting iteratively builds an additive model as follows. For suitable convex functions $W(\cdot)$, $F_{k+1} = F_k + \hat{\gamma}\hat{h}$ is the argmin of

$$\frac{1}{n}\sum_{i=1}^{n} W\big(Y_i(F_k(Z_i) + \gamma h(Z_i))\big) \tag{3.8}$$

over $h \in \mathcal{H}$ and γ, where $\mathcal{H}$ is a large (or infinite) dictionary of functions of Z, (originally specified as "weak learners", classifiers themselves), for instance, candidate covariates or decision trees. In particular, $W(t) = e^{-\alpha t}$ leads us to AdaBoost. We can think of each iteration as representing a $\theta_k(P_n)$. Indeed, $F_k(Z)$, the kth population iterate converges to $F(Z) = q(2m(Z) - 1)$, as discussed. In particular for $W(t) = e^{-t}, q(v) = \log\big(\frac{1+v}{1-v}\big)$ and $F_\infty(Z) = \log\big(\frac{m(Z)}{1-m(Z)}\big)$. Thus if $\hat{F}_{\hat{k}}(Z)$ is an estimate of $F_\infty(Z)$, $\hat{m}_k \equiv exp\big(\hat{F}_{\hat{k}}(Z)\big)/\big(1 + exp\big(\hat{F}_{\hat{k}}(Z)\big)\big)$ estimates $m_k \equiv exp\big(F_k(Z)\big)/\big(1 + exp\big(F_k(Z)\big)\big)$.

Choice of regularization parameter in boosting. However, the $\theta_k(P_n)$ do not converge, since, for suitable $\mathcal{H}$, $\inf_{F\in\mathcal{H}} E_{P_n} W(YF(Z)) = 0$ and is

not achieved. The choice of k plays a critical role. Zhang and Yu (2005) suggested an *early stopping* rule to pick k in terms of the ℓ_1-norm of the boosting aggregation coefficients. Specifically, a sequence of suitably decaying tuning bounds $(b_k : k = 1, 2, \cdots)$ are chosen beforehand. Stopping occurs $(\hat{k} = k^*)$ as soon as the ℓ_1-norm of the coefficients (corresponding to the sparsest representation in the library $\mathcal{H}$), say, $\|\beta^{(k^*)}\|_1$, exceeds the predetermined bound b_{k^*}. Choosing $(b_k : k = 1, 2 \cdots)$ is evidently a problem and optimality in any sense is unclear. A simple way based on the "lasso", the Lagrange multiplier form of using the ℓ_1 penalty can be analyzed as follows. If γ is fixed , define $\theta_k(P)$ as the kth step, F_k in the population version of the algorithm for minimizing

$$\left\{ \int W(yF(z))\mathrm{d}P(z,y) + \gamma \sum_{j=1}^{k'} |\beta_j| : \right.$$
$$\left. F_k = F_{k-1} + \sum_{j=1}^{k'} \beta_j h_j \text{ sparsely represented for some } k' \right\} \quad (3.9)$$

and let $\theta_\gamma(P)$ be the minimizer which, in general, doesn't agree with $\theta_0(P)$, the true parameter. Then $\theta_k(P)$ do not converge to $\theta(P)$ defining the Bayes rule in general, but rather to $\theta_\gamma(P)$. But if $\gamma \to 0$, as we move from k to $k+1$, it is not hard to show that they do, provided that we have convergence if $\gamma = 0$. It is interesting to note that the phenomenon of failure to convergence of the algorithm described (or other algorithms for minimizing convex functions) for the sample and original objective function does not hold in the penalized case for any fixed $\gamma > 0$, since the objective function plus the convex penalty has a positive and achieved minimums. Thus, for the empirical version, the $\theta_k(P_n)$ do converge in probability to $\theta_\gamma(P)$ for γ fixed and, under suitable conditions on $\mathcal{H}$, to $\theta(P)$, if $\gamma_n \to 0$.

Various other ways of stopping based on versions of the classical model selection criteria, Bühlmann and Yu (2006) and Bühlmann (2006) have been recently proposed and their properties studied. Bickel et al. (2006) proposed yet another methods of early stopping which can achieve the appropriate rate bounds for Sobolev spaces. Their methods, save for the construction of a sieve of lower dimensional models to pass through, is the one primarily used in practice, V fold cross validation that we discuss later. Another approach to avoid early stopping is to regularize on each boosting step, as done in Lugosi and Vayatis (2004), in which minimization (3.8) is

constrained to the convex hull of $\mathcal{H}$. In order to select an optimal tuning parameter, their regularization scheme entails many ℓ_1-norm constrained optimizations, and is computationally problematic.

Optimality. Optimality for classification is more subtle than for squared error. If one uses as measure $0-1$ loss in the two classes case as above and $\delta(Z; X_1, \cdots, X_n) \in \{-1, 1\}$ is a rule, then, the Bayes regret is

$$R(P, \delta) - R(P, \delta_B) = E_P |r_P(Z_{n+1})| 1(\delta\delta_B < 0) \tag{3.10}$$

Here $r_P(Z) = 2P(Y = 1|Z) - 1$, and $\delta_B = \text{sgn}(r_P)$ is the Bayes rule, see Devroye et al. (1996) for instance. This expression reveals that the distribution of $r_P(Z)$ in a neighborhood of $\{z : r_P(z) = 0\}$ is as important as the estimation of $r_P(Z)$ by $\hat{F}(Z)$ if $\delta = \text{sgn}(\hat{F}(Z))$. Bayes regret can take very different values, in particular, it can be dramatically small if, for instance, $\{z : -\epsilon \leq r_P(z) \leq \epsilon\} = \emptyset$ for some ϵ. This is related by Tsybakov and others to the empirical margin between the sets $\{Z_i : Y_i = 1\}$ and $\{Z_i : Y_i = 0\}$. This quantity is defined through the hyperplane which, in the most balanced way, separates the sets committing at most ϵ_n errors. On the other hand, it is reflected in the population margin conditions of Tsybakov (2004). This empirical margin plays a major role in the oracle inequalities produced in the machine learning literature with minimax optimality counterparts in the work of Tsybakov (2004).

3.3 Selection of regularization parameter via cross validation

We have touched several ways to select γ (or k) in our previous discussion. We now address cross validation, as a most general model selection rule. An extensive review of model selection has been given by Wang (2004). Shao (1997) provided an interesting taxonomy of various model selection schemes in linear regression context.

Leave-one-out cross validation. A general approach is leave one out cross validation. Let $\mathbf{X}_{(-i)} = \{X_j : j \neq i\}$ and consider the predictor of Y_i, $\hat{m}_\gamma^{(-i)}(Z_i)$, trained from $\mathbf{X}_{(-i)}$ by penalizing with $\gamma Pen(\beta)$. Then the cross validation estimate of error is just

$$CV(\gamma) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_\gamma^{(-i)}(Z_i))^2 \tag{3.11}$$

The "optimal" $\hat{\gamma}$ is defined as giving the smallest cross validation error.

The motivation here is reasonably clear and goes back to the work of Stone (1974). $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{m}_\gamma^{(-i)}(Z_i))^2$ is an unbiased estimate of the actual risk of $\hat{m}_\gamma^{(-i)}(Z_i)$ which we expect is very close to that of $\hat{m}_\gamma(X_1,\cdots,X_n;Z_{n+1})=\hat{m}_\gamma(Z_{n+1})$ for which we want to compute $E(Y_{n+1}-\hat{m}_\gamma(Z_{n+1}))^2$.

For a linear estimator $\big(\hat{m}_\gamma(X_1),\cdots,\hat{m}_\gamma(X_n)\big)^T = H(\gamma)(Y_1,\cdots,Y_n)^T$, generalized cross validation minimizing

$$GCV(\gamma)=\frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i-\hat{m}_\gamma(Z_i))^2}{(1-tr(H(\gamma))/n)^2} \tag{3.12}$$

was proposed by Craven and Wahba (1979) for computational reasons, as an approximation to leave-one-out cross validation, since the computation of $\hat{m}_\gamma^{(-i)}$ $(i=1,\cdots,n)$ multiplies computation time by a factor of n.

Efron (2004) showed that all the methods we have discussed in this section so far correspond to the estimation of the expected *optimism*,

$$E(Y_i^0-\hat{m}_\gamma(Z_i))^2-E(Y_i-\hat{m}_\gamma(Z_i))^2 \tag{3.13}$$

in an approximately unbiased fashion. Using a Rao-Blackwell type argument, he further showed that the model-based penalty methods (C_p, AIC, SURE) outperformed the nonparametric methods such as leave 1 out CV, assuming the model is believable. He also gave similar connections between parametric and nonparametric bootstrapping methods.

The extent to which the use of CV and GCV yield procedures satisfying our optimality criteria has been studied (Li, 1985, 1986, 1987). Birgé and Massart (1997) showed that leave one out cross validation is equivalent to Mallows C_p in regression,making it optimal for nested models but selecting too large a model if all 2^p submodels are considered.

V-fold cross validation. In fact, few of these methods for selecting γ have been used in machine learning practice. The standard approach is to choose V dividing n, divide the sample into V disjoint parts of size $m=n/V$, say, $\Psi^{(1)},\cdots,\Psi^{(V)}$, and then use the $n-m$ observations in $V-1$ of the parts to calculate $\hat{m}_\gamma(\Psi^{(-t)})=\hat{m}_{\gamma,t}$ and evaluate

$$Q_t(\gamma)=\frac{1}{m}\sum_{j\in\Psi^{(t)}}(\hat{m}_{\gamma,t}(Z_j)-Y_j)^2 \tag{3.14}$$

an unbiased estimate of the risk of the prediction based on $n-m$ observations. Then, although looking at more than a single partition is not necessary for theory, form $Q(\gamma) = \frac{1}{V}\sum_{t=1}^{V} Q_t(\gamma)$, and choose $\hat{\gamma}$ by minimizing $Q(\gamma)$. Leave 1 out CV is also of this form with $V = n$. However, taking, say, $V = \frac{n}{\Omega(n)}$, where $\Omega(n)$ is slowly varying, can be shown to work very generally to establish both oracle and minimax results, see Györfi et al. (2002), Bickel et al. (2006). Some further discussion is in Dudoit and van der Laan (2005). A great advantage of both leave 1 out CV and V-fold CV is that they immediately generalize to any prediction question, such as generalized linear model prediction as in Fan and Li (2006), or more general model selection. V-fold cross validation is closely related to the m out of n bootstrap and subsampling we shall discuss in Section 5.

This discussion of the choice of γ in classification has been entirely in the context of Goal (I). When we turn to Goal (II), in which we assume there is a true model $\mathcal{P}_k$, the situation is different. If we choose γ via BIC, or in more complex situations, the closely related Bayesian, MDL criterion of Rissanen (1984), we can obtain the true k with probability tending to 1 and thus safely act as if $\hat{k}$ gave us the true model. On the other hand, as we have noted previously, AIC and the Goal (I) oriented criteria end up picking models that are larger than necessary.

3.4 Bayes and regularization

It is asserted, with some justification, that Bayesian methods regularize automatically. To see why this is so consider ridge regression with a large number of variables or the lasso in the same situation. If we assume as *a priori* that β_j are i.i.d $N(0, \frac{\sigma^2}{\gamma})$ and Y_i given Z_i are $N(Z_i^T\beta, \sigma^2)$, then the posterior density of β is proportional to

$$exp\{-\frac{1}{2}[\frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - Z_i^T\beta)^2 + \gamma\sum_{j=1}^{p}\beta_j^2]\} \tag{3.15}$$

Thus ridge regression can be thought of as finding the posterior mode of β and then plugging in to $\hat{m}_\gamma(Z)$. The lasso can be thought of similarly but with i.i.d double exponential β_j with density $f(\beta_j) = \frac{\gamma}{2}exp[-\gamma|\beta_j|]$. Of course, we are still left with the choice of γ. We can, in principle, put a fixed prior on γ also or alternatively use an empirical Bayes approach and estimate γ by maximum likelihood from $(Z_i, Y_i), i = 1, \cdots, n$, viewed

as having the marginal distribution obtained by integrating β out. The Gaussian prior and the empirical Bayes approach lead to the celebrated James-Stein Estimator (James and Stein, 1961).

Whether one thinks of the first stage regularization, putting prior distribution on β, as Bayesian or not seems immaterial. The second, however, is more problematic since the effect of integrating out β to get an estimate of γ requires caution. More significantly, making inference as in Goal (II) about β using the posterior leaves one asking questions about sensitivity to the choice of prior. The success of empirical Bayes methods used in the context of the Gaussian white noise model Johnstone and Silverman (2005) suggests that the frequentist behavior of Bayesian procedures in a prediction context, including using other posterior features such as the posterior mean rather than mode, should be studied further. This has become particularly attractive since MCMC (see Robert and Casella, 2004, for an introduction) makes the generation of approximate samples from the posterior, and hence of means rather than modes, computationally relatively easy.

General Bayesian model selection is mainly based on the Bayes factor (Kass and Raftery, 1995)

$$B(\gamma_1, \gamma_2) = \frac{P(M_{\gamma_2}|\mathbf{X})}{P(M_{\gamma_1}|\mathbf{X})} \div \frac{P(M_{\gamma_1})}{P(M_{\gamma_2})} = \frac{P(\mathbf{X}|M_{\gamma_1})}{P(\mathbf{X}|M_{\gamma_2})} \tag{3.16}$$

where M_{γ_1} and M_{γ_2} correspond to models with parameter γ_1 and γ_2 respectively. Kass and Wasserman (1995) showed that BIC can be refined by a more careful analysis of the asymptotics of the Bayes factor than that of Schwarz (1978).

3.5 Large n, large p

There has been relatively little work in this context for the model suggested by the introduction to our paper $Y = m(Z_1, \cdots, Z_p, \cdots) + \varepsilon$, where essentially we think of $(Z_1, \cdots, Z_p, \cdots)$ is as being infinitely dimensional with the variable $Z_1, \cdots, Z_p$ being all that is observed or more satisfactorily have $p \to \infty$, with n in our analysis. The major work in the context of Goal (I) has been the work of Greenshtein and Ritov (2004), Greenshtein (2006), Meinshausen (2005), and to some extent in Bickel and Levina (2004). In the context of Goal (II), Fan and coworkers (Fan and Li, 2006, and refer-

ences therein) have also looked at many generalizations of the regression model we have focussed on in the large n, p context.

3.6 Computational issues

It is important to note the computational savings of the Lasso and the usual forward stagewise algorithm. A major insight is in the work of Efron et al. (2004), in which it is shown that a modification of their fast Least Angle Regression (LAR) gives the complete path of the Lasso problem with varying penalty parameter. On the other hand, Hunter and Li (2005) proposed to use minorization-maximization (MM) algorithms for optimization involving nonconcave penalties and justified their convergence. Whether the latter algorithms will be computationally effective when there are many local minima remains to be seen.

3.7 Discussion

We have left out of our discussion many important methods such as local fitting of nonparametric methods (Fan and Gijbels, 1996) and tensor spline fitting (Stone et al., 1997), and, of course, neural nets, which involve nonlinear methods of estimation. We've also neglected other topics such as selecting γ, if interest focusses on other parameters which can be estimated at the $n^{-1/2}$ rate with curves, usually derivatives of regression function, and density functions viewed as nuisance parameters, estimated in some regularized way. For a discussion of difficulties which can arise if one is not careful, see Chen (1988) and the discussion in Bickel et al. (1998). Perhaps the outstanding issue in this area is the reconciliation of the theoretical optimality results with the exponential increase of methods proposed in practice and the production of a consistent overview. We have only scratched the surface.

4 Estimating large covariance matrices

Estimation of large covariance matrices, sometimes accompanied by the assumption that the data is $p-$variate Gaussian $N_p(\mu, \Sigma)$, plays an important role in various parts of statistics. The principal components (leading eigenvectors) of the empirical matrix have been used for data visualization

and reduction, by using only the principal components corresponding to the first few eigenvalues in order of absolute magnitude. In other directions, inverses of covariance matrices are important for determining important conditional relationships and for the construction of Kalman filters. The goal in all of these directions is of type (II), inference. But type (I) also appears, see Bickel and Levina (2004). The common feature of such analyses is, not surprisingly, that p and n are of the same order and frequently, as in microarrays, p is much larger than n, see, for instance, Dudoit et al. (2002), Kosorok and Ma (2006). As we mentioned earlier not only does the empirical covariance matrix become singular for $p > n$, but as pointed by Wigner (1955), Wachter (1978), Johnstone (2001), Johnstone and Lu (2006), Paul (2005), Bair et al. (2006), Bickel and Levina (2004) and others, if $\frac{p}{n} \to c, 0 < c \leq \infty$, the empirical eigenvectors and eigenvalues are grossly inconsistent in terms of estimating the corresponding population quantities.

If we think of X as an infinite sequence such that Σ, the variance-covariance matrix of the process $(X_1, X_2, \cdots)$ is a well conditioned operator on ℓ_2, see Böttcher and Silbermann (1999), and Σ_p is the variance-covariance matrix of the first p coordinates, then

$$\|\Sigma_p y - \Sigma y\| \to 0 \tag{4.1}$$

for all $y \in \ell_2$ as $p \to \infty$. Or equivalently if $y \in (y_1, y_2, \cdots), \sum_{j=1}^{\infty} y_j^2 = 1$, then $\mathrm{Var}(\Sigma_{j=p+1}^{\infty} X_j y_j) \to 0$. On the other hand, $\hat{\Sigma}_p y$ does not converge if $\frac{p}{n} \to \infty$. So we are led to regularization. Various methods have recently been proposed, Daniels and Pourahmadi (2002), Wu and Pourahmadi (2003), Huang et al. (2006), Ledoit and Wolf (2004), Furrer and Bengtsson (2006). Wu and Pourahmadi (2003) and Huang et al. (2006) use the remark of Pourahmadi (1999), Pourahmadi (2000) that fitting Σ by maximum likelihood fitting can be thought of as consecutively fitting inhomogeneous autoregressions of order $1, 2, \cdots, n-1$ to the data, and viewing the estimates of the autoregression parameters as estimates of the entries of the unique lower triangular matrix of the Cholesky decomposition of Σ^{-1}. If $p > n$, then Σ^{-1} is only defined in the Moore-Penrose senses. Both sets of authors assume $p < n$, and follow the Fan and Li (2001) prescription of penalizing the log likelihood viewed as fitting autoregressions. In one case, Wu and Pourahmadi (2003) do so, by selecting the maximum order of the autoregression fitted as $t < p$, using the Akaike model selection criteria, which can be viewed as a generalization of the Mallows criterion we dis-

cussed earlier. Huang et al. (2006) use the Lasso of Tibshirani (1996) as an L_1 penalty on the coefficients of the autoregression. Furrer and Bengtsson (2006) attack the problem differently using linear filters which preserve positive definiteness of the empirical covariance matrix. These filters have the effect of diminishing the absolute values of entries $\hat{\sigma}_{ij}$ of $\hat{\Sigma}$, according to their distance, from the diagonal.

All asymptotics, other than those of Furrer and Bengtsson (2006) and Johnstone and Lu (2006), were as $p, n \to \infty$, but $\frac{p}{n} \to 0$, and were essentially statements about the rate of convergence of individual regularized $\tilde{\sigma}_{ij} - \sigma_{ij}$ to 0, where $\Sigma = \|\sigma_{ij}\|$. Furrer and Bengtsson (2006) showed the much more useful convergence of the regularized matrices in the Frobenius norm $\sum_{i,j}(\tilde{\sigma}_{ij} - \sigma_{ij})^2$, but obtain results only if $\frac{p^2}{n} \to 0$. Johnstone and Lu (2006) devised a method for regularizing principal components for special types of Σ, where the number of large eigenvalues is bounded for all p, which gave convergence even if $\frac{p}{n} \to c > 0$.

In Bickel and Levina (2004), followed by Bickel and Levina (2006) (in preparation), one of the authors and E. Levina showed that by the crude method of regularization called banding, replacing $\hat{\Sigma}_p = \|\hat{\sigma}_{ij}\|$ by $B(\hat{\Sigma}) = \|\hat{\sigma}_{ij} 1(|i-j| \leq k)\|$, consistent estimation in the operator norm was possible as long as $\frac{\log p}{n} \to 0$. Note that the Frobenius norm is much larger than the operator norm if p is large. It implies convergence of eigenstructures since the operator norm does, but requires $\frac{p}{n} \to 0$. We view $\frac{\log p}{n} \to 0$ as remarkable since it covers situations such as microarrays where $n << p$. On the other hand, in microarrays, there is no one metric which corresponds to closeness to the diagonal. The methods we have developed so far however do permit application to situations, such as climate forecasting, with a similar imbalance between n, the ensemble size and p, positions of measurements in the atmosphere, where we can think of i, j corresponding to spatial points and it is reasonable to assume that covariances diminish in absolute value as the distance between points increases. We are in the process of deriving the analogue of methods based on Wu and Pourahmadi (2003)'s approach to fitting Σ^{-1} as well as to situations where we can apply the lasso, that is, Σ^{-1} is assumed to be sparse but the structure of the 0's has to be determined. Note that Σ_p can remain very well conditioned no matter what the relationship between p and n is. Essentially if one thinks of Σ as an operator from ℓ_2 to ℓ_2 as we have suggested, then we require that Σ is a bounded operator and invertible and that Σ^{-1} also be bounded. This is

satisfied by all stationary ergodic ARMA processes. We conclude by stating a result essentially from Bickel and Levina (2006) giving the flavor of our results.

Theorem 4.1. *Suppose* $\frac{\log p}{n} \to 0$. *Let* $\mathcal{T}_0$ *be a uniformly well conditioned set of covariance matrices. Then* $\exists k_n \uparrow \infty$ *such that* $\forall \epsilon > 0$,

$$\sup_{\Sigma \in \mathcal{T}_0} P\left[\|BAND_{k_n}(\hat{\Sigma}_p) - \Sigma_p\| \geq \epsilon\right] \to 0 \tag{4.2}$$

$$\sup_{\Sigma \in \mathcal{T}_0} P\left[\|[BAND_{k_n}(\hat{\Sigma}_p)]^{-1} - \Sigma_p^{-1}\| \geq \epsilon\right] \to 0 \tag{4.3}$$

The issue of choice of the regularization parameter k remains. Wu and Pourahmadi (2003), in a different context, use the Akaike criterion to select k in the method we have mentioned which is equivalent to approximating the covariance matrix by that of a kth order autoregression. Bickel and Levina (2006) investigate this approach further as well as the analogous approach of estimating the order of a moving average approximation,which banding the covariance matrix itself corresponds to.

It is interesting to note that if we are interested in a classification goal such as implementing the Fisher linear discriminant function, then an alternative approach which consider classifiers, based on linear predictors as we discussed, without reference to an underlying distribution such as the Gaussian, then results comparable to Bickel and Levina (2004) have been obtained by Greenshtein and Ritov (2004) and Greenshtein (2006).

We note also that there is an extensive literature on using Bayesian methods in estimation of Σ under parametric assumptions (Smith and Kohn, 2002). The sense that Bayesian methods regularize is present here also but the connection with $p, n \to \infty$ needs to be investigated under the very mild assumptions one can employ.

5 Subsampling and the m out of n bootstrap

Our main emphasis so far has been on the need for regularization in prediction, our Goal (I), although Goal (II) arises in this context as well. For instance,once we have a nonparametric estimate of a regression function, we would like to have a confidence band around it as well. Inferential problems of setting confidence bounds and testing become central as soon as we formulate semiparametric models whose parameters we interpret.

A central tool for making inferential statements in a non and semiparametric context is nonparametric maximum likelihood, specifically in the "bootstrap" form suggested by Efron (1979). Recall that this method essentially extends the scope of our previous discussion about plug in estimates to estimating a sample size dependent parameter such as the $1-\alpha$ quantile of the distribution of some complicated function of the data and P, such as the pivot $T_n(P_n, P) = \frac{\sqrt{n}(\bar{X}-\mu(P))}{\hat{\sigma}(P_n)}$, where $\mu(P) = E_P X, \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2(P_n) = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. If we call the quantile $\theta_n(P)$ then it is defined for all P and we have the "plug-in" bootstrap estimate $\theta_n(P_n)$, the $1-\alpha$ quantile of $T_n(P_n^*, P_n)$, where P_n^* is the distribution of a sample of size n from P_n, treating P_n as known. The success of the bootstrap is, we believe, due to the following features,

a) $\theta_n(P_n)$ can, in principle, be computed numerically $\theta_n(P) = \mathcal{L}_n^{-1}(1-\alpha, P)$ where

$$\mathcal{L}_n(t, P) = \frac{1}{n^n}\sum_{(i_1,\cdots,i_n)} 1(T_n(X_{(i_1)}, \cdots, X_{(i_n)}; P) \leq t) \tag{5.1}$$

But this, in practice impossible. However, as Efron pointed out, Monte Carlo simulation can yield (5.1) with arbitrarily good rate of precision. That is, we can approximate $\mathcal{L}_n$ by

$$\mathcal{L}_{nB}(t) = \frac{1}{B}\sum_{b=1}^{B} 1\big(T_n(X_{1b}^*, \cdots, X_{nB}^*; P_n) \leq t\big) \tag{5.2}$$

where $(X_{1b}^*, \cdots, X_{nb}^*), b = 1, \cdots, B$ is an i.i.d sample from P_n. This is the bootstrap as practiced.

b) The resulting estimates tend to be consistent and have nice higher order properties — Hall (1992). They share the general feature of maximum likelihood procedures that no choice of tuning constant is required.

Evidently, Efron's bootstrap can only be applied where it makes sense to talk of $\theta_n(P_n)$ so that the situations we have discussed previously do not arise. In fact it has made sense in situations where $\theta_n(P) \to \theta(P)$ with $\theta_n(P)$ defined for all P but $\theta(P)$ was not defined for all P. A prime example (Efron, 1979) is the suitably normalized variance of the

sample median, which converges only if the density of P, f, exists and is positive. That is $\theta_n(P) \equiv n\mathrm{Var}_P(X_{(\frac{n}{2})}) \to \frac{1}{4f^2(\mu(P))} \equiv \theta(P)$, where $\mu(P)$ is the population median. Efron showed that, even in this case $\theta_n(P_n) \xrightarrow{P} \theta(P)$. Yet, suppose that P corresponds to a bounded random variable with upper bound $\nu(P) = F^{-1}(1)$ and $f(\nu(P)-) > 0$. Then, while $\mathcal{L}_n(n(\nu(P) - X_{(n)})) \Rightarrow Exponential(f(\nu(P)-))$, the bootstrap distribution of $n(X_{(n)} - X^*_{(n)})$ does not converge to any fixed distribution.

A solution advocated early on, in cases such as this one, was to use the bootstrap distribution of $m(X_{(n)} - X^*_{(m,m)})$ as our estimate, where $X^*_{(m,m)}$ is the maximum of a sample of size $m < n$, and $m \to \infty$, but $m/n \to 0$. The rationale is that the joint distribution of $(X^*_1, \cdots, X^*_m)$ from P_n, $\mathcal{L}_m(P_n)$ is a much more stable estimate of $\mathcal{L}_m(P)$ which is close to $\mathcal{L}(P)$ and in turn to $\mathcal{L}_n(P)$. A better approximation may be to use $\tilde{\mathcal{L}}_m(P_n)$, where $\tilde{\mathcal{L}}_m$ is the distribution of the function of interest when P is replaced by P_n and the sample of size m is drawn without replacement. One reason is that the empirical distribution of a sample of m observations without replacement exhibits no ties unless there are points of mass in the support of P, while a bootstrap sample does with high probability, and, in that way, can be a poor approximation to an underlying P which is continuous.

Thinking carefully about this situation we see that we are again dealing with regularization. We assume that $\theta_n(P) \to \theta(P)$ on $\mathcal{P}$ in a suitable sense. We know that $\theta_m(P_n) \xrightarrow{P} \theta_m(P)$ for all fixed m. Thus we are essentially proposing that $\theta_{\hat{m}}(P_n)$ be used where $\hat{m} \to \infty, \hat{m}/n \xrightarrow{P} 0$. The key choice here is that of $\hat{m}$ since $\theta_m(P)$ are given by the problem. The generality of this approach is brought out by a remarkable theorem discovered independently by Politis and Romano (1994) and Götze (1993).

Theorem 5.1. *Suppose $T_n(P_n, P) = T_n(P_n)$ only, an ordinary statistic, and suppose that if $\mathcal{L}_n(P)$ is the distribution of $T_n(P_n, P)$, then*

$$\mathcal{L}_n(P) \to \mathcal{L}(P) \text{ on } \mathcal{P} \tag{5.3}$$

(Convergence here is in the weak of some other suitable sense.) Define $\mathcal{L}_n(P)$ itself as $\theta_n(P)$. Let $\tilde{\mathcal{L}}_m(P_n)$ be the distribution of $T_m(\tilde{X}_1, \cdots, \tilde{X}_m; P_n)$, the distribution of T_m where $(\tilde{X}_1, \cdots, \tilde{X}_m)$ are a sample without replacement from $X_1, \cdots, X_n$. Then if $m \to \infty, m/n \to 0$,

$$\tilde{\mathcal{L}}_m(P_n) \to \mathcal{L}(P) \tag{5.4}$$

without any further conditions.

In fact, under very weak conditions, the same is true of $\mathcal{L}_m(P_n)$. The subsampling approach (without replacement) is pursued extensively by Politis, Romano and workers in Politis et al. (1999). In particular, there are important and extensive generalization to simulation for statistics of stationary processes, following up the block bootstrap of Künsch (1989). We give some results on the choice of m in regularization for the m out of n bootstrap, rather than subsampling, because it permits us to think of choosing $\hat{m}$ to give consistency in an optimal way even when the ordinary bootstrap is consistent, which subsampling cannot.

Götze and Račkauskas (2001) and Bickel and Sakov (2005) analyze a general regularization method suggested in Bickel et al. (1997), which can be shown to give the best rates of convergence of $\mathcal{L}_{\hat{m}}(P_n)$ to $\mathcal{L}(P)$, whether the Efron bootstrap is or is not consistent. The methods rely on the following observations,

i) If $\mathcal{L}_n(T_n(P_n^*, P_n))$ doesn't converge to $\mathcal{L}(P)$, then it normally misbehaves seriously. It can be viewed as a probability distribution (as $X_1, \cdots, X_n$ vary) on the set of all probability distributions. As such it converges weakly, not to a point mass at $\mathcal{L}_P$, as it should when the Efron bootstrap is correct, but to a nondegenerate random probability distribution on the space of all probability distributions.

ii) If we put $m = n\pi^k$, for appropriately chosen $0 < \pi < 1$, $k = 1, 2, \cdots, r$, r fixed, $\mathcal{L}_m(T_m(P_m^*, P_n))$ misbehaves in exactly the same way as in (i), but convergence is generally to a different distribution for each $k > 0$.

iii) If m is fixed

$$\mathcal{L}_m^* = \mathcal{L}_m(T_m(P_m^*, P_n)) \to \mathcal{L}_m(P) \tag{5.5}$$

the limiting distribution of $T(P_m, P)$. Again, we expect $\mathcal{L}_{m_1}(P) \neq \mathcal{L}_{m_2}(P)$ if $m_1 \neq m_2$. The common exceptional cases are where $\mathcal{L}_m(P) \equiv \mathcal{L}(P)$ for all m in which case any fixed choice of m will give the same answers, so that any reasonable $\hat{m}$ will do well.

These remarks prompt our rule.

(1) Choose a metric ϱ on the space of probability distribution of T.

(2) Choose $\hat{m} = n\pi^{\hat{k}}$, where $\hat{k} = \text{argmin}_k \varrho(\mathcal{L}^*_{n\pi^k}, \mathcal{L}^*_{n\pi^{k+1}})$.

Interestingly enough, it is shown in Götze and Račkauskas (2001) and Bickel and Sakov (2005) that under suitable assumptions, $\varrho(\mathcal{L}^*_{\hat{m}}, \mathcal{L}_m(P))$ converges to 0, at the same rate as that given by $m_n = n\pi^{k_n}, k_n = \text{argmin}_k\{\varrho(\mathcal{L}^*_m, \mathcal{L}(P) : m = n\pi^k, k = 0, 1, \cdots\}$. This is evidently the best that an oracle could do. A major application is given in Bickel and Sakov (2005), to setting confidence bounds on extreme percentiles, $F^{-1}(1 - \frac{1}{n})$, where F is the cdf of P. Simulations suggest that the method which is not very sensitive to the choice of π, works as well as others where more knowledge of the tails of F is assumed, e.g., Breiman et al. (1990). A substantial difficulty of the approach is that we need the exact scale of $T_n(P_n, P)$, for instance, that n is right for $(\nu(P) - X_{(n)})$, where $\nu(P)$ is the upper endpoint of the distribution of X_1, since we need to know how to rescale when we form $T_m(P^*_m, P_n)$. Incorrect rescaling will lead to our estimate converging to point mass at 0 or $\pm\infty$. There are, however, various ways to estimate the correct scale by interpolating, between different values of m. For more and further references, see Bickel and Sakov (2005).

An important question is to how to apply more traditional models of regularization. In those cases where $\theta_m(P) - \theta(P)$ could genuinely be viewed as bias this has been done (see Hall et al. (1995) and Datta and McCormick (1995)). Otherwise it is unclear how to proceed.

Acknowledgement

We are grateful for Alexander Tsybakov for his kind comments.

References

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22:203–217.

Bair, E., Hastie, T. J., Paul, D., and Tibshirani, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7(1):1–31. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995).

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1998). *Efficient and adaptive estimation for semiparametric models. Reprint of the 1993 original*. Springer-Verlag, New York.

Bickel, P. J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

Bickel, P. J. and Levina, E. (2006). Regularized estimation of large covariance matrices. Technical Report 716, Department of Statistics, University of California, Berkeley, CA.

Bickel, P. J., Ritov, Y., and Zakai, A. (2006). Some theory for generalized boosting algorithms. *Journal of Machine Learning Research*. To appear.

Bickel, P. J. and Sakov, A. (2005). On the choice of m in the m out of n bootstrap and its application to confidence bounds for extreme percentiles. Unpublished.

Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In D. Pollard, E. Torgessen, and G. Yang, eds., *A Festschrift for Lucien Le Cam: Research papers in Probability and Statistics*, pp. 55–87. Springer–Verlag, New York.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268.

Böttcher, A. and Silbermann, B. (1999). *Introduction to large truncated Toeplitz matrices*. Universitext. Springer-Verlag, New York.

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

Breiman, L., Stone, C. J., and Kooperberg, C. (1990). Robust confidence bounds for extreme upper quantiles. *Journal of Statistical Computation and Simulation*, 37(3-4):127–149.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583.

Bühlmann, P. and Yu, B. (2006). Sparse boosting. *Journal of Machine Learning Research*, 7:1001–1024.

Bunea, F., Wegkamp, M. H., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference*, 136(12):4349–4364.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *The Annals of Statistics*, 16(1):136–146.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403.

Daniels, M. J. and Pourahmadi, M. (2002). Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*, 89(3):553–566.

Datta, S. and McCormick, W. P. (1995). Bootstrap inference for a first-order autoregression with positive innovations. *Journal of the American Statistical Association*, 90(432):1289–1300.

Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*, Vol. 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York.

Donoho, D. L. (2000). High dimensional data analysis: the curses and blessings of dimensionality. In *Math Challenges of 21st Centuary (2000)*. American Mathematical Society. Plenary speaker. Available in: `http://www-stat.stanford.edu/ donoho/Lectures/AMS2000/`.

Donoho, D. L. and Johnstone, I. M. (1998). Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921.

Draper, N. R. and Smith, H. (1998). *Applied regression analysis*. Wiley Series in Probability and Statistics: Texts and References Section. John Wiley & Sons, New York, 3rd ed.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.

Dudoit, S. and van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussions). *Journal of the American Statistical Association*, 99(467):619–642.

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

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*, Vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, eds., *Proceedings of the International Congress of Mathematicians, Madrid 2006*, Vol. III, pp. 595–622. European Mathematical Society Publishing House.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Furrer, R. and Bengtsson, T. (2006). Estimation of high-dimensional prior and posteriori covariance matrices in Kalman filter variants. *Jounal of Multivariate Analysis*. To appear.

Götze, F. (1993). Asymptotic approximation and the bootstrap. *I.M.S. Bulletin*, p. 305.

Götze, F. and Račkauskas, A. (2001). Adaptive choice of bootstrap sample sizes. In *State of the art in probability and statistics (Leiden, 1999)*, Vol. 36 of *IMS Lecture Notes Monograph Series*, pp. 286–309. Institute of Mathematical Statistics, Beachwood, OH.

Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under ℓ_1-constraint. *The Annals of Statistics*, 34(5). To appear.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York.

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hastie, T. J., Tibshirani, R., and Friedman, J. H. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York. Data mining, inference, and prediction.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, 93(1):85–98.

Hunter, D. R. and Li, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics*, 33(4):1617–1642.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the 4th Berkeley Sympos. Math. Statist. and Probability*, Vol. I, pp. 361–379. Univ. California Press, Berkeley, Calif.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327.

Johnstone, I. M. and Lu, A. Y. (2006). Sparse principle component analysis. *Journal of the American Statistical Association*. To appear.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics*, 33(4):1700–1752.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Kosorok, M. and Ma, S. (2006). Marginal asymptotics for the "large p, small n" paradigm: with applications to microarray data. Unpublished.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411.

Li, K.-C. (1985). From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, 13(4):1352–1377.

Li, K.-C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 15(3):958–975.

Lugosi, G. and Nobel, A. B. (1999). Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864.

Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of regularized boosting methods. *The Annals of Statistics*, 32(1):30–55.

Mallows, C. L. (1973). Some comments on c_p. *Technometrics*, 15(4):661–675.

Mammen, E. (1992). *When Does Bootstrap Work?*. Springer–Verlag, New York.

Mammen, E. and Tsybakov, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829.

Meinshausen, N. (2005). Lasso with relaxation. Unpublished.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 10:186–190.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33:1065–1076.

Paul, D. (2005). Asymptotics of the leading sample eigenvalues for a spiked covariance model. Unpublished.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):2031–2050.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics. Springer–Verlag, New York.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, 87(2):425–435.

Rissanen, J. (1984). Universal coding, information, prediction, and estimation. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 30(4):629–636.

Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer–Verlag, New York, 2nd ed.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27:832–837.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussions). *Statistica Sinica*, 7(2):221–264.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97(460):1141–1153.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussions). *The Annals of Statistics*, 25(4):1371–1470.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussions). *Journal of the Royal Statistical Society. Series B*, 36:111–147.

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

Tikhonov, A. N. (1943). On the stability of inverse problems. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 39:176–179.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166.

Vapnik, V. N. (1998). *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, New York. A Wiley-Interscience Publication.

Wachter, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *The Annals of Probability*, 6(1):1–18.

Wang, Y. (2004). Model selection. In *Handbook of computational statistics*, pp. 437–466. Springer–Verlag, Berlin.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā. Series A*, 26:359–372.

Wigner, E. P. (1955). Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics. Second Series*, 62:548–564.

Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R., and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association*, 99(467):659–672.

Zhang, T. and Yu, B. (2005). Boosting with early stopping: convergence and consistency. *The Annals of Statistics*, 33(4):1538–1579.

Zou, H. and Hastie, T. J. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, 67(2):301–320.

DISCUSSION

Alexandre B. Tsybakov
Laboratoire de Probabilités et Modèles Aléatoires
Université Paris VI, France

In their paper, Peter Bickel and Bo Li give an interesting unified view of regularization methods in statistics. The literature on this subject is immense, so they outline a general conceptual approach, and then focus on some selected problems where regularization is used, such as regression and classification, or more generally, prediction. In this context, they discuss in detail a number of recently emerging techniques, in particular, boosting, estimation of large covariance matrices, estimation in the models where the dimension is larger than the sample size.

It is difficult to overestimate the importance of regularization in statistics, especially in nonparametrics. Most of nonparametric estimation problems are ill-posed, and common estimators (kernel, histogram, spline, orthogonal series etc.) are nothing but regularized methods of solving them. The corresponding regularization parameters are just smoothing parameters of the estimators.

The main ideas of statistical regularization can be very transparently explained for prediction problems. Assume that $X_1, \dots, X_n$ are i.i.d. observations taking values in a space $\mathcal{X}$, and assume that the unknown underlying function $f^\star$ that we want to estimate belongs to a space $\mathcal{F}$. Consider a loss function $Q : \mathcal{X} \times \mathcal{F} \to \mathbb{R}$ and the associated prediction risk

$$R(f) = \mathbf{E}\, Q(X, f)$$

where X has the same distribution as X_i. Assume that $f^\star$ is a minimizer of the risk $R(f)$ over $\mathcal{F}$. Then a classical, but not always reasonable, estimator of $f^\star$ is a minimizer over $f \in \mathcal{F}$ of the corresponding empirical risk

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} Q(X_i, f).$$

Clearly, if $\mathcal{F}$ is too large, this can lead to overfitting and the minimizers can be nonsense. On the other extreme, if $\mathcal{F}$ is chosen to be too small, we cannot be sure that the true function $f^\star$ belongs to $\mathcal{F}$. So, continuing in the logic of empirical risk minimization, we need to know rather precisely a class $\mathcal{F}$ (the smaller, the better) where $f^\star$ lies. This, of course, is not very realistic in practice, but minimizing R_n over a suitable restricted class $\mathcal{F}$ yields us a first way of statistical regularization. For example, we can minimize R_n over the class of twice differentiable functions such that $\int (f'')^2 \le L$ where L is a given constant. Closely related is the second way of statistical regularization where a "roughness" penalty $\mathrm{pen}(f)$ is added to $R_n(f)$, for example, $\mathrm{pen}(f) = \lambda \int (f'')^2$ where $\lambda > 0$ is a smoothing parameter, and the estimator of $f^\star$ is defined as a minimizer of $R_n(f) + \mathrm{pen}(f)$.

These examples illustrate a construction of estimators for a given (fixed) smoothness of the underlying function. To get adaptation to unknown smoothness or other types of adaptation, we need one more stage of regularization, typically realized as penalization but this time over the complexity (smoothing) parameter appearing at the first stage. For instance, the famous Mallows – Akaike or cross-validation type schemes can be used.

Such a two-stage procedure works well in many cases. However, it has been recently realized that often it does not take advantage of the sparseness property. On the other hand, sparseness is believed to be an inalienable feature of many modern problems of signal processing and classification where "p is larger than n", in the terminology of Peter Bickel and Bo Li. A remedy can be then suggested in the form of alternatative regularization

procedures, with one stage only, which turn out to have optimal properties both in "classical" and "sparse" cases. One of the main ideas is to use an ℓ_1 penalization of the empirical risk or, on a closely related note, minimization of the empirical risk under an ℓ_1 constraint. In its earliest and simplest form, this idea appears in soft thresholding of Donoho and Johnstone for the gaussian sequence model. For other models, e.g., in regression and classification, it is realized in more recent procedures, such as Lasso, various versions of boosting or mirror averaging.

Let us focus here on boosting and mirror averaging. Consider a dictionary $\mathcal{H}$ of functions on $\mathcal{X}$. Assume without loss of generality that the dictionary is finite: $\mathcal{H} = \{h_1, \dots, h_M\}$, but M can be very large, for example, much larger than the sample size n. We believe that the underlying function $f^\star$ is well approximated either by the linear span $\mathcal{L}(\mathcal{H})$ of $\mathcal{H}$ or by its convex hull $\mathcal{C}(\mathcal{H})$. The aim is then to find an estimator $\tilde{f}_n$ such that its risk $R(\tilde{f}_n)$ would be close to the oracle risks $\inf_{f\in\mathcal{L}(\mathcal{H})} R(f)$ or $\inf_{f\in\mathcal{C}(\mathcal{H})} R(f)$. To get such an estimator $\tilde{f}_n$, we can implement ℓ_1 regularization, in particular, some versions of boosting. We can also use the method of mirror averaging.

Boosting. It will be convenient to distinguish between *linear boosting* where the output $\tilde{f}$ of the procedure belongs to the linear span of $\mathcal{H}$ (not restricted to its convex hull), and *convex boosting* where $\tilde{f}$ belongs to the convex hull $\mathcal{C}(\mathcal{H})$. Convex boosting methods can be viewed as ℓ_1 penalized procedures since the set $\mathcal{C}(\mathcal{H})$ is determined by an ℓ_1 constraint. Peter Bickel and Bo Li describe a basic linear boosting algorithm for the problem of classification (cf. (3.8)). Clearly, it can be also written for a general prediction problem:

- initialize: pick $F_0 \in \mathcal{L}(\mathcal{H})$,
- for $k = 0, 1, \dots, k^*$, find

$$(\hat{\gamma}_k, \hat{h}_k) = \operatorname{argmin}_{\gamma\in\mathbb{R}, h\in\mathcal{H}} R_n(F_k + \gamma h)$$

  and set $F_{k+1} = F_k + \hat{\gamma}_k \hat{h}_k$,
- output $\tilde{f}_n = F_{k^*+1}$.

Here the stopping time $k^* \le M - 1$ is a regularization parameter of the algorithm. It can be selected by classical methods, as mentioned above,

by adding a second stage of the procedure, i.e., a minimization of some criterion penalizing for large values of k. This is realized for the regression problem with squared loss by Bühlmann and Yu (2006), Bickel et al. (2006), Barron et al. (2005), and for classification with convex loss by Zhang and Yu (2005). Peter Bickel and Bo Li suggest in (3.9) another boosting method which is based on ℓ_1 penalization. They also provide its heuristic motivation. Some questions remain open here: how to choose k' in (3.9)? Does the method require a "second stage", i.e., a model selection step for early stopping?

For the regression problem with squared loss and for some linear boosting procedures $\tilde{f}_n$, Barron et al. (2005), under mild assumptions on the functions h_j from the dictionary, prove oracle inequalities of the form

$$\mathbf{E}\{R(\tilde{f}_n)\} \le C \inf_{f \in \mathcal{C}(\mathcal{H})} R(f) + \Delta_n \tag{1}$$

where $\Delta_n > 0$ tends to 0, but not faster than $n^{-1/2}$, and $C > 1$ is a constant. This shows that, in fact, their linear boosting procedures $\tilde{f}_n$ mimic the convex oracle.

Mannor et al. (2003), Lugosi and Vayatis (2004) and Klemelä (2006) establish oracle inequalities similar to (1) for some convex boosting procedures. However, there is no evidence that boosting mimics well linear oracles. For a particular linear boosting scheme, an inequality similar to (1) but involving linear oracle risk $\inf_{f \in \mathcal{L}(\mathcal{H})} R(f)$ has been obtained by Zhang and Yu (2005), however, with a remainder term Δ_n far from optimality. It would be, indeed, interesting to investigate whether boosting can achieve optimal rates of aggregation given in Tsybakov (2003). This can be, in principle, obtained as a consequence of *sparsity oracle inequalities*, i.e., inequalities of the form

$$\mathbf{E}\{R(\tilde{f}_n)\} \le C \inf_{f \in \mathcal{L}(\mathcal{H})} \left\{ R(f) + \frac{M(f)}{n} \log M \right\} \tag{2}$$

where $C \ge 1$ and $M(f)$ is the number of non-zero coefficients in the $\mathcal{H}$-representation of f:

$$M(f) = \min \left\{ \sum_{j=1}^{M} \mathbb{I}_{\{\lambda_j \ne 0\}} : f = \sum_{j=1}^{M} \lambda_j h_j \right\}$$

An open question is whether there exist a boosting procedure $\tilde{f}_n$ satisfying (2). Note that, in fact, (2) can be proved for other procedures: a first

example is given in Bunea et al. (2005, 2006) where (2) is established for a Lasso type $\tilde{f}_n$ in the regression model with squared loss.

Mirror averaging. A competitor of boosting is the mirror averaging (MA) algorithm Juditsky et al. (2005a,b). It aims to achieve the same goal as the boosting procedures discussed above which is to mimic the convex or linear oracles associated to a given dictionary of functions $\mathcal{H}$ (or to mimic the model selection oracle). The following two properties give evidence in favor of MA, as compared to boosting:

- unlike boosting, MA is an on-line method: it is applicable with streaming data. The computational cost of MA is of the same order or even smaller than that of boosting;
- in the theory, at least at its actual stage, better oracle inequalities are available for MA than for boosting.

To define the MA algorithm we introduce some notation. For any $\theta = (\theta^{(1)}, \ldots, \theta^{(M)}) \in \Theta \subseteq R^M$ set $\mathsf{f}_\theta = \sum_{j=1}^M \theta^{(j)} h_j$ and assume that Θ is convex and that $\theta \mapsto Q(X, \mathsf{f}_\theta)$ is convex for all $X \in \mathcal{X}$, with (sub)gradient $\nabla_\theta Q(X, \mathsf{f}_\theta)$. Given a sequence of positive numbers β_i, the basic MA algorithm is defined as follows:

- $i = 0$: initialize values $\zeta_0 \in \mathbb{R}^M$, $\bar{\theta}_0 \in \Theta$, $\tilde{\theta}_0 \in \Theta$,
- for $i = 1, \ldots, n$, iterate:

$$\begin{array}{rcll} \zeta_i & = & \zeta_{i-1} + \nabla_\theta Q(X_i, \mathsf{f}_{\bar{\theta}_{i-1}}) & (\text{GRADIENT DESCENT}) \\ \bar{\theta}_i & = & G(\zeta_i/\beta_i) & (\text{MIRRORING}) \\ \tilde{\theta}_i & = & \tilde{\theta}_{i-1} - (\tilde{\theta}_{i-1} - \bar{\theta}_{i-1})/i & (\text{AVERAGING}) \end{array}$$

- output $\tilde{\theta}_n$ and set $\tilde{f}_n = \mathsf{f}_{\tilde{\theta}_n}$.

Here $G : \mathbb{R}^M \to \Theta$ is a specially chosen "mirroring" mapping. When Θ is the simplex, $\Theta = \Lambda^M = \left\{\theta = (\theta^{(1)}, \ldots, \theta^{(M)}) : \ \theta^{(j)} \geq 0, \ \ \sum_{j=1}^M \theta^{(j)} = 1\right\}$, a possible choice of G is

$$G(z) = \left(\frac{\exp\left(-z^{(1)}\right)}{\sum_{j=1}^M \exp\left(-z^{(j)}\right)}, \ldots, \frac{\exp\left(-z^{(M)}\right)}{\sum_{j=1}^M \exp\left(-z^{(j)}\right)} \right), \tag{3}$$

where $z = (z^{(1)}, \dots, z^{(M)})$. Remark that choosing Θ as a simplex Λ^M can be viewed as an ℓ_1 regularization, this point is in common with the convex boosting procedures. Note also that the recursive averaging step of the MA algorithm is equivalent to

$$\tilde{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \bar{\theta}_{i-1}.$$

Therefore, when $\Theta = \Lambda^M$, the vector of weights $\tilde{\theta}_n$ belongs to the simplex Λ^M, so that $\tilde{f}_n$ is a convex mixture of the functions h_j with data-dependent weights.

Under the appropriate choice of β_i, the MA estimator $\tilde{f}_n$ with $\Theta = \Lambda^M$ and with $G(\cdot)$ as defined in (3) satisfies the following oracle inequality Juditsky et al. (2005a):

$$\mathbf{E}\{R(\tilde{f}_n)\} \le \inf_{f \in \mathcal{C}(\mathcal{H})} R(f) + 2\sqrt{Q^\star} \sqrt{\frac{\log M}{n}} \tag{4}$$

where

$$Q^\star = \sup_{\theta \in \Lambda^M} \mathbf{E} \|\nabla_\theta Q(Z, \mathsf{f}_\theta)\|_\infty^2.$$

Here and below $\|\cdot\|_p$ stands for the ℓ_p norm in $\mathbb{R}^M$. Inequality (4) shows that the MA algorithm mimics the convex oracle with optimal rate $\sqrt{\frac{\log M}{n}}$. It is sharper than the corresponding bound for boosting (1) because the risk of the oracle $\inf_{f \in \mathcal{C}(\mathcal{H})} R(f)$ in (4) appears with the minimal possible constant $C = 1$. Furthermore, (1) is only proved for the regression model with squared loss, while (4) is valid for any prediction model with convex loss.

In general, MA applies to any convex loss function whereas boosting is usually operational with the squared loss or with some special loss functions [an exception seems to be the gradient boosting of Mason et al. (2000) but not much is known about its theoretical properties].

There are also some computational advantages of MA as compared to boosting. The computational cost of boosting with finite dictionary of cardinality M is of the order nM^2: the cost of each iteration is of the order nM since we have to compare M different values of R_n, and this is multiplied by M since we need to run M iterations in order to select

the stopping time k^* by comparing their outputs. For some versions of boosting the cost is of the order nMk^* where the random stopping time $k^* \le M-1$ cannot be evaluated in advance. The computational cost of MA is just $O(nM)$, i.e., n iterations with vectors of dimension M. If M is very large, for example, $M \gg n$, the difference between the two costs becomes substantial.

For a general convex set Θ, the mirror mapping G is defined as

$$G(z) = \operatorname{argmin}_{\theta\in\Theta}\Big\{(z,\theta) + V(\theta)\Big\} \tag{5}$$

where $(\cdot,\cdot)$ denotes the scalar product in $\mathbb{R}^M$ and $V : \Theta \to \mathbb{R}^M$ is a convex penalty which is strongly convex w.r.t. the ℓ_1 norm in $\mathbb{R}^M$. The last requirement makes it impossible to take V as the ℓ_1 norm of θ, but a sensible choice (Juditsky et al., 2005a) is to use a penalty based on a norm that are quite close to the ℓ_1 norm, for example,

$$V(\theta) = \frac{1}{2}\|\theta\|_p^2, \quad p = 1 + \frac{1}{\log M}\,. \tag{6}$$

With this penalty and $\Theta = \mathbb{R}^M$, the mirror mapping G in (5) has the form

$$G(z) = -\left(\sum_{j=1}^M |z^{(j)}|^{\frac{p}{p-1}}\right)^{1-\frac{2}{p}} \left(|z^{(1)}|^{\frac{1}{p-1}}\operatorname{sign} z^{(1)}, \ldots, |z^{(M)}|^{\frac{1}{p-1}}\operatorname{sign} z^{(M)}\right).$$

To compare, the function G with exponential weights defined in (3) is a solution of (5) with $\Theta = \Lambda^M$ and the entropic penalty $V(\theta) = \sum_{j=1}^M \theta^{(j)} \log \theta^{(j)}$. This penalty also satisfies the strong convexity property w.r.t. the ℓ_1 norm (see Juditsky et al., 2005a). We see that MA with exponential weights operates with two types of penalization: the first of them is an ℓ_1 penalization due to a restriction of θ to the simplex $\Theta = \Lambda^M$, and the second one comes with the entropic penalty $V(\theta)$.

It would be interesting to understand whether the sparsity oracle inequalities of the type (2) can be proved for the MA algorithm. Some additional conditions on the loss function Q, such as strong convexity, might be needed to make it possible.

Additional references

Barron, A., Cohen, A., Dahmen, W., and DeVore, R. (2005). Approximation and learning by greedy algorithms. Manuscript.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2005). Aggregation for gaussian regression. *The Annals of Statistics*. Tentatively accepted.

Bunea, F., Tsybakov, A. B., and Wegkamp, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. In H. U. Simon and G. Lugosi, eds., *Proceedings of 19th Annual Conference on Learning Theory (COLT 2006)*, Vol. 4005 of *Lecture Notes in Artificial Intelligence*, pp. 379–391. Springer-Verlag, Berlin-Heidelberg.

Juditsky, A., Nazin, A., Tsybakov, A., and Vayatis, N. (2005a). Recursive aggregation of estimators by mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384.

Juditsky, A., Rigollet, P., and Tsybakov, A. B. (2005b). Learning by mirror averaging. Preprint, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 – Paris 7. `https://hal.ccsd.cnrs.fr/ccsd-00014097`.

Klemelä, J. (2006). Density estimation with stagewise optimization of the empirical risk. Manuscript.

Mannor, S., Meir, R., and Zhang, T. (2003). Greedy algorithms for classification – consistency, convergence rates, and adaptivity. *Journal of Machine Learning Research*, 4:713–742.

Mason, L., Baxter, J., Bartlett, P. L., and Frean, M. (2000). Functional gradient techniques for combining hypotheses. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, eds., *Advances in Large Margin Classifiers*, pp. 221–247. MIT Press, Cambridge, MA.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In B. Schölkopf and M. Warmuth, eds., *Proceedings of 16th Annual Conference on Learning Theory (COLT 2003) and 7th Annual Workshop on Kernel Machines*, Vol. 2777 of *Lecture Notes in Artificial Intelligence*, pp. 303–313. Springer-Verlag, Berlin-Heidelberg.

Sara A. van de Geer
Seminar für Statistik
ETH Zürich, Switzerland

Regularization has become a major statistical tool since computers have made it possible to analyze large amounts of data in various ways. The authors of this wonderful review paper have put regularization in its general perspective, ranging from classical Tikhonov regularization, to analysis of high-dimensional data and to bootstrap procedures. A trend one may observe over the last years is to apply many different algorithms to the same data set. (In fact, most statistical software present you numerous outcomes and statistics you never even asked for.) Regularization is crucial in the subsequent analysis where the outcomes of the estimation or testing methods are combined. As is pointed out in the paper, one should not use up all data in the first step, and take into account beforehand what validation procedure is used in the second step (e.g. V-fold cross validation).

The authors present a very general description on what regularization actually is. It formulates the concept in an asymptotic sense, with in the first step a sequence of approximating parameters θ_k converging to the target parameter ϑ, and for each k a sequence of estimators $\hat{\theta}_k$ converging to θ_k. The second step is then choosing a data dependent value $\hat{k}$ for k. It is to be noted that many regularization techniques are "embedded" ones, i.e., the first and second step are not strictly separated.

The (squared) distance between θ_k and ϑ may be called approximation error (bias^2) and the (squared) distance between $\hat{\theta}_k$ and θ_k may be called estimation error (variance). When the approximation error and estimation error are approximately balanced for a data dependent choice $\hat{k}$, one often speaks of a so-called oracle inequality. The problem in practice is that both types of error cannot be observed, as they depend on the underlying distribution. An important aspect of regularization is the estimation of the variance (or a more general concept of variability) of a collection of estimators. Let us illustrate this for the special case of empirical risk minimization. We only present the rough idea. For a good overview, see Boucheron et al. (2005), and also, see Koltchinskii (2006) for general oracle results.

Let the sample $X_1, \ldots, X_n$ be i.i.d. copies of a random variable $X \in \mathcal{X}$ with unknown distribution P, and let $\gamma_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$ be a given loss function. The theoretical risk is $R(\theta) := P\gamma_\theta$, and the empirical risk is $R_n(\theta) := P_n\gamma_\theta$. Our target parameter is $\vartheta := \arg\min_{\theta\in\Theta} R(\theta)$.

Consider a collection of model classes $\{\Theta_k\}$ with $\Theta_k \subset \Theta$. The empirical risk minimizer over Θ_k is

$$\hat{\theta}_k := \arg\min_{\theta\in\Theta_k} R_n(\theta).$$

The excess risk is $\mathcal{E}(\theta) := R(\theta) - R(\vartheta)$. The best approximation of ϑ in the model Θ_k is

$$\theta_k := \arg\min_{\theta\in\Theta} R(\theta)$$

The approximation error is now $B_k^2 := \mathcal{E}(\theta_k)$, and the estimation error is $V_k := R(\hat{\theta}_k) - R(\theta_k)$. Let us define the oracle as

$$k^* := \arg\min_k \left\{ B_k^2 + \mathbf{E}V_k \right\}.$$

Our aim is now to find an estimator $\hat{\theta} = \hat{\theta}_{\hat{k}}$ which mimics the oracle, i.e. which satisfies an oracle inequality of the form

$$\mathcal{E}(\hat{\theta}) \leq \text{const.}\big(B_{k^*}^2 + \mathbf{E}V_{k^*}\big)$$

with large probability (or in expectation).

This can be done by complexity regularization, invoking a penalty on the empirical risk, equal to a good bound for the estimation error. Let us briefly examine why. Consider the empirical process $\nu_n(\theta) := R_n(\theta) - R(\theta)$, $\theta \in \Theta$, and define $\mathcal{V}_k := -\big[\nu_n(\hat{\theta}_k) - \nu_n(\theta_k)\big]$. It is easy to see that $\mathcal{V}_k \geq V_k$, i.e., $\mathcal{V}_k$ is a bound for the estimation error V_k. Consider the penalized empirical risk minimizer $\hat{\theta} = \hat{\theta}_{\hat{k}}$, with

$$\hat{k} := \arg\min_k \left\{ R_n(\hat{\theta}_k) + \hat{\pi}(k) \right\}.$$

Suppose that with probability $1 - \epsilon$, we have for some constants A, and a_n,

$$\mathcal{V}_k \leq \hat{\pi}(k) \leq A\mathbf{E}V_k \ \forall\ k,$$

as well as

$$|\nu_n(\theta_k) - \nu_n(\vartheta)|/\sigma(\gamma_{\theta_k} - \gamma_\vartheta) \leq a_n \ \forall\ k,$$

where $\sigma^2(\gamma) = \mathrm{var}\big(\gamma(X)\big)$. Suppose moreover that

$$\mathcal{E}(\theta) \geq G[\sigma(\gamma_\theta - \gamma_\vartheta)] \;\forall\; \theta \in \Theta, \tag{1}$$

where G is a strictly convex function function with conjugate H. Then it is not hard to show that with probability at least $1 - \epsilon$, for all $0 < \delta < 1$,

$$(1 - \delta)\mathcal{E}(\hat{\theta}) \leq (1 + \delta)B_{k^*}^2 + A\mathbf{E}V_{k^*} + 2\delta H(a_n/\delta).$$

Thus, good bounds for the estimation error can result in an oracle inequality. Recent work in this area makes use of (local) Rademacher complexities (see Koltchinskii, 2006, and its references). An alternative approach is using V-fold cross validation. In general, oracle behavior through penalization or cross validation requires knowledge of the margin behavior, i.e. the function G in (1). Such knowledge is not required when using for example aggregation (see Tsybakov, 2004).

The bootstrap fits in nicely in regularization theory, as method to estimate the variability of an estimator. Alternatively, the authors view the distribution of a normalized estimator as θ_k and the bootstrap distribution as $\hat{\theta}_k$. For the bootstrap to "work", the limit ϑ of θ_k is assumed to exist. However, to me there is now no clear reason to balance the bias (distance between θ_k and ϑ) and the $\sqrt{\text{variance}}$ (distance between $\hat{\theta}_k$ and θ_k) in this setup.

Additional references

Boucheron, S., Bousquet, O., and Lugosi, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM. Probability and Statistics*, 9:323–375 (electronic).

Koltchinskii, V. (2006). 2004 IMS Medallion Lecture: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6). To appear.

Bin Yu
Department of Statistics
University of California at Berkeley, USA

First of all we would like to thank the authors for an insightful and coherent synthesis of regularization methods in statistical inference as diverse as sieves, model selection, penalized regression and classificaiton, and m out of n bootstrap. With the advent of information technology age, we encounter high dimensional data no matter where we are. Regularization has emerged as the key to extract meaningful information at the face of high dimensionality.

It is extremely interesting that the authors start off with the Tikhonov paper in 1943 which gives essentially the method of penalized regression as we might call it today in statistics. Tikhonov was concerned with solving an integral equation in a numerically stable manner. His formulation was through a regularized Least Squares (LS) optimization. In this discussion, we would like to share some thoughts on this connection between regularization and numerical stability.

Most statistical procedures or estimators can be derived as the solution to an optimization problem. The objective funtion is data dependent, hence random. In the classical domain, this random objective function stabalizes, as the sample size increases or in asymptopia, to a deterministic function at the courtesy of some version of the Law of Large Numbers. The minimizer of this deterministic function is the true parameter under certain smoothness conditions, resulting in the consistency of the estimator. In this classical setting, in asymptopia, when the data is a duplicate of the previous string, the objetive function doesn't change much – because both versions are close to the deterministic function in the limit.

For high dimensional data that we encouter nowadays, we still look for such "stability", but the asymptopia is not as well defined or established. That is, once the data is replicated or disturbed, we would like to ask the solutions to the objective functions to stay more or less the same, to a certain degree. We believe that this requirement is the most essential for a statistical procedure to make sense because if a procedure can not endure such a pertubation, then there is nothing like a "law" useful for

anything because things will just keep changing like noise. It follows that a meaningful statistical procedure has to be "stable" – this is almost the same as numerical stability except that it is a fixed and small pertubation for numerical stability and in our case the pertubation could be a probability distribution.

As pointed by Breiman (1996), with a large number of parameters in the high dimensional data situations, procedures are often unstable. When a procedure is not stable, then "regularization" is needed. To be precise, given data $Z = (Z_1, ..., Z_n)$ with a distribution $P(Z)$, a statistical procedure $\hat{\theta}(Z)$ is a function of this vector Z. When another data vector Z^* comes sharing the same distribution, we get $\hat{\theta}(Z^*)$. We would want $\hat{\theta}(Z^*)$ to be close to $\hat{\theta}(Z)$ in a distributional sense and properly defined relative to the desirable precision of the specific problem. On the other hand, the numerical stability is defined such that $Z^* = Z + \epsilon$ where ϵ is a small fixed pertubation vector. More generally, the distribution of Z^* might not be identical to the original distribution. Instead, it could represent a situation for the statistical "law", that we are after, to hold as well. Hence in statistics we "perturb" by Z^* in the form varying from bootstrap samples (cf. m out of n bootstrap in Bickel and Li), to permutation samples, and to cross-validation samples. Results have been obtained in statistical machine learning (empirical risk minimization) that a properly defined "stable" algorithm is proven to have good generalization performance, tying "stability" with statistical performance at a very concrete level (cf. Bousquet and Elisseeff, 2002; Kutin and Niyogi, 2002, and references therein). Even though we all agree by now that regularization is necessary for high dimensional problems, these results from machine learning are the beginning of directly justifying the use of stability (hence regularization) for statistical gains. They are dervied using McDiarmid's concentration inequality and its variants: the condition is the existence of a bound on the empirical risk when one component of data is pertubed by an independent copy – a stability condition. Then, the empirical risk concentrates on its expectation and good generalization error bounds follow.

The goal in achieving numerical stability is to turn an ill-posed problem into a well-posed one. One prominent example is the Tikhonov regularization which started the Bickel and Li paper under discussion. For statistical regularization, the well-posed problem needs to be related back to the original ill-posed problem in the sense that 1. the solutions of the well-posed problems with original Z and disturbed Z^* are close; 2. when

the regularization parameter goes to zero, the solution to the well-posed approximation gets close to the "optimal" solution in the population case. In essence, these two considerations generalized from numerical stability are reflected in the definition of a "regularization process" in the paper under discussion.

We have so far explored mainly the conceptual connection between statistical regularization and numerical stability, the latter being a numerical optimization concept. Not only that the Tikhonov regularization corresponds to penalized regression in modern statistics, the implict regularization by early stopping as in Boosting has also a counterpart in numerical regularization which is known as Landweber Iteration. It would be interesting to investigate further the connections of statistical regularization and numerical optimization at a computational or algorithmic level.

Additional references

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2(3):499–526.

Kutin, S. and Niyogi, P. (2002). Almost-everywhere algorithmic stability and genearalization error. Technical Report TR-2002-03, Department of Computer Science, University of Chicago, Chicago, IL.

Teófilo Valdés and Carlos Rivero
Department of Statistics and Operational Research
Complutense University of Madrid, Spain

A common question that researchers of any subject have to face in many cases is: What should be done when an irregular situation arises and the standard procedures do not work? The answer is simple: use common sense (that is, think up a plausible rationale) to make them work. In short, use regularization to be able to handle irregular (atypical, or unstable, or ill-posed) situations. From this point of view, the need of regularization appears in any science and it is clearly based on pragmatism. In Statistics, for example, this happens when fitting parametric or non-parametric models with a great number of parameters (even more than the number of data), or

when estimating large covariance matrices, or when estimating densities, or when treating with missing or incomplete data, and so on. Since the basis underlying regularization is pragmatism, its techniques depend mainly on the particular situations that need to be tackled. That is, the regularization techniques, like the heuristic procedures, comprise mainly of *ad hoc* procedures which need to be empirically assessed or validated (in most of the cases, by simulation). Although this is true to a great extent, Bickel and Li have made a worthy effort to analyse the conceptual insides of regularization in statistics, the intention being to present a solid and comprehensible unified theory (methodology is a more precise term on seeing the scope of the paper) of it. This methodology is frame-worked by the following two initial conditions: (1) the data available is a random sample, and (2) the attention is centred on asymptotics when the sample size tends to ∞. Under these circumstances, their abstraction task has been successful in discovering and presenting the common aspects of regularization which are corroborated by a wide range of examples. We congratulate the authors for their efforts in unifying different regularization processes, concerning both data and models, which were developed to treat a great variety of unstable situations up to date considered unlinked. The interest of unifying different theories and techniques under a common conceptual approach has been constant throughout centuries. Several examples and tries have occurred, with more or less success, in scientific areas such as Physics, Mathematics, Economics, Medicine, Psychology..., and, also, in Statistics. In fact, the search of unified theories has been the motor of basic research. Under this perspective, the paper of Bickel and Li undoubtedly deserves a special praise, which must be added to that merited by the clear presentation and good organization of the paper, the numerous examples discussed and the large number of references included.

All praises having been said (and meant), we will switch to the role of critic commenting on some remarks and suggestions that came to our minds after reading the paper.

1. As was indicated above, the authors have made a valuable effort in presenting a unified methodology to treat irregular cases in statistics. With this methodology, one may be conscious of the sequence of steps that may lead us to solve an irregular situation. However, since the paper includes no global results from which different particular cases

may be tackled, we think that some way additional effort needs to be done before a unified theory of regularization is present. This task is, in our opinion, challenging and we encourage the authors to continue researching into this area sketched in the paper.

2. The authors maintain that a generic regularization process consists of two different activities. The first (the second will be considered later) is the *sequence of approximations* in which the objective is to construct the sequence θ_k, which needs to be defined on the set of all possible underlying distributions as well as on the set of discrete distributions to guarantee that $\theta_k(P)$ can be estimated by the natural "plug in" estimate $\theta_k(P_n)$. The authors impose consistency at the two levels:

$$\theta_k(P) - \theta(P) \xrightarrow{P} 0, \text{ and } \theta_k(P_n) - \theta_k(P) \xrightarrow{P} 0.$$

However, in many typical situations, mainly when Θ is a Euclidean space, only convergence in law (to a certain known distribution) is needed for

$$\theta_k(P_n) - \theta(P),$$

thus, the convergence in probability may be weakened in one of them (usually in the second). It is clear that the authors are mainly interested in non-parametric statistical methods, in which function valued parameters are present and convergence in law may be pointless. Since regularization appears in both non-parametric statistics and parametric statistics, to contemplate other possible types of convergences may help to widen the conceptual scope of the paper.

3. The second activity is what is called in the paper the *selection of approximations* in which the "plug in" estimate of the regular parameter θ_k is approximated from the data. In an ideal situation, it would be desirable that

$$\theta_{\widehat{k}_n(X_1,\ldots,X_n)}(P_n) - \theta_k(P_n) \xrightarrow{P} 0$$

and a longer decomposition, similar to (2.2), should be established and interpreted. The authors do not mention this (although something similar may be intuited from (3.5)), probably thinking that this activity is highly dependent on the particular case under study. Finally, as a minor remark, the names given to the activities are a little

confusing, since the word approximations appears in both without qualification. Likely, "sequence of estimators" and "selection of approximations" would be better terms and more descriptive, although we do not wish to argue on semantics.

4. Section 3 provides an authoritative review of non-parametric regression and classification. It constitutes an example of clear and broad exposition, and profound analysis of the topics mentioned above. It is also an excellent pedagogical work, since all is articulated under the perspective of the regularization methodology described at the end of Section 2. The sole point that we consider arguable is to consider the Bayesian methods as an automatic regularization. The fact that the ridge regression (Hoerl and Kennard, 1970) and the "lasso" (Tibshirani, 1996) result as a particular case of posterior mode is a weak argument, and the selection of the approximations tremendously controversial. In spite of this, we congratulate the authors for this pedagogical exposition which we extend to the rest of the examples with which the paper is brought to a close.

Finally, we will like to highlight that after reading the paper we have found out that certain problems of inference with missing or incomplete observations fall within the scope of regularization. This happens, for example, in the context of linear models or panel data models with general errors, not necessarily Gaussian, and interval censored data (see Rivero and Valdés, 2004, 2006). Although we do not use the "plug in" estimation of the regular parameter $\theta_k(P)$, it is clear that the concept of regularization may be extended to experiments in which the data available does not constitute a random sample.

We would like to thank the editors of TEST for having given us the opportunity to read and discuss the insightful paper of Bickel and Li, in which their unifying view of the asymptotics of regularization is revealed and magnificently displayed.

Additional references

Rivero, C. and Valdés, T. (2004). Mean based iterative procedures in linear models with general errors and grouped data. *Scandinavian Journal of Statistics*, 31(3):469–486.

Rivero, C. and Valdés, T. (2006). A procedure to analyse covariance panel data models with grouped observations. Submitted.

Jianqing Fan
Department of Operations Research and Financial Engineering
Princeton University, USA

I would like to wholeheartedly congratulate Bickel and Li for their comprehensive, stimulating, and successful overview of the regularization methods in statistics. Their attempt to integrate diversified statistical methods from a regularization point of view is intriguing, and their paper demonstrates convincingly and surprisingly how seemingly unrelated techniques, from nonparametric function estimation, model selection, and machine learning to Bayesian inference, covariance matrix estimation, and bootstrap, can indeed be thought as some aspects of regularization. I appreciate the opportunity to comment and expand the discussions by Bickel and Li.

1 Regularization and sparsity

As Bickel and Li insightfully suggested, the regularization method is to construct a more regularized sequence $\theta_k(P)$ to approximate $\theta(P)$. The approximation error

$$\theta_k(P) - \theta(P) \tag{1}$$

is usually the bias of the estimator $\theta_k(P_n)$, with the P_n being the empirical distribution. The ways to approximate $\theta(P)$ are far from unique. For example, a nonparametric regression function $m(x)$ in L^2 admits an expansion

$$m(x) = \sum_{i=1}^{\infty} \theta_i \phi_i(x), \tag{2}$$

where $\{\phi_i(\cdot)\}$ is a family of basis functions in L^2. In the situation of estimating the conditional probability in supervised learning, a known link g should be applied before the expansion. Commonly-used basis functions include Fourier, wavelets, and splines. Thinking of k in (1) as the number of terms chosen from expansion (2), we hope that a basis is chosen such

that approximation errors in (1) are as small as possible. To achieve this, the representation in (2) should be as sparse as possible — namely, most coefficients θ_i should be small.

With a sparse representation, the regularization can be effective. It substantially reduces the dimensionality by focusing only the non-sparse elements in the expansion. This mitigates the variance of the estimation. For a smooth function with a Fourier basis, it is expected that the energy at high frequencies is nearly zero and therefore the estimation focuses only on the first k coefficients (Efromovich, 1999). For functions with discontinuities or different degrees of smoothness, the Fourier expansion is not effective; instead, wavelet representation can achieve sparsity. With the sparse representation, the regularization basically becomes a model selection problem (Antoniadis and Fan, 2001). The hard- and soft-threshold procedures in Donoho and Johnstone (1994) are simple and effective model selection approaches when the design matrix is orthogonal. When a spline basis is used, model selection techniques are frequently employed to select non-sparse elements, as exemplified in the work by Stone and his collaborators (Stone et al., 1997).

The sparsest possible representation is the one in which the basis contains a function that is parallel to the unknown function m and the rest are orthogonal complements. In this ideal basis, the representation is the sparsest possible with only one nonzero coefficient. This means that there does not exist an orthogonal basis that can universally sparsely represent a family of functions. Over-complete bases have been sought to make the sparse representation possible over a larger family of functions (Chen et al., 1998).

2 Model selection

As Bickel and Li correctly pointed out, model selection is also a regularization. I agree with their two main objectives in model selection: risk minimization and causal inference. Model selection usually assumes that there is a finite-dimensional (possible dependence on the sample size) correct submodel, while nonparametric function estimation does not. Hence, the former imposes exact sparsity, with many coefficients exactly zero, while the latter requires approximate sparsity, with many small coefficients.

The developments of model selection and nonparametric function estimation have influenced each other over the last twenty years. The model selection community has helped the nonparametric one to develop procedures that select non-sparse elements, while the nonparametric community has helped parametricians to understand modelling biases and their consequences in parametric inferences.

In achieving the first goal of risk minimization, the optimal predictors are not necessarily the ones with non-zero coefficients. Setting some small coefficients to zero or shrinking them toward zeros reduces the variance and instability of the prediction. This comes at the cost of a possible increase in the biases. The situation is very similar to that of optimal smoothing in nonparametric function estimation.

In achieving the second goal of causal inference, a more concise relationship between covariate and response variables is needed. In this case, the usual idealization of the model is that some covariate variables contribute to the response variables while others do not, and the statistical task is to identify the correct submodel and to estimate their associated coefficients. Fan and Li (2001) outline three properties that a model selection procedure should ideally have.

Sparsity: Some coefficients are estimated as precisely zero, which reduces the model complexity.

Continuity: Estimated coefficients should be a continuous function of data to avoid instability in model prediction.

Unbiasness: For coefficients that are statistically large enough, the estimation procedure should not try to shrink these coefficients to avoid unnecessary biases.

In addition, a model selection procedure should allow valid statistical inferences: the stochastic errors in the model selection processes should be accounted for in constructing confidence intervals and in other statistical inferences. Fan and Li (2001) proposed a penalized likelihood using the SCAD penalty to achieve the postulated properties. It corresponds to Bayesian estimation with an improper prior to achieve the unbiasedness property and with irregular 'density' function at the origin to achieve sparsity.

3 High-dimensional semiparametric problems

The issue of sparsity arises naturally in microarray and proteomic applications. Among tens of thousands of genes, it is believed that there are at most hundreds of genes differently expressed between the treatment and control arrays. For example, in the normalization of a microarray (Fan et al., 2005a,b), the following model

$$y_{gi} = \mu_g + f_i(x_{gi}) + \varepsilon_{gi}, \quad g = 1, \cdots, G; i = 1, \cdots, n$$

is proposed, in which y_{gi} represents the observed log-ratio of the expressions of gene g between the treatment and control in the i^{th} array, x_{gi} is the associated log-intensity, μ_g is the treatment effect on gene g, and $f_i(\cdot)$ is the intensity effect on the i^{th} array. The normalization is to estimate the intensity effects $f_i(\cdot)$ and remove them from the log-ratios. Other parameters can be added to the model account for the block effect (Fan et al., 2005b). Hence, for the normalization purposes, the parameters $\{\mu_g\}$ are nuisance ones. In the microarray applications, it is helpful to think that the total number of genes G tends to ∞. Biological sparsity means that most of μ_g equals to zero. This information helps more accurately estimate $\{f_i(\cdot)\}$ and poses new methodological and theoretical challenges on how to efficiently estimate them.

In tumor classification using microarrays (Tibshirani et al., 2002), it is desirable to choose tens of genes to construct classification rules among tens of thousands of genes. Using the generalized view of Bickel and Li, this can also be regarded as a regularization problem. Efficient construction of classification rules and statistical understanding of these rules pose new challenges in statistics.

4 High-dimensional covariance matrix

Covariance matrices are very important in portfolio management and asset allocation. Suppose that we have 500 stocks to be managed. The covariance matrix involves 125,250 elements. Therefore, regularization is necessary. Let $Y_1, \cdots, Y_p$ be the excessive returns of p assets over the risk-free interest rate. Derived by Ross (1976) using the arbitrage price theory and Chamberlain and Rothschild (1983), these excessive returns can be written

approximately as

$$Y_i = b_{i1} f_1 + \cdots + b_{iK} f_K + \varepsilon_i, \quad i = 1, \cdots, p, \tag{3}$$

where $f_1, \cdots, f_K$ are the returns of the K factors that influence the returns of the assets, b_{ij}, $i = 1, \cdots, p$, $j = 1, \cdots, K$, are unknown factor loadings, and $\varepsilon_1, \cdots, \varepsilon_p$ are uncorrelated idiosyncratic errors. That is to say that given the K factors, the cross-sectional market risk is captured by these K factors. Model (3) is called a multi-factor model in financial econometrics. Assume that the factors $f_1, \cdots, f_K$ are observable such as those in the famous Fama-French three-factor or five-factor model (Fama and French, 1993). Then, there are $(K + 1)p$ instead of $p(p + 1)/2$ parameters. The number of factors K can also depend on the number of assets p. This can also be regarded as a regularization problem, according to Bickel and Li, as some factor loadings can be very small. Assume that we have the data observed on n periods (e.g. days); then the covariance matrix can be estimated using the factor structure (3). The question then arises if the factor structure helps us better estimate the covariance matrix under a relevant norm.

5 Inference using regularization

They are many statistical inference questions that require regularization. For example, after fitting the linear model

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \tag{4}$$

one may naturally ask if model (4) is adequate. For this kind of question, the alternative hypothesis is usually vague. Similar problems arise in many scientific investigations in which the null model is usually well formulated while the alternative model is not. For example, one may ask if a stochastic volatility model fits returns of certain assets, or if a biological model is consistent with observed data.

A natural alternative to model (4) is the following:

$$Y = m(X_1, \cdots, X_p) + \varepsilon, \tag{5}$$

where m is unspecified. While this family of models is wide enough to include the true regression function, consistent tests have little power due to

the curse of dimensionality. Another possibility is to impose the alternative model of additive structure:

$$Y = m_1(X_1) + \cdots + m_p(X_p) + \varepsilon. \tag{6}$$

In both cases, regularization is needed for constructing an omnibus test. The testing problem is essentially a parametric null hypothesis versus a nonparametric alternative hypothesis.

The problem of testing nonparametric hypothesis against a larger nonparametric hypothesis can also arise naturally. Under the additive model (6), one may ask if the first two variables are statistically significant. The null model becomes

$$Y = m_3(X_3) + \cdots + m_p(X_p) + \varepsilon.$$

Again, regularization is needed for this type of hypothesis. Fan et al. (2001) introduced the generalized likelihood ratio test for handling both types of testing problems. Detailed development for these problems in the additive model can be found in Fan and Jiang (2005).

Additional references

Antoniadis, A. and Fan, J. (2001). Regularized wavelet approximations (with discussion). *Journal of the American Statistical Association*, 96:939–967.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51:1281–1304.

Chen, S., Donoho, D. L., and Saunders, M. A. (1998). Automatic decomposition by basis pursuit. *SIAM Journal on Scientific Computting*, 20:33–61.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer–Verlag, New York.

Fama, E. and French, K. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56.

Fan, J., Chen, Y., Chan, H. M., Tam, P., and Ren, Y. (2005a). Removing intensity effects and identifying significant genes for Affymetrix arrays in MIF-suppressed neuroblastoma cells. *Proceedings of the National Academy of Sciences of the United states of America*, 103:17751–17756.

Fan, J. and Jiang, J. (2005). Nonparametric inference for additive models. *Journal of the American Statistical Association*, 100:890–907.

Fan, J., Peng, H., and Huang, T. (2005b). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. *Journal of the American Statistical Association*, 100:781–813.

Fan, J., Zhang, C. M., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, 29:153–193.

Ross, S. (1976). The arbitrage theory of capital asset pricing. *Journal of Economic Theory*, 13:341–360.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United states of America*, 99:6567–6572.

Aad van der Vaart
Department of Mathematics
Vrije Universiteit Amsterdam, Netherlands

Bickel and Li are to be congratuled with this review of the use of regularization methods in statistics. The treatment is insightful and broadminded. My remarks are focused on some aspects of regularization that I feel deserve more emphasis.

1 Approximation theory

Approximation theory has developed in a separate specialism within mathematics, applied and pure. The aim is to represent "general" functions by simpler ones up to an approximation error. Theoretical interest is in the maximal approximation error on a given class of functions achieved by a scheme of given complexity. "Constructive approximation" methods (e.g. DeVore and Lorentz, 1993) make such schemes practical. There is a clear relation with efficient coding of functions (e.g. Cohen et al., 2001; Kerkyacharian and Picard, 2004). Wavelets are the most recent successful example, but older examples as polynomials, Fourier series or splines also belong to this area.

Many examples of regularization in statistics are based on such approximation methods. A "true" parameter (regression function, density) is replaced by a simpler one, and next the simpler one is estimated from the data. In an unpolite manner one could say that statistics is only studying the effects of adding noise on the approximation error, although bias-variance thinking appears to yield new insight even in approximation itself.

Using an approximation scheme that is suited to the application at hand is very important. For instance, smoothness, periodicity, locality, sparsity, spatial distribution, etc. appear all non-statistical aspects. A modern view of penalization methods (expressed e.g. in Barron et al. (1999), Birgé and Massart (2001) or Birgé (2006)) is to set up a (very) large number of models with good approximation properties and next make a data-driven choice of these models. With the right penalties this will result in "adaptive" estimators that work well whenever the "truth" is close to one of the models. A remarkable finding is that (at least in theory) it is possible to use huge numbers of models (e.g. exponential in the number of replicated data) in these schemes. Still at least one of the approximating models must be good, where "good" will depend on the type of application.

It will certainly be profitable for statisticians to follows the many new schemes developed in approximation theory (cf. DeVore et al., 2006) and engineering. At least if they are interested in aim I mentioned by Bickel and Li: prediction. For aim II, causal inference, approximation theory appears to be of little help.

2 Bayesian methods

Bayesian methods for non- and semiparametric models have long been looked upon with suspicion, based on the finding that many (or most, depending on definition and taste) priors in these settings lead to posterior distributions that are a very poor reflection of the distribution underlying the data. For instance, the posterior distribution may not contract to the "truth" as the number of replicated observations increases indefinitely. Through the use MCMC-schemes Bayesian methods are nevertheless increasingly implemented, also in nonparametric and inverse problems. More recent theoretical research seem to indicate that many priors may give good results after all (e.g. Ghosal et al. (2000) and Ghosal and van der Vaart (2006)).

In relation to the discussion by Bickel and Li it is of interest to know whether it is possible to regularize by purely Bayesian methods. Given the close link between penalization and prior modelling the answer should of course be affirmative, but there are still many open questions. A fully Bayesian approach (as opposed to the empirical Bayes methods mentioned in Section 3.4) would spread priors on each model in a set of models deemed reasonable (e.g. models with different sets of regression variables, models based on different approximation schemes, models based on approximation schemes of different dimensionality) and combine these with prior weights on the models. One asks under what circumstances does the Bayesian machine yield good posteriors? First results have been obtained in Ghosal et al. (2003), Huang (2004), Belitser and Ghosal (2003), Lember (2004), but there are many open questions. Of special interest is to know whether "objective" priors, not dependent on arbitrary parameters, give the desired result. It may be noted that Bayesian methods average over regularization values, rather than select, which potentially should be advantageous.

Bayesian methods are connected to penalization both through direct interpretation of a penalty as a prior density and through BIC. BIC was developed for choosing between finitely many smoothly parametric models. Recent research appears to indicate that it cannot be used unchanged for regularizing (many) infinite-dimensional models. Of course, BIC penalizes more and hence leads to smaller models, but it seems unclear whether the usual interpretations of BIC versus AIC etc. are valid in more complicated settings. For instance, in Bayesian model selection there is interaction

between the manner by which prior is spread over the model and the weights given to models.

More research is needed in this area.

3 Confidence sets

Bickel and Li touch on the issue of confidence sets mainly in their discussion of the m out of n bootstrap. Some indication of the precision of (regularized) estimators is very desirable for their use in practice. Reality here appears to be not favorable. While estimation methods may through regularization adapt to a large number of models, honest confidence regions are typically mostly determined by the union of all models used. Thus adaptive procedures necessarily come with wide confidence margins, unless there is much a-priori information, no matter how smart the statisticians who implement it. See e.g. Cai and Low (2004), Cai and Low (2005), Juditsky and Lambert-Lacroix (2003), Hoffmann and Lepski (2002), and Robins and van der Vaart (2006).

As a side conclusion, one can mention that the sizes of credible regions attached to the Bayesian procedures based on model selection mentioned previously, appear to adapt to the underlying models, and hence such credible sets cannot be used as "honest" indications of uncertainty.

4 Cross validation

Bickel and Li discuss cross validation in Section 3.3, but appear not to come to a clear conclusion. This may well reflect the fact that the literature on cross validation is confusing and incomplete. Many general claims are made, but often seem to refer to specific situations. It appears that V-fold cross validation (with e.g. $V = 10$; the choice $V = n/\log n$ mentioned by Bickel and Li appears a theoretical choice only) is most popular in practice. Whether it is better than e.g. leave-one-out is not altogether clear, and probably depends on the setting.

Recent work in Keleş et al. (2004) extends cross validation to settings that do not immediately have the form of a prediction problem.

5 Aggregation

Aggregation of estimators aims at combining a given set of estimators, for instance through a (convex) linear combination, rather than choosing a "best one". If applied to a set of estimators obtained under various regularity levels, then it can be viewed as another method of regularization, which averages rather than selects. Recent and promising theoretical work is given in Yang (2000), Juditsky et al. (2005), Yang (2004), Nemirovski (2000), Juditsky and Nemirovski (2000), Tsybakov (2004). A striking feature of some of these methods is that they are highly constructive, giving very simple explicit algorithms.

6 Functionals on semiparametric models

Though the general definition of regularization in Section 2 is not restricted to this, their examples in Section 3 concern exclusively nonparametric estimation problems, such as regression and classification. Regularization also appears important for estimating certain functionals on large semiparametric models. As an example consider the semiparametric regression model $y = \theta x_1 + f(x_2) + e$, where x_2 may be a very high-dimensional covariate and the interest is in the effect θ of the one-dimensional covariate x_1. If f is suitably restricted, then θ can be estimated at the rate $n^{-1/2}$ if n replicates from the model are available. For instance, if f is a smooth function on a d-dimensional domain, then smoothness larger than $d/2$ would suffice. Semiparametric theory (Bickel et al., 1998; van der Vaart, 2002) as developed in the 1990s has mostly been concerned with such "regular" cases. However, particularly if the dimension d of x_2 is large, an a-priori assumption that the nonparametric part is smoother than $d/2$ appears problematic. This observation appears particularly relevant for the analysis of observational data in epidemiology or econometrics, where a large number of covariates may have to be included in the model to correct for possible confounding (Robins, 1997; van der Laan and Robins, 2003). It is not clear that nonsmooth cases can be easily treated through changes in the penalized likelihood or Bayesian paradigms (see the discussion of Murphy and van der Vaart (2000)), as regularization appears to require a bias-variance trade-off that is not easy to describe directly through the likelihood itself.

Some promising results using a new type of estimating equations have been obtained in Li et al. (2005).

Additional references

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113(3):301–413.

Belitser, E. and Ghosal, S. (2003). Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution. *The Annals of Statistics*, 31(2):536–559.

Birgé, L. (2006). Statistical estimation with model selection (*Brouwer lecture*). Preprint.

Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840.

Cai, T. T. and Low, M. G. (2005). On adaptive estimation of linear functionals. *The Annals of Statistics*, 33(5):2311–2343.

Cohen, A., Dahmen, W., Daubechies, I., and DeVore, R. (2001). Tree approximation and optimal encoding. *Applied and Computational Harmonic Analysis. Time-Frequency and Time-Scale Analysis, Wavelets, Numerical Algorithms, and Applications*, 11(2):192–226.

DeVore, R., Kerkyacharian, G., Picard, D., and Temlyakov, V. (2006). Approximation methods for supervised learning. *Foundations of Computational Mathematics. The Journal of the Society for the Foundations of Computational Mathematics*, 6(1):3–58.

DeVore, R. A. and Lorentz, G. G. (1993). *Constructive approximation*, Vol. 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531.

Ghosal, S., Lember, J., and Van Der Vaart, A. (2003). On Bayesian adaptation. *Acta Applicandae Mathematicae. An International Survey Journal on Applying Mathematics and Mathematical Applications*, 79(1-2):165–175.

Ghosal, S. and van der Vaart, A. W. (2006). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 34.

Hoffmann, M. and Lepski, O. (2002). Random rates in anisotropic regression. *The Annals of Statistics*, 30(2):325–396.

Huang, T.-M. (2004). Convergence rates for posterior distributions and adaptive estimation. *The Annals of Statistics*, 32(4):1556–1593.

Juditsky, A. and Lambert-Lacroix, S. (2003). Nonparametric confidence set estimation. *Mathematical Methods of Statistics*, 12(4):410–428 (2004).

Juditsky, A., Nazin, A., Tsybakov, A., and Vayatis, N. (2005). Recursive aggregation of estimators by mirror descent algorithm with averaging. *Problems of Information Transmission*, 41(4):368–384.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. *The Annals of Statistics*, 28(3):681–712.

Keleş, S., van der Laan, M., and Dudoit, S. (2004). Asymptotically optimal model selection method with right censored outcomes. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 10(6):1011–1037.

Kerkyacharian, G. and Picard, D. (2004). Entropy, universal coding, approximation, and bases properties. *Constructive Approximation. An International Journal for Approximations and Expansions*, 20(1):1–37.

Lember, J. (2004). On Bayesian adaptation. Preprint.

Li, L., Tchetgen, E., Robins, J., and van der Vaart, A. (2005). Robust inference with higher order influence functions: Parts I and II. Joint Statistical Meetings, Minneapolis, Minnesota.

Murphy, S. A. and van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–485.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, Vol. 1738 of *Lecture Notes in Mathematics*, pp. 85–277. Springer–Verlag, Berlin.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane, ed., *Latent variable modeling and applications to causality (Los Angeles, CA, 1994)*, Vol. 120 of *Lecture Notes in Statistics*, pp. 69–117. Springer–Verlag, New York.

Robins, J. M. and van der Vaart, A. W. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253.

van der Laan, M. J. and Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer Series in Statistics. Springer-Verlag, New York.

van der Vaart, A. W. (2002). Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, Vol. 1781 of *Lecture Notes in Mathematics*, pp. 331–457. Springer–Verlag, Berlin.

Yang, Y. (2000). Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87.

Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 10(1):25–47.

Rejoinder by Peter J. Bickel and Bo Li

We are grateful to all the discussants for making us think more deeply of the issues we have raised and face some issues raised by them.

1 The heuristic formulation of regularization in our sense

Tsybakov and van de Geer implicitly point out that the connection between our framework and prediction is obscure save for L_2 regression where the

identity

$$E(Y-\delta(Z))^2 = E(\delta(Z)-E(Y|Z))^2 + E(Y-E(Y|Z))^2$$

makes the problem of minimizing $E(Y-\delta(Z))^2$ equivalent to the problem of estimating $E(Y|Z)$ in the L_2 sense, as well as possible. Indeed, the general formulation can be extended, but only conditionally, to cover more general prediction.

Following the notation in Section 3 and the remarks of Tsybakov and van de Geer, the prediction problem is: given observations $X_i = (Z_i, Y_i), i = 1, \cdots, n, Z \in \mathcal{Z}, Y \in \mathcal{Y}, X \in \mathcal{X} = \mathcal{Z} \times \mathcal{Y}$ i.i.d P on $\mathcal{X}$,

1) Define an action space $\mathcal{A}$ (e.g. $\mathcal{A} = \{1, -1\}$ for 2 classification, $\mathcal{A} = \mathbb{R}$ for regression).

2) Define a loss function $\ell : \mathcal{X}^n \mapsto \mathcal{A}^{\mathcal{X}}$, where $\mathcal{A}^{\mathcal{X}}$ is all (measurable) function from $\mathcal{X}$ to $\mathcal{A}$.

3) Find a rule, $\delta(Z : \mathbf{X}_n)$, where $\mathbf{X}_n := (X_1, \cdots, X_n)$ and $X_{n+1} = (Z, Y)$ is a new observation such that

$$R(P, \delta) := \int \ell(y, \delta(z, \mathbf{X}_n)) \mathrm{d}P(z, y) \tag{1}$$

is "small".

A critical role is played by the Bayes risk $R_B(P)$ and procedure $\delta_{B,P}$ defined by

$$R_B(P) = \min_\delta R(P, \delta) = R(P, \delta_{B,P})$$

where

$$\delta_{B,P}(z) = \mathrm{argmin}_a \int \ell(y, a) \mathrm{d}P(y|z).$$

We could, at this point, define $\theta_k(P) = R(P, \delta_k(\cdot, \mathbf{X}_n))$, but in that case

$$\begin{aligned} \theta_k(P_n) &= \int \int \ell(y, \delta_k(z : x_1, \cdots, x_n)) \mathrm{d}P_n(x) \Pi_{i=1}^n \mathrm{d}P_n(x_i) \\ &= E^* \left(\frac{1}{n} \sum_{i=1}^n \ell(Y_i, \delta_k(Z_i : \mathbf{X}_n^*)) \right) \end{aligned}$$

the bootstrap expectation of our rule or equivalently the empirical risk of the rule obtained by "bagging" δ_k. However, if we define the conditional risk

$$R(P, \delta, \mathbf{X}_n) = \int \ell(y, \delta(z, \mathbf{X}_n)) \mathrm{d}P(z, y)$$

and let $\theta_k(P, \mathbf{X}_n) = R(P, \delta_k, \mathbf{X}_n)$ for a suitable sequence of δ_k than we can again require at a minimum,

(i) $\theta_k(P, \mathbf{X}_n) \xrightarrow{P} \theta_\infty(P) = R_B(P)$

(ii) $\theta_k(P_n, \mathbf{X}_n) - \theta_k(P, \mathbf{X}_n) \xrightarrow{P} 0$

and have a conditional variance-bias decomposition,

$$\theta_k(P_n, \mathbf{X}_n) - \theta_\infty(P) = (\theta_k(P_n, \mathbf{X}_n) - \theta_k(P, \mathbf{X}_n)) + (\theta_k(P, \mathbf{X}_n) - \theta_\infty(P)).$$

In fact, this point of view, conditioning on the training sample, is natural. For simplicity, we shall follow the notation of Tsybakov and write $R_n(\delta)$ for $R(P_n, \delta, \mathbf{X}_n)$ and $R(\delta)$ for $R(P, \delta, \mathbf{X}_n)$.

As both Tsybakov and van de Geer point out, the δ_k in the machine learning literature are usually obtained by specifying classes $\mathcal{D}_k$ such that δ_B can be arbitrarily well approximated by suitable $\delta_k \in \mathcal{D}_k$, and choosing

$$\delta_k(\cdot : \mathbf{X}_n) = \mathrm{argmin}_{\mathcal{D}_k} R_n(\delta) \tag{2}$$

The $\mathcal{D}_k$ sometimes correspond to the Bayes rules for P belonging to a sieve element $\mathcal{P}_k$, but empirical risk minimization is not usually maximum likelihood.

Again, as Tsybakov and van de Geer point out, the principal machine learning focus is on comparing $R(\delta_k)$, the actual posterior risk of δ_k with $\min_{\mathcal{D}_k} R(\delta)$ and the corresponding δ_k^*, the rule and (posterior) risk that an oracle knowing P as well as $\mathbf{X}_n$, but still forced to choose one of the δ_k.

The oracle chooses k^* so as to minimize $R(\delta_k) - R_B$. Imitating the oracle means choosing $\hat{k}$ so that, for moderate C depending weakly on P,

$$R(\delta_{\hat{k}}) - R_B \leq C(R(\delta_{k^*}) - R_B) \tag{3}$$

in expectation or with high probability.

As van de Geer points out,

$$R(\delta_k) - R(\delta_k^*) \leq -\big[(R_n(\delta_k) - R(\delta_k)) - (R_n(\delta_k^*) - R(\delta_k^*))\big]$$

each of which is a variance term. Writing

$$R(\delta_k) - R_B = (R(\delta_k) - R(\delta_k^*)) + (R(\delta_k *) - R_B)$$

we see that obtaining oracle properties:

(i) Requires control of the variance term through complexity control of $\mathcal{D}_k$ needed to apply empirical process theory, and

(ii) Having P and $\{\mathcal{D}_k\}$ such that the bias term is commensurate or at lest boundable by some known function of n and the (expected) variance term.

Van de Geer suggests and Koltchinskii (2006) discusses in detail various ways how to construct penalties $\hat{\pi}(k)$ such that

$$\hat{k} = \text{argmin}\{R_n(\delta_k) + \hat{\pi}(k)\}$$

achieves (3). The choices of $\mathcal{D}_k$ in terms of complexity and approximation properties depend on what properties, sparseness in particular representations, smoothness, $\cdots$, we assume about P leading to δ_B and the homogeneity of P to which P is assumed to belong.

2 A particular response to Tsybakov

Tsybakov and coworkers' remarkable inequality (2) which they have established for some methods and the even more surprising (5) established for mirror averaging, permit an easy choice of regularizing k if one considers δ_B for which the order of the approximation by convex or linear combinations of $\leq k$ functions from a dictionary $\mathcal{H}$ is known. If it is not, then is adaptation easily achieved? Of course, the choice of $\mathcal{H}$ remains an act. Mirror averaging seems a remarkable technique for on line classification or regression, but one presumably wants M to increase with n, and it seems plausible then to make M sequentially data-determined as well. Analyzing such procedures should be of importance. Since the computationally

impossible procedure obtained by averaging mirror average estimates over all permutations (or more realistically an average over a sample of permutations) should do strictly better for convex losses. It seems to be less compelling for a complete sample. Finally, we'd like to, correct a misconception about (3.9). k' is a dummy index. We should have just said $\mathcal{L}(\mathcal{H})$. Our heuristics suggest one should regularize in both k and γ.

3 A particular response to van de Geer

We address van de Geer's last point that, since there is a limit, $\theta(P)$, to $\theta_k(P)$ regularization is not involved in the m out of n bootstrap. We note that there is a limit of $\theta_k(P)$ in our general formulation also. The limit is, however, not defined for discrete P. That is the case here as well. We elaborate on an example of Section 5. Consider a $1-\alpha$ bootstrap confidence bound $\theta_n^*(P_n)$ for the upper point of the support c of a distribution with unknown density f at c (continuous from the left). The distribution of the pivot $n(c - X_{(n)})$ where $X_{(n)}$ is the sample maximum converges to an $\mathcal{E}(f(c))$ distribution where $\mathcal{E}(\lambda)$ is the exponential distribution with density $p(x, \lambda) = \lambda e^{-\lambda x}$. Thus the quantity that an upper $1 - \alpha$ confidence bound based on $X_{(n)}$ should estimate is $-\frac{\log \alpha}{f(c-)}$. But, as was already noted by Bickel and Freedman (1981), the bootstrap distribution of $n(X_{(n)} - X^*_{(n)})$ doesn't converge to a fixed distribution at all. It converges in law to a random probability distribution which with positive probability has mass at 0—see also Bickel and Sakov (2005). As one might expect, in estimating a density one needs to regularizing, use the m out of n bootstrap with $m \to \infty, \frac{m}{n} \to 0$. And if one sees one goal as estimating $-\frac{\log \alpha}{f(c-)}$ as well as possible, then indeed the best choice of m reflects the appropriate balance between "bias" and "variance" to get for instance rate $n^{-1/3}$ if one assumes $|f'(c-)| \leq M < \infty$. The same consideration applies to Efron's famous example $\theta_n(P_n) = n\mathrm{Var}X_{(\frac{n}{2})}$, where $X_{(\frac{n}{2})}$ is the median, which tends to $\frac{1}{4f^2(F^{-1}(\frac{1}{2}))}$. The ordinary bootstrap is consistent but does not converge at optimal rates.

4 Response to Yu

Yu points out a very interesting new direction in the machine learning literature in which prediction is related to "stability" of a rule in terms of

the effect of small changes in the training sample $\mathbf{X}_n$. One version due to Devroye and Wagner (1979) can be formulated in terms of the conditional decomposition above. They define "error stability" of δ by, in our notation,

$$|R(P_n, \delta, \mathbf{X}_n) - R(P_n, \delta_{-i} : \mathbf{X}_n)| \leq \beta$$

for $i = 1, \cdots, n$, where

$$\delta_{-i}(z : \mathbf{X}_n) := \delta(z : \mathbf{X}_n^{-i})$$

and $\mathbf{X}_n^{-i} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n)$, and we assume that δ is defined for all $\mathbf{X}_n$ and n.

The notion of stability of parameters $\theta(P)$ was developed in statistics by Hampel, following work of Huber (1964) and Hodges (1967)— see Hampel et al. (1986). The goal was to construct procedures which would be robust to gross errors and other data perturbations. The "influence function" formulation is to consider for a parameter $\theta(P)$,

$$\Psi(Q, P) = \frac{\partial}{\partial \epsilon} \theta((1-\epsilon)P + \epsilon Q)|_{\epsilon=0} \tag{1}$$

the Gateaux derivative of θ in the "direction" Q. In analogy to the total differential in $\mathbb{R}^p$, one expects that

$$\Psi(Q, P) = \int \Psi(x, P) \mathrm{d}Q(x) \tag{2}$$

where $\Psi(x, P) := \Psi(\delta_x, P)$, δ_x is the Dirac measure, and $\int \Psi(x, P) \mathrm{d}P(x) = 0$. The influence function measures the impact of point mass at x on the parameter. Rigorous justification of statements such as (2) (under conditions) may be found in various places, including van der Vaart (1998) and Bickel et al. (1998).

Robustness is generated by $|\Psi(x, P)| \leq M < \infty$ for all x, P. For a discussion of the implication of this assumption and its relation to nonparametric inference, see Bickel and Ritov (2000).

The relation to stability comes by noting that if we write

$$V(P_n, P) := \int \ell(y, \delta(z, \mathbf{X}_n)) \mathrm{d}P(z, y)$$

when we identify $X_1, \cdots, X_n$ with P_n, i.e., assume that δ is symmetric in $\mathbf{X}_n$, then stability says that

$$\left|V(P_n, P_n) - V(P_n^{-i}, P_n)\right| \leq \beta \tag{3}$$

where P_n^{-i} is the empirical distribution of $\mathbf{X}_n^{-i}$.

Writing $P_n = (1 - \frac{1}{n})P_n^{-i} + \frac{1}{n}\delta_{X_i}$, we see that if $P \mapsto V(P, P_n)$ has influence function $\Psi(\cdot, P)$ for all P (we suppress $\mathbf{X}_n$) with $|\Psi| \leq M$, then (3) holds with $\beta = \frac{M}{n}$. This follows from the identity,

$$V(Q, P_n) - V(P, P_n) = \int_0^1 \Psi(x, (1-\lambda)P + \lambda Q)\mathrm{d}\lambda \tag{4}$$

Here is an application of this formulation. We consider rules δ obtained by empirical risk minimization with penalty, $\delta = \delta_{\gamma(P_n)}(\cdot)$, where $\gamma(Q)$ minimizes

$$W(\gamma, Q) = \int \ell(y, \delta_\gamma(z))\mathrm{d}Q(z, y) + \lambda K(\gamma) \tag{5}$$

where $\gamma \in \mathbb{R}^p$. Then δ is stable with $\beta = \frac{M}{n}$ for $M = \frac{[M']^2}{\epsilon}$ where

$$\sup_{z,y} \|D_\gamma \ell(y, \delta_\gamma(z)\|_\infty \leq M' < \infty \tag{6}$$

and

$$\inf_{z,y} \|D_\gamma^2 \ell(y, \delta_\gamma(z) + \lambda D_\gamma^2 K(\gamma)\|_\infty \geq \epsilon > 0 \tag{7}$$

We use D_γ for a total differential in γ and D_P for the operator corresponding to the calculation (1) of the influence function, Frechet differentiation under our conditions.

Note that ridge regression is stable according to these conditions, but L_1 penalties are not.

Proof. Under these conditions it is easy to check that the influence function of $W(\gamma(P), P)$ is

$$\Psi(x, P) = D_P W(\gamma(P), P)(\delta_x) + D_\gamma W(\gamma(P), P) D_P \gamma(P)(\delta_x)$$

and

$$D_P \gamma(P)(\delta_x) = -\frac{D_{\gamma P} W(\gamma(P), P)}{D_\gamma^2 W(\gamma(P), P)}(\delta_x)$$

where differentiation of $\theta(P)$ with respect to P in direction Q is given by (1). In our case

$$D_P W(\gamma(P), P)(Q) = \int \ell(y, \delta_{\gamma(P)}(z))\mathrm{d}(Q - P)(z, y)$$

$$D_{P\gamma}W(\gamma(P),P)(Q) = \int D_\gamma \ell(y,\delta_{\gamma(P)}(z))\mathrm{d}(Q-P)(z,y)$$

$$D^2_\gamma W(\gamma(P),P) = \int D^2_\gamma \ell(y,\delta_{\gamma(P)}(z))\mathrm{d}P(z,y) + \lambda D^2_\gamma K(\gamma)$$

and the result follows. □

Evidently, a Lipschitz condition,

$$|W(\gamma(P),P) - W(\gamma(Q),Q)| \leq \beta\|P-Q\|$$

where $\|\cdot\|$ is variational norm suffices for stability.

Since the jackknife is known to be closely connected to numerical analysis — see Gray and Schucany (1972) on the one hand and to robustness in statistics on the other, the connection Yu suggests should be followed seems promising.

We would also argue that obtaining oracle inequalities in terms of stability suggests that one is concerned with P of the form $(1-\epsilon)F+\epsilon G$, where the conditional distribution of Y given $Z=z$ are very different under F and G, the supports of the marginal distributions of Z under F and G are nearly disjoint, and ϵ is very small. That is, one is dealing with the same sort of concerns as in the statistical literature.

5 Response to Valdés and Rivero

Valdés and Rivero bring up an important point to correct a possible implication of our remark, that viewing the result of (1.4) as the mode of a Bayes posterior distribution provides automatic regularization. In fact, as Theorem 1.5.3 of Wahba (1990) shows, the usual smoothing spline estimate resulting from penalized least squares with $\lambda = \frac{c}{n}$ is, for each fixed n, an improper Bayes estimate. Many results, for instance, of Cox (1993) and Freedman (1999) indicate problems that can arise with Bayesian inference in nonparametric models. The analysis of Bayesian methods, proper or improper, in nonparametric situations requires much more research.

On the other hand, the issues they raise with types of convergence needed in the parametric case are, we believe, mistaken. Regularization is not needed for smooth Euclidean parameters, since they can be defined by plug in directly.

6 Response to Fan

Jianqing Fan's discussion enlarging on each of our examples was a pleasure. We heartily second his comments on the important of sparsity and also his caution that the hardest and most subject dependent questions have to be do with finding the sense in which the situation at hand can be well approximated by a sparse representation. For instance, to say that high dimensional covariates live on a very low dimensional smooth manifold and that a regression is a smooth function of these is a sparse description. So is saying the covariates are independent of each other and the regression is an additive model. But sparseness in the second sense is complexity in the first.

Although we obviously agree on the different goals of prediction and causal inference in model selection, we don't find the criteria of sparsity continuity and unbiasedness equally compelling. Sparsity and continuity resonate but not unbiasedness. It's clear that one wants to isolate factors that are important (large). But important has to be judged using knowledge outside the data.

Fan's successful application of a semiparametric model to microarrays is impressive and an excellent illustration of the importance of regularization even when the parameters of interest are estimable at classical rates. The covariance estimation methods he considers are in the same spirit as ours, though the approximations are not through banding. He proposes a sieve more appropriate to the finance applications he considers.

The effect of unknown bias, a necessary consequence of regularization, are more troubling in our view. One of us discussed these issues in terms he now finds too extreme in Bickel and Ritov (2000). But concerns about interpretation remain.

7 Response to van der Vaart

We are grateful to Aad van der Vaart for extending the discussion to several topics untouched by us or previous discussants.

To the topics he lists under more research being needed in the Bayesian framework we would add one he surprisingly omits, the validity of inference at the Bernstein-von Mises level for regular functionals. Initial interesting

work in this direction is due to Kleijn and van der Vaart (2005). We also found of particular interest his remarks on priors putting discrete masses on a sieve of finite dimensional models, which is just what BIC is based on. We presume that his reference to the contrast between BIC and AIC is between BIC's always choosing the lowest dimensionality model if a member of the sieve is generating the data and AIC's doing best in terms of prediction if no member of the sieve is entirely correct. Results on model selection for sieves of infinite dimensional models would certainly be interesting.

We now address the point raised by van der Vaart in his indication that we do not give a clear statement in favor of $V-$fold cross validation versus leave-1-out CV. Consider selecting among all linear spaces generated by subsets of a wavelet basis of size $\log 2n$ — see Birgé and Massart (1997, p. 61, 62). Mallows C_p which does not yield minimax results for Sobolev spaces (or optimizes oracle inequalities for classes of smooth functions) is equivalent to leave-1-out cross validation, Birgé and Massart (1997), and thus is suboptimal. However, $V-$fold CV corresponding to, say, a training set of size $n - \log n$ and a test set of size log n, does yield minimax results in the Birgé and Massart situation. Apply Theorem 6 of Bickel et al. (2006), for instance, to see this. This theorem and similar results for the white noise model indicate $V-$fold cross validation with a test sample size which is growing, but slowly with n, works extremely generally. The only example where leave-1-out is better we believe is estimation in smooth parametric models where regularization is not needed. Most of the other points touched on by van der Vaart, such as cautions on inference in non and semiparametric models, also appeared in the other discussions including ours, though we are intrigued to hear of potential solutions such as Li et al. (2005).

In conclusion, we again thank all the discussants for their very stimulating comments.

Additional references

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Bickel, P. J. and Ritov, Y. (2000). Non- and semiparametric statistics: compared and contrasted. *Journal of Statistical Planning and Inference*,

91(2):209–228. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

Cox, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *The Annals of Statistics*, 21(2):903–923.

Devroye, L. P. and Wagner, T. J. (1979). Distribution-free performance bounds for potential function rules. *Institute of Electrical and Electronics Engineers. Transactions on Information Theory*, 25(5):601–604.

Freedman, D. (1999). On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1140.

Gray, H. L. and Schucany, W. R. (1972). *The generalized jackknife statistic*, Vol. 1 of *Statistics Textbooks and Monographs*. Marcel Dekker Inc., New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York. The approach based on influence functions.

Hodges, J. L., Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)*, pp. Vol. I: Statistics, pp. 163–186. Univ. California Press, Berkeley, Calif.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101.

Kleijn, B. and van der Vaart, A. (2005). The Bernstein-Von-Mises theorem under misspecification. Unpublished.

Koltchinskii, V. (2006). 2004 IMS Medallion Lecture: Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6). To appear.

Li, L., Tchetgen, E., Robins, J., and van der Vaart, A. (2005). Robust inference with higher order influence functions: Parts I and II. Joint Statistical Meetings, Minneapolis, Minnesota.

VAN DER VAART, A. W. (1998). *Asymptotic statistics*, Vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

WAHBA, G. (1990). *Spline models for observational data*, Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.