Oracle Inequalities and Nonparametric Function Estimation

Iain M. Johnstone

Abstract. In non-parametric function estimation, partial prior information about the unknown function is often expressed by a family of models or estimators, among which a choice must be made. Oracle inequalities bound the mean squared error of a given estimator in terms of the (unknowable) best possible choice of model for the unknown function. This survey concentrates on three examples: the James Stein estimator, soft thresholding, and complexity penalized least squares and as illustrations, we describe some consequences for adaptive estimation.

1991 Mathematics Subject Classification: 62G07, 62G20, 62C20
Keywords and Phrases: adaptive estimation, complexity penalty, James-Stein estimator, minimax estimation, thresholding, unconditional basis, wavelet shrinkage

1 Introduction

Statistical theory aims in part to articulate when and why certain applied methods of data analysis succeed. With emergence of large, often instrumentally acquired datasets, recent decades have seen a focus on "nonparametric" models in which the number of model parameters grows with the size of available data. Here we focus on the estimation (or "recovery" or "denoising") of functions observed in additive noise and describe some relatively simple inequalities that encode information on the effect of sparse representation on the quality of estimation.

A common caricature is to posit observed data $y \in \mathcal{R}^n$ with structure $y = \mu + \epsilon z$. Here μ is an unknown function which one desires to "estimate" or "recover", and $z \in \mathcal{R}^n$ is a vector of standard Gaussian noise of known scale ϵ. When expressed in terms of coefficients in an orthonormal basis $\{\psi_i, I \in \mathcal{I}_n\}$, the model becomes

$$y_I = \mu_I + \epsilon z_I \qquad\qquad I \in \mathcal{I}_n. \tag{1}$$

Here z_I are independent Gaussian noises of mean zero and variance one. This sequence form of the "Gaussian white noise model", whether finite as here, or infinite, as in Section 1.1 below, is the conceptually and technically simplest model of nonparametric estimation. Extensions to correlated noise and indirect data $y = K\mu + \epsilon z$ are possible, but not covered here.

Examples 1. (a) The *equispaced, fixed design* in which $y_I = f(t_I) + \sigma z_I$, with f an unknown function defined on $[0,1]$ and $t_I = I/n, I = 1, \dots n$.

(b) An initial segment of the *continuous Gaussian white noise* model. Suppose that W_t is a standard Brownian motion (or sheet) and that one observes $dY_t = f(t)dt + \epsilon dW_t$ for $t \in D$, a compact set in $\mathbb{R}^d$. If $d = 1$ and $D = [0,1]$, this may be interpreted as $Y_t = \int_0^t f(s)ds + \epsilon W_t$ for $0 \le t \le 1$. Take inner products with elements $\{\psi_I, I \in \mathcal{I}\}$ of a complete orthonormal basis for $L_2[0,1]$ and set $y_I = \langle \psi_I, dY \rangle$, $\theta_I = \langle \psi_I, f \rangle$ and $z_I = \langle \psi_I, dW \rangle$. This gives an infinite sequence version of (1), to be used in Section 1.1 below. To recover precisely (1), consider an initial segment of cardinality n of the index set $\mathcal{I}$. A discrete orthogonal wavelet transform of model (i) yields an approximation to this initial segment (after calibrating $\epsilon = \sigma n^{-1/2}$, cf. [10]).

(c) *Redundant regression.* Suppose that there are given vectors (or signals) $x_1, \dots x_p \in \mathbb{R}^n$, and that it is thought useful, for reaons of parsimony, interpretability or otherwise, to represent μ in terms a few of the x_i. Collecting x_i as columns of a "design matrix" $X = [x_1 \cdots x_n]$, one obtains the standard, homoscedastic Gaussian linear regression model $y = X\beta + \epsilon z$. In traditional parametric regression analysis, it is supposed that $p < n$ and that $\mu \in \text{span}\{x_i\}$. However, we specifically consider two "non-parametric" cases: a) $p = n$ and x_i orthogonal (i.e. equivalent to (1)), and b) $p > n$ and *not* orthogonal - here the x_i might be a class of basic signals from a (possibly highly) redundant *dictionary* $\mathcal{D}$ and we seek a parsimonious representation of μ in terms of as few elements of $\mathcal{D}$ as possible.

Assessing error. An estimator $\hat{\mu} = \hat{\mu}(y)$ is a function of observed data y: we wish to quantify and compare the quality of estimation as $\hat{\mu}$ varies. Simplest to work with is mean squared error (MSE):

$$r_\epsilon(\hat{\mu}, \mu) = E_\mu|\hat{\mu} - \mu|^2 = \int |\hat{\mu} - \mu|^2 \phi_\epsilon(y - \mu) dy. \tag{2}$$

Here $\phi_\epsilon(z)$ denotes the probability density function of ϵz. The notation $r_\epsilon(\hat{\mu}, \mu)$, mnemonic for "risk", hints at the possible and frequently desirable use of more general error norms $\|\hat{\mu} - \mu\|$ or loss functions $L(\hat{\mu}, \mu)$.

The error $\hat{\mu} - \mu$ is usually decomposed into a zero-mean *stochastic* component $\hat{\mu} - E_\mu \hat{\mu}$ and a deterministic component, the *bias*, $E_\mu \hat{\mu} - \mu$. For quadratic error measures, these components are uncorrelated, so that the MSE is the sum of variance and squared bias terms. In particular, for a *linear* estimator $\hat{\mu}_L(y) = Ly$,

$$r(\hat{\mu}_L, \mu) = \epsilon^2 \text{tr } LL^t + |L\mu - \mu|^2. \tag{3}$$

The quality of approximation of μ by the operator L is thus balanced against the complexity of L, as measured by the variance term, which for example becomes $\epsilon^2 m$ in the case of orthogonal projection onto a subspace of dimension m. Already visible here is the important role that approximation theory plays in analysing the deterministic component of error. For non-linear estimators that, implicitly or explicitly, involve a choice among linear estimators, the analysis of the stochastic term is facilitated by the concentration of measure phenomenon (Section 4).

Models and Estimators. A model is a subset M of the full parameter space $\mathbb{R}^n$. A family of models $\{M_\alpha, \alpha \in \mathcal{A}\}$ is one device commonly used to represent imperfect and partial information about the unknown μ. Often there is a natural estimator $\hat{\mu}_\alpha$ associated with each model and in this paper we simplistically conflate choice of model with choice of the associated estimator.

Examples 2. *(a) Spheres and linear shrinkage.* For positive α, let M_α be the sphere $|\mu| = \alpha$: this might correspond to prior information about the signal-to-noise ratio. Natural corresponding estimators are given by *linear shrinkage*: $\hat{\mu}_\alpha = \gamma y$ where $\gamma = \gamma(\alpha)$ is obtained by minimizing the MSE in (3), namely $n\gamma^2 + (1-\gamma)^2|\mu|^2$, on M_α to obtain the Wiener filter $\gamma(\alpha) = \alpha^2/(n+\alpha^2) \in (0,1)$.

(b) Subspaces and projections. In the regression setting of Example 1(c) above, to each subset $J \subset \{1, \ldots, p\}$ of the full variable list is associated a linear model $M_J = \text{span}\,\{x_j, j \in J\}$. The corresponding estimators are orthogonal projections P_J on M_J: these are the least squares estimators on the assumption that $\mu \in M_J$.

Ideal Risk Given a family $\mathcal{A}$ of models (or corresponding estimators), and for a given unknown μ, the best attainable MSE is given by the *ideal risk*

$$\mathcal{R}_\epsilon(\mu, \mathcal{A}) = \inf_\alpha R(\hat{\mu}_\alpha, \mu).$$

Thus, in example (a), the ideal linear shrinkage risk is

$$\mathcal{R}_\epsilon(\mu, LS) = n\epsilon^2|\mu|^2/(n\epsilon^2 + |\mu^2|). \tag{4}$$

Outline of paper. Of course, μ is not known, and without access to an oracle who divulges the best α, the ideal risk is not attainable by an estimator depending on the data y alone. Nevertheless, it acts as a useful benchmark, and we seek estimators that in an appropriate sense optimally mimick the ideal risk. Such estimators turn out to be non-linear, and in particular, not members of the family $\hat{\mu}_\alpha$. For three settings and estimators, oracle inequalities are presented in Theorems 3, 5 and 8 – we emphasize that the inequalities are non-asymptotic and uniform in character, holding for all n, ϵ and for all $\mu \in \mathbb{R}^n$.

Corresponding lower bounds (although asymptotic in n) show that without some restriction on, or further information about μ, the inequalities cannot be improved, and thus represent in some sense the necessary "price" for searching over a class of models/estimators of a given size.

Oracle inequalities are neither the beginning nor the end of a theory, but when available, are informative tools. For example, Theorems 3, 5 and 8 may also be used to derive asymptotic (i.e. low noise ϵ) results within a framework of adaptive minimax estimation: this class of applications is considered in a connected sequence of "illustrations" in the continuous Gaussian white noise model, which we now introduce.

1.1 Illustration: Asymptotic Minimax Estimation.

The continuous Gaussian white noise model is that of Example 1(b). Because of Parseval's inequality $\int_0^1 (\hat{f} - f)^2 = \sum_I (\hat{\theta}_I - \theta_I)^2 = \|\hat{\theta} - \theta\|^2$, estimation error can equally well be measured in the sequence domain. To evaluate estimators, we use the minimax principle - although inherently conservative and not universally accepted, we find that it leads to clear structures and informative results. Thus, estimators are assessed by their worst case risk over a given Θ. The *minimax risk* measures the best attainable such maximum risk, within a class $\mathcal{E}$ of estimators: $R_{\mathcal{E}}(\Theta, \epsilon) = \inf_{\hat{\theta} \in \mathcal{E}} \sup_{\theta \in \Theta} r_\epsilon(\hat{\theta}, \theta)$. The symbols $\mathcal{E} = N, L, D, \ldots$ refer to specific estimator classes: all non-linear, all linear, all threshold rules etc. Finally, estimator $\hat{\theta}$ is called *asymptotically* $\mathcal{E}-$ *minimax* if

$$\sup_{\theta \in \Theta} r_\epsilon(\hat{\theta}, \theta) = R_{\mathcal{E}}(\Theta, \epsilon)(1 + o(1)), \qquad \epsilon \to 0.$$

In order to describe a flexible and scientifically meaningful class of parameter spaces Θ, we employ a *dyadic sequence* notation, in which $I = (j, k)$, with $j = 0, 1, \ldots$ and $k = 1, \ldots, 2^j$. The primary motivation comes from orthonormal bases of wavelets $\{\psi_{jk}\}$, which, under suitable regularity and decay conditions on the wavelets, and with suitable modifications to handle intervals, form unconditional bases for many function spaces of interest ([22, 15, 3]). Thus their norms may be characterized in terms of conditions on $|\theta_I|$. For example, let χ_I denote the indicator function of the interval $[(k-1)2^{-j}, k2^{-j}]$: the sequence of (quasi-)norms $\| \cdot \|_{\alpha,p}$, defined for $0 < \alpha < \infty, 0 < p \le \infty$ by

$$\|\theta\|_{\alpha,p}^p = \int_0^1 [\sum_I (2^{aj} |\theta_I| \chi_I)^2]^{p/2}, \qquad a = \alpha + 1/2,$$

are equivalent, (for $p > 1$ and $\alpha \in \mathbb{N}$) to the traditional Sobolev norms $\|f\|_{W_p^\alpha}^p = \int_0^1 |f^{(\alpha)}|^p + |f|^p$. In the Hilbertian case $p = 2$, these take the simpler form

$$\|\theta\|_{\alpha,2}^2 = \sum_{j \ge 0} 2^{j\alpha} |\theta_j|^2, \qquad |\theta_j|^2 = \sum_k |\theta_{jk}|^2.$$

As parameter spaces, we thus use norm balls: $\Theta_{\alpha,p}(C) = \{(\theta_I) : \|\theta\|_{\alpha,p} \le C\}$, which are analogs of size restrictions on derivatives, but measured in L_p norms.

In practice, the values of (α, p, C) will not be known, and rather than seeking a minimax estimator for a single such $\Theta_{\alpha,p}(C)$, we look for estimates with an adaptive minimaxity property. Thus, suppose that a *scale* of spaces $\mathcal{S} = \{\Theta_\nu(C) : \nu \in \mathcal{V}, C > 0\}$ is given, where ν is an order parameter, such as (α, p) above, and C a scale parameter. Then $\hat{\theta}$ is *adaptively* $\mathcal{E}-$*minimax* if (i) the definition of $\hat{\theta}$ is independent of (ν, C), and (ii) $\hat{\theta}$ is asymptotically $\mathcal{E}-$minimax *for all* (ν, C).

2 Linear Shrinkage and Orthogonal Invariance

A celebrated result in parametric statistics, due to Stein [24], is the inadmissibility of the maximum likelihood estimator $\hat{\mu}^0(y) = y$ in model (1) as soon as $n \ge 3$. Indeed, [17] showed that adaptive linear shrinkage

$$\hat{\mu}^{JS+}(y) = (1 - \hat{\gamma})_+ y, \qquad \hat{\gamma} = (n-2)\epsilon^2 / |y|^2,$$

is everywhere better than $\hat{\mu}^0$, in the sense that for all $\mu \in \mathbb{R}^n$, $r_\epsilon(\hat{\mu}^{JS+}, \mu) < r_\epsilon(\hat{\mu}^0, \mu) \equiv n\epsilon^2$. Here $a_+ = \max(a, 0)$. The result was and remains surprising because it can seem counterintuitive that combining data from statistically completely independent problems, represented by each coordinate in (1), leads to better MSE properties.

A simple proof was later given by Stein [25], using his unbiased estimate of risk to show that $\hat{\mu}^{JS}(y) = (1 - \hat{\gamma})y$, necessarily worse than $\hat{\mu}^{JS+}$, satisfies

$$r(\hat{\mu}^{JS}, \mu) = E_\mu\{n - (n-2)^2|y|^{-2}\} < n. \tag{5}$$

(where, for simplicity, $\epsilon = 1$ here.) Using in (5) the fact that the distribution of $|y|^2$ can be represented as the mixture of central chi-squared distributions χ^2_{n+2P} with P distributed as a Poisson variate with mean $|\mu|^2/2$, and applying Jensen's inequality, one obtains our first oracle inequality.

Theorem 3 ([7]). *In model* (1)*, suppose* $n \geq 3$. *For all* $\mu \in \mathbb{R}^n$,

$$E|\hat{\mu}^{JS} - \mu|^2 \leq 2\epsilon^2 + \frac{(n-2)\epsilon^2|\mu|^2}{(n-2)\epsilon^2 + |\mu|^2}. \tag{6}$$

In view of (4)*, this implies*

$$r_\epsilon(\hat{\mu}^{JS+}, \mu) \leq 2\epsilon^2 + \mathcal{R}_\epsilon(\mu, LS). \tag{7}$$

Thus, the classical James-Stein estimator comes within an additive penalty of $2\epsilon^2$ of mimicking the ideal linear shrinkage estimator. This performance is impressive when calibrated against the minimax risk $R_N(\mathbb{R}^n, \epsilon)$, in this problem $n\epsilon^2$.

However it should be noted that this inequality is orthogonally invariant, and makes no use of the particular basis in which the unknown signal μ is represented.

2.1 Illustration: Levelwise shrinkage in the dyadic sequence model.

In the dyadic sequence model of Section 1.1, group coefficients by level $j : y_j = (y_{jk})_{k=1}^{2^j}$. Form a *levelwise* James Stein estimator $\hat{\theta}^{LJS}$ by applying James-Stein shrinkage to y_j: $\hat{\theta}_j^{LJS} = \hat{\theta}^{JS+}(y_j)$, at least for levels j below a cutoff $J = \log_2 \epsilon^{-2}$, above which $\hat{\theta}_j^{LJS}$ simply estimates zero. [Recall the calibration $n = \epsilon^{-2}$ of Example 1(b).] The MSE of the $\hat{\theta}^{LJS}$ may then also be represented levelwise:

$$E\|\hat{\theta}^{LJS} - \theta\|^2 = \sum_{j < J(\epsilon)} E|\hat{\theta}^{JS+}(y_j) - \theta_j|^2 + \sum_{j \geq J(\epsilon)} |\theta_j|^2.$$

The oracle inequality (7) may be applied to each level j in the first sum, while the geometric weights 2^{aj} used to define $\Theta_{\alpha,2}$ imply that the second sum is negligible for small ϵ. For the scale of *Hilbert* spaces $\mathcal{S}_2 = \{\Theta_{\alpha,2}(C) : \alpha > 0, C > 0\}$:

Theorem 4 ([7]). $\hat{\theta}^{LJS}$ *is adaptively minimax over* $\mathcal{S}_2$.

This recovers and extends a notable result of Efroimovich & Pinsker [14], originally formulated in the Fourier basis. In fact, one verifies relatively easily that $\hat{\theta}$ is adaptively minimax among linear estimates (from the ideal linear shrinkage risk) and then appeals to the celebrated theorem of Pinsker [23]), which shows that for the *ellipsoids* occurring in $\mathcal{S}_2$, linear minimax rules are actually asymptotically minimax among all non-linear estimates.

This levelwise application of an oracle inequality is shows how the dyadic sequence model allows a "lifting" of results from a symmetric and "parametric" setting (an exchangeable multivariate normal law at each level) to a non-parametric, infinite-dimensional model. Other examples of this type may be found in [7, 9].

3 Orthogonal Regression and Thresholding

To this point, we have considered only orthogonally invariant estimators. However, a basic principle is that sparsity of representation of a signal in a given basis leads to better estimation, and to exploit such sparsity, non-linear estimators are needed.

Thus, assume the orthonormal basis leading to coefficients (1) is chosen so that $\{\mu_i\}$ contains few large coefficients, although of course it is not known in advance *which* among the co-ordinates are important.

In this orthogonal regression setting, the least squares subset selection estimators have a simple co-ordinatewise representation: the $j-$th component of $\hat{\mu}_J(y)$ equals y_j if $j \in J$ and 0 otherwise. Thus, the least squares estimators have the form of diagonal projections (DP below). The mean squared error of $\hat{\mu}_J$ is then the sum of terms which measure either variance or bias:

$$r(\hat{\mu}_J, \mu) = \sum_{j \in J} \epsilon^2 + \sum_{j \notin J} \mu_j^2.$$

The ideal risk for among all such diagonal projection estimators can therefore be found by minimizing termwise:

$$\mathcal{R}_\epsilon(\mu, DP) = \inf_J r(\hat{\mu}_J, \mu) = \sum_j \mu_j^2 \wedge \epsilon^2.$$

To quantify sparsity, order the squared magnitudes of the components of μ via $\mu_{(1)}^2 \geq \mu_{(2)}^2 \geq \ldots \geq \mu_{(n)}^2$ and define *compression numbers* $c_j^2 = \sum_{k>j} \mu_{(k)}^2$. The number of large coefficients is measured by $N(\epsilon) = \#\{j : |\mu_j| > \epsilon\}$, and we have

$$\mathcal{R}_\epsilon(\mu, DP) = \epsilon^2 N(\epsilon) + c_{N(\epsilon)}^2,$$

which shows an intimate connection between ideal risk and the compressibility of the signal in this basis.

Various forms of thresholding estimator can be introduced: here we consider soft thresholding:

$$\hat{\mu}_j^{ST}(y) = \operatorname{sgn}(y_j)(|y_j| - \lambda)_+.$$

The key points are that the estimator acts co-ordinatewise and that there is a threshold zone $[-\lambda, \lambda]$ in which the data is interpreted as noise and "discarded".

Theorem 5 ([6]). *If $\lambda = \sqrt{2\log n}$, then for all $\mu \in \mathbb{R}^n$,*

$$r_\epsilon(\hat{\mu}^{ST}, \mu) \leq (2\log n + 1)[\epsilon^2 + \mathcal{R}_\epsilon(\mu, DP)]. \tag{8}$$

Since the logarithmic penalty is of small order relative to n, the result shows that sparsity, as measured by ideal risk, implies good estimation. The bound is valid for all sample sizes and all μ. There has been much work on the choice of threshold λ - the choice given here is attractive for its conservatism: since for independent and identically distributed $N(0,1)$ variates z_i, $P(\max_{1\leq j\leq n}|z_j| > \sqrt{2\log n}) \to 0$, it follows that $P(\hat{\mu}^{ST} = 0|\mu = 0) \to 1$. For more on these issues and numerical examples, see [11]. Smaller choices of λ, even depending on the data y, lead to better mean squared error in exchange for less conservatism [7]. Natural extensions of Theorems 5 and 6 to correlated noise exist [19]

Optimality. Absent extra restrictions on μ, the factor $2\log n$ is optimal:

Theorem 6 ([6]). *As $n \to \infty$,*

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^n} \frac{r(\hat{\mu}, \mu)}{\epsilon^2 + \mathcal{R}_\epsilon(\mu, DP)} \geq (2\log n)(1 + o(1)).$$

The lower bound arises from the difficulty of distinguishing rare true signal components from the also infrequent extremes of the white Gaussian noise z_i. Indeed, suppose $\epsilon = 1$ and that the values μ_i are drawn independently from a two point prior distribution with masses of probability $1 - \delta_n$ at 0 and δ_n at $\bar{\mu}_n$. Choosing $\delta_n = \log n/n$ and $\bar{\mu}_n \sim (2\log\delta_n^{-1})^{1/2}$, it turns out that the posterior distribution of μ_i, having observed even a value of $y_i > \bar{\mu}_n$, is still concentrated on $0 : P(\mu = 0|y = \bar{\mu}_n + z) \approx 1$, for z large and fixed, as $n \to \infty$. Hence, with probability δ_n, the estimator is forced to make an error of order $\bar{\mu}_n^2 \sim 2\log n$.

3.1 Illustration: thresholding in the dyadic sequence model.

Return to the dyadic sequence model, and apply soft thresholding at $\lambda = \epsilon\sqrt{2\log\epsilon^{-2}}$ to the first $n = \epsilon^{-2}$ coefficients. In other words, $\hat{\theta}_I^T(y) = \eta_{ST}(y_I, \lambda)$ for all I with $j < J(\epsilon)$. Applying the thresholding oracle inequality (8) to the first n co-ordinates,

$$r_\epsilon(\hat{\theta}^T, \theta) \leq c \cdot \log\epsilon^{-2} \cdot [\epsilon^2 + \mathcal{R}_\epsilon(\theta, DP)] + \sum_{j \geq J(\epsilon)} |\theta_j|^2 \tag{9}$$

In contrast with the scale $\mathcal{S}_2$ of Section 2.1, consider now a broader scale of Sobolev-type parameter spaces: $\mathcal{S} = \{\Theta_{\alpha,p}(C) : \alpha > 1/p - 1/2, p > 0, C > 0\}$. For such spaces there is a bound relating ideal to minimax risk. First, the geometric weights in the definition imply ([12]) that for $\Theta = \Theta_{\alpha,p}(C)$ and on setting $r = 2\alpha/(2\alpha+1)$,

$$\mathcal{R}_\epsilon(\Theta, DP) := \sup_\Theta \mathcal{R}_\epsilon(\theta, DP) = \sup_{\theta\in\Theta} \sum \theta_I^2 \wedge \epsilon^2 \leq c_\alpha C^{2(1-r)}\epsilon^{2r}.$$

Second, the minimax risk over Θ is minorized by that over any inscribed hypercube of dimension m and side length ϵ : $R_N(\Theta, \epsilon) \geq c_0 m\epsilon^2$. Optimizing over the dimension m and combining with the previous display, we obtain the basic *ideal to minimax* risk inequality:

$$\mathcal{R}_\epsilon(\Theta, DP) \leq c_\alpha C^{2(1-r)} \epsilon^{2r} \leq c'_\alpha R_N(\Theta, \epsilon). \tag{10}$$

In combination with the oracle inequality and negligibility of the tail sum in (9), this yields an adaptive *near*-minimaxity property for thresholding:

Theorem 7. *For all* $\Theta_{\alpha,p}(C) \in \mathcal{S}$,

$$\sup_{\Theta_{\alpha,p}(C)} r_\epsilon(\hat\theta^T, \theta) \leq c_{\alpha,p} \cdot \log \epsilon^{-2} \cdot R_N(\Theta_{\alpha,p}(C), \epsilon).$$

The term near-minimaxity refers to the logarithmic term in the upper bound, which is negligible with respect to the algebraic rate ϵ^{2r}. In fact, this logarithmic term can also be removed by a lower, data-dependent choice of threshold [7, 18].

Important here is that in contrast to the linear adaptivity of Theorem 4, this result applies for all $p > 0$, and in particular for $p < 2$. These latter spaces contain spatially inhomogeneous functions with localized discontinuities or other singularities. The ability of an estimator to adapt to such functions is in practice more important than the attractive, but limited adaptation of the levelwise James-Stein estimator, and its cousins, the spatially homogeneous kernel methods, even with bandwidth selected from data. This is discussed further in [11].

4 Redundant Dictionaries & Complexity Penalized Model Selection

In seeking a sparse representation for a signal, one may build build rich dictionaries $\mathcal{D} = \{x_1, \ldots, x_p\}$ in various ways: for example by combining many orthonormal bases (as in libraries of wavelet and cosine packets, [4]), or by considering redundant discretizations of continuously parametrized families, or by allowing products (interactions) of many simple elements, such as B-splines with knots at individual data locations (e.g. [16]). In all these cases, the dictionary size p greatly exceeds that data size n, and estimation methods will have to allow for the effects of searching over such a vast domain (in principle, 2^p models).

Recalling Examples 1(c) and 2(b), the data may be represented in the form $y = X\beta + \epsilon z$, where we now assume that $\text{span}(X) = \mathbb{R}^n$. Thus, the models of interest correspond to subsets $J \subset \{1, \ldots, p\}$, $M_J = \text{span}\,\{x_j : j \in J\}$, and $\hat\mu_J = P_J y$, orthogonal projection on M_J. The risk of individual projection estimators is given by (3), so the ideal risk of subset selection from dictionary $\mathcal{D}$ becomes

$$\mathcal{R}_\epsilon(\mu, SS(\mathcal{D})) = \min_J r_\epsilon(\hat\mu_J, \mu) = \min_J |\mu - P_J\mu|^2 + \epsilon^2 \text{rank}(P_J).$$

To obtain an estimator that mimicks ideal risk, we use the penalized least squares principle. This balances the fit of the estimate, which in the absence of any penalty could be made arbitrarily close to the data, against some measure of complexity of the estimate:

$$\hat\mu_P = \text{argmin}_{\tilde\mu} |y - \tilde\mu|^2 + \epsilon^2 P(\tilde\mu).$$

In the orthonormal basis setting, $\hat{\mu}_P$ can be evaluated explicitly when the penalty P has an additive form: for example, $P(\mu) = c \sum \mu_i^2$ implies linear shrinkage, $P(\mu) = 2\lambda \sum |\mu_i|$ implies soft thresholding, and $P(\mu) = \lambda^2 \sum I\{\mu_i \neq 0\}$ implies $\hat{\mu}_{P,i}(y) = y_i I\{|y_i| > \lambda\}$, or hard thresholding. For the *redundant linear model* $y = X\beta + \epsilon z$, we generalize the third case by setting

$$P(\mu) = \lambda^2 N(\mu), \qquad N(\mu) = \min\{|J| : \mu = \sum_{j \in J} \beta_j x_j\}.$$

The resulting penalized least squares estimator may expressed in terms of the residual sums of squares $RSS_J = |y - \hat{\mu}_J|^2$ of the possible models:

$$\min_{\tilde{\mu}} |y - \tilde{\mu}|^2 + \lambda^2 \epsilon^2 N(\tilde{\mu}) = \min_J RSS_J + \lambda^2 \epsilon^2 \mathrm{rank}(P_J).$$

Hence we call this the *Complexity Penalized Residual Sum of Squares* (CPRSS) estimate. Certain choices of the factor λ lead to well known estimators: $\lambda^2 = 2 \ (AIC), \ \log p \ (BIC), \ 2\log n \ (RIC)$ (For details and references see [8]).

THEOREM 8 ([8]). *Let $\zeta > 1, \beta > 0$ and $\lambda = \lambda_p = \zeta[1 + \sqrt{2(1+\beta)\log(p+1)}]$. Then for all $n, p \geq n$, and $\mu \in \mathbb{R}^n$,*

$$r_\epsilon(\hat{\mu}_{CPRSS}, \mu) \leq L_p[(2 + \gamma_p)\epsilon^2 + \mathcal{R}_\epsilon(\mu, SS(\mathcal{D}))], \tag{11}$$

where $L_p = (1 - \zeta^{-1})^{-1}\lambda_p^2$, and $\gamma_p = \gamma(p, \beta) \to 0$ as $p \to \infty$.

The penalty factor λ_p^2 is slightly larger than $2 \log p$, where p is the cardinality of the dictionary. We emphasize that the result holds for all μ, n and $p \geq n$, and in particular the inequality depends only on p, not n! Building on the remarkable [1], Birgé & Massart are conducting a thorough study of penalties $P(\mu)$ for which such oracle inequalities and improvements hold. While the constant L_p in (11) is certainly not optimal, there is a lower bound similar to Theorem 6:

THEOREM 9 ([8]). *For each fixed $r \in \mathbb{N}$, there exists a sequence of dictionaries $\mathcal{D}_n$ with $p(n) = |\mathcal{D}_n| \asymp n^r$ such that as $n \to \infty$,*

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{R}^n} \frac{E|\hat{\mu} - \mu|^2}{\epsilon^2 + \mathcal{R}_\epsilon(\mu, SS(\mathcal{D}))} \geq [2 \log p(n)](1 + o(1)).$$

ROLE OF CONCENTRATION INEQUALITIES. The stochastic part of the proof of Theorem 8 depends on an early example (due to Cirelson-Ibragimov-Sudakov [2, 21]) of what are now in probability called concentration (or deviation) inequalities. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitz with $\|f\|_{Lip} = L$. If $Z \sim N_n(0, I)$, then

$$P\{f(Z) \geq Ef(Z) + t\} \leq \exp\{-t^2/2L^2\}.$$

The key points are the Gaussian tail behaviour of $f(Z)$ and the fact that it does not depend on dimension n - hence the dimension-free aspect of Theorem 8. This inequality can then be applied to all projections onto model subsets of cardinality $|J| = \ell$, and then summed over ℓ. Thus, since $f(z) = \|P_J z\|$ has $\|f\|_{Lip} = 1$, and since $Ef(Z) \leq \sqrt{\ell}$, we have, on setting $t = \sqrt{2\ell(1+\beta)\log p}$,

$$P\{\sup_{|J|=\ell} \|P_J z\| \geq \sqrt{\ell} + t\} \leq \binom{p}{\ell} p^{-\ell(1+\beta)} \leq \frac{1}{p^{\ell\beta}\ell!}.$$

4.1 Illustration: Minimaxity for non-standard function classes

The penalized least squares formalism can be applied in situations where no unconditional basis exists. To give a simple example, consider again the model $dY_t = f dt + \epsilon dW_t$, where now $t \in [0,1]^2$, and the *horizon model* for edges in images, studied earlier by, for example, Korostelev and Tsybakov [20]. It is supposed that f takes only the values 0 and 1, and further that the boundary is such that $f(t_1, t_2) = I\{t_2 \le \theta(t_1)\}$. The boundary, or *horizon*, is supposed to be Hölder continuous: more specifically, we say that $f \in \text{HÖLDER}_s(B)$ if $\|\theta\|_\infty + \|\theta^{(r)}\|_\beta \le B$, where $r \in \mathbb{N}$, $\beta = s - r \in (0,1]$ and $\|g\|_\beta = \sup |g(t) - g(t')|/|t-t'|^\beta. \ldots$]

Dictionaries and minimax risk. While $\mathcal{D}$ is often conceptually infinite, in practice one must work with a family of finite subdictionaries $\mathcal{D}_\epsilon$ with cardinality $m(\epsilon)$ being at most a polynomial function of $\epsilon^{-2} : m(\epsilon) \le \beta_1 \epsilon^{-2\beta_2}$. [8] defines a notion of *universal dictionary* for a scale $\mathcal{S} = \{\mathcal{F}_\nu(C)\}offunctionclasses$, which has as consequence the same type of *ideal to minimax risk inequality* as used in the orthobasis case (compare (10)): for all $\mathcal{F}_\nu(C) \subset \mathcal{S}$ and $\epsilon < \epsilon(\nu, C)$, there exists $r = r(\nu)$ such that

$$\mathcal{R}_\epsilon(\mathcal{F}_\nu(C), \mathcal{D}_\epsilon) \le K_\nu C^{2(1-r)} \epsilon^{2r} \le K'_\nu R_N(\mathcal{F}_\nu(C), \epsilon).$$

This may then be combined with the oracle inequality of Theorem 8 to obtain adaptive near-minimaxity.

Thus, in the horizon example, we start with a continuum *trapezoid* dictionary, parametrized by $\gamma = (a,b,c,d)$, representing a function taking value 1 on the trapezoid in $[0,1]^2$ with abscissae $a < b$ and corresponding ordinates c, d. Thus $\mathcal{D}_{Trap} = \{T_\gamma : \gamma \in [0,1]^4, b \ge a\}$. To obtain finite subdictionaries, discretize the unit interval into $I_N = \{i/N : 0 \le i \le N\}$ and set $\mathcal{D}_N = \{T_\gamma : \gamma \in I_N^2 \times I_{N^2}^2\}$. Choose $N(\epsilon) = \epsilon^{-2}$, and set $\mathcal{D}_\epsilon = \mathcal{D}_{N(\epsilon)}$. It can be verified [8] that $\mathcal{D}_{Trap}$ is universal for $\mathcal{S} = \{\text{HÖLDER}_s(B) : 0 < s \le 2, 0 < B\}$, with $\nu = s/2, C = B^{1/2}$.

Corollary 10 ([8]). *On* $\text{HÖLDER}_s(B)$*, for* $0 < s \le 2$*, and setting* $r = s/(s+1)$*,*

$$r_\epsilon(\hat{f}_{CPRSS}, f) \le c_0 \cdot \log_2 \epsilon^{-2} \cdot B^{1-r} \epsilon^{2r}.$$

A key remark is that this adaptively (near minimax) rate of convergence is better than the rate attainable using a two dimensional tensor product wavelet basis when $s > 1$.

Nevertheless, a serious practical defect of Theorem 8 is the combinatorial search implicit in the definition of $\hat{\mu}_{CPRSS}$. The development of fast algorithms suitable for specific cases is an active direction of current research [5, 13].

Acknowledgements The results described here were obtained in collaboration with David Donoho, and some also in joint work with Gérard Kerkyacharian, Dominique Picard, and Bernard Silverman. Work on this manuscript and talk was supported in part by NSF, NIH and the Guggenheim Foundation.

References

[1] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Its Applications*, 00:in press, 1999.

[2] B.S. Cirel'son, I.A. Ibragimov, and V.N. Sudakov. Norm of gaussian sample function. In *Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory*, Lecture Notes in Mathematics, 550, pages 20–41, 1976.

[3] A. Cohen, I. Daubechies, and P. Vial. Wavelets and fast wavelet transform on an interval. *Applied Computational and Harmonic Analysis*, 1:54–81, 1993.

[4] R.R. Coifman, Y. Meyer, and M.V. Wickerhauser. Wavelet analysis and signal processing. In B. Ruskai et. al., editor, *Wavelets and their Applications*, pages 153–178. Jones and Bartlett, 1992.

[5] R.R. Coifman and M.V. Wickerhauser. Entropy-based algorithms for best-basis selection. *I.E.E.E. Transactions on Information Theory*, 38:713–718, 1992.

[6] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[7] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90:1200–1224, 1995.

[8] D. L. Donoho and I. M. Johnstone. Empirical atomic decomposition. Technical report, Stanford University, 1995.

[9] D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 1998. To appear.

[10] D. L. Donoho and I. M. Johnstone. Asymptotic minimaxity of wavelet estimators with sampled data. *Statistica Sinica*, 00:in press, 1999.

[11] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society, Series B*, 57:301–369, 1995. With Discussion.

[12] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Universal near minimaxity of wavelet shrinkage. In Pollard D., Torgersen E., and Yang G.L., editors, *Festschrift for L. Le Cam*, pages 183–218. Springer Verlag, 1997.

[13] D.L. Donoho. Wedgelets: Nearly minimax estimation of edges. Technical report, Dept. of Statistics, Stanford University, 1997.

[14] S.Yu. Efroimovich and M.S. Pinsker. A learning algorithm for nonparametric filtering. *Automat. i Telemeh.*, 11:58–65, 1984. (in Russian).

[15] M. Frazier, B. Jawerth, and G. Weiss. *Littlewood-Paley Theory and the study of function spaces.* NSF-CBMS Regional Conf. Ser in Mathematics, 79. American Mathematical Society, Providence, RI, 1991.

[16] J.H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67, 1991. (with discussion).

[17] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability Theory*, pages 361–380. University of California Press, 1961.

[18] I. M. Johnstone. Wavelet shrinkage for correlated data and inverse problems: adaptivity results. *Statistica Sinica*, 00:in press, 1999.

[19] I. M. Johnstone and B. W. Silverman. Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society, Series B.*, 59:319–351, 1997.

[20] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction: Lecture Notes in Mathematics.* Springer Verlag: New York, 1993.

[21] M. Ledoux. Isoperimetry and gaussian analysis. In P. Bernard, editor, *Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilities de Saint Flour, 1994.* Springer Verlag, 1996.

[22] P.G. Lemarié and Y. Meyer. Ondelettes et bases Hilbertiennes. *Revista Matematica Iberoamericana*, 2:1–18, 1986.

[23] M.S. Pinsker. Optimal filtering of square integrable signals in gaussian white noise. *Problems of Information Transmission*, 16:120–133, 1980. originally in Russian in *Problemy Peredatsii Informatsii* 16 52-68.

[24] C. Stein. Efficient nonparametric estimation and testing. In *Proc. Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pages 187–195. University of California Press, Berkeley, CA., 1956.

[25] Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9:1135–1151, 1981.

Iain M. Johnstone
Department of Statistics
Stanford University
Stanford CA 94305 U.S.A.