Acta Numerica (2006), pp. 1–69
DOI: 10.1017/S0962492904000xxx

Modern statistical estimation via oracle inequalities

Emmanuel J. Candès
Applied and Computational Mathematics,
California Institute of Technology,
Pasadena, CA 91125, USA
E-mail: emmanuel@acm.caltech.edu

A number of fundamental results in modern statistical theory involve thresholding estimators. This survey paper aims at reconstructing the history of how thresholding rules came to be popular in statistics and describing, in a not overly technical way, the domain of their application. Two notions play a fundamental role in our narrative: sparsity and oracle inequalities. Sparsity is a property of the object to estimate, which seems to be characteristic of many modern problems, in statistics as well as applied mathematics and theoretical computer science, to name a few. 'Oracle inequalities' are a powerful decision-theoretic tool which has served to understand the optimality of thresholding rules, but which has many other potential applications, some of which we will discuss.

Our story is also the story of the dialogue between statistics and applied harmonic analysis. Starting with the work of Wiener, we will see that certain representations emerge as being optimal for estimation. A leitmotif throughout our exposition is that efficient representations lead to efficient estimation.

CONTENTS

1. Introduction

1.1. Foreword

This paper is a survey article based on a series of lectures I gave at the Institute of Mathematical Sciences at the National University of Singapore in August 2004. The theme of these lectures was the interactions between applied harmonic analysis and statistical estimation. I feel that it is important to state upfront that these lectures were by no means conceived as an extended review of recent developments in the theory and practice of nonparametric estimation but merely as an account of some important ideas I had learned as a PhD student in the Department of Statistics at Stanford University during the years 1995–1998. More to the point, these lectures owe much to the scientific vision proposed by David Donoho and his colleagues in a series of papers published in the early and mid-1990s, which have influenced my thinking enormously, and continue to do so. I would also like to acknowledge inspiration from a course I took called 'Function Estimation in White Noise' taught by Iain Johnstone, and from a set of notes written for this course, which have been updated since then, namely, Johnstone (2002) in the reference section. This paper makes repeated references to Johnstone's unpublished manuscript, as the latter deals with many of the topics we discuss here. I might have achieved something, should this paper merely serve the purpose of encouraging the curious reader to take a look at Donoho's papers and Johnstone's manuscript.

1.2. Interactions between statistical estimation and harmonic analysis

The interactions between harmonic analysis and statistical estimation have, of course, a long history. Although it is amusing to note that Joseph Fourier, the founding father of harmonic analysis, spent a significant fraction of his research career studying statistical problems (see Stigler (1990) for an excellent account of Fourier's contribution to early statistics), this history cannot be traced quite that far back. Instead, the credit for bringing both these topics together should probably go to Norbert Wiener where our story begins. In the late 1930s and early 1940s, Wiener studied the problem of filtering out noise (by statistical means) that has corrupted a time series. He developed a solution by requiring information regarding the spectral content of the original signal and the noise, and by creating a filter, which, for stationary signals, filters selected frequencies. This filter was proposed in the 1940s and first published in Wiener (1949). Since this fundamental contribution, Fourier analysis has always played an important role in the filtering literature and, more generally, in the analysis of time series.

Harmonic analysis and statistical estimation also remained connected via the theory of splines (Wahba 1990), via the theory of estimation in statistical inverse problems and via key theoretical developments in function estimation in the white noise model, to name a few examples. Having said that, it is nevertheless fair to say that the subject has been completely revitalized by Donoho and his colleagues. In the early 1990s, Donoho and his team realized that recent advances in applied harmonic analysis such as the theory of wavelets had very significant implications for statistical estimation. They developed *wavelet shrinkage* and established many of its spectacular properties, showing that, perhaps surprisingly, this algorithm has universal properties in the sense that it solves many statistical estimation problems simultaneously. I am sure that everyone reading this paper has heard about wavelet shrinkage as this has almost become a household word, and is perhaps the greatest application of wavelets to this date. But beyond wavelet shrinkage, Donoho also showed that efficient representations lead to efficient estimations, and that certain representations emerge as optimal. In doing so, he has linked statistical estimation and harmonic analysis in a durable and profound way. There is something remarkable about the timeliness of this discovery, since it occurred during a period marked and followed by intense research in computational harmonic analysis. On the one hand, applied mathematicians were energized by the prospect of new applications for the tools they were constructing, and on the other hand, statisticians had access to a brand new and powerful toolbox to refine and extend Donoho's ideas.

1.3. Our preoccupations

Such a broad subject imposes a selection of topics that will be covered and others that will not. As emphasized earlier, we will focus on ideas that have shaped my thinking; our focus is on the key structures and tools that bind statistical estimation and harmonic analysis. For example, we will explore the consequences of sparsity and emphasize the key role played by oracle inequalities – a new, fruitful and enlightening concept with an almost unlimited range of applications.

Our focus on sparsity and oracle inequalities serves a simple purpose: we wish to provide the reader with the necessary ideas for understanding an important fraction of the literature on modern statistical estimation, and with tools for future research in this area. Our point of view is that both these notions are fundamental, and that many decision-theoretic results are, in fact, easy consequences from rather simple oracle inequalities. To make this point, the reader perhaps already knows that wavelet shrinkage, discussed above, is asymptotically optimal for recovering objects taken from certain functional classes, such as the so-called Besov spaces – a result which has attracted a lot of attention. In truth, this is an automatic consequence of the fact that (1) wavelets provide optimally sparse representations of such functional classes, and that (2) a fundamental oracle inequality relates the performance of thresholding rules to the sparsity of such wavelet representations. Although there exist other ways to comprehend these types of results – note that we are not saying that these alternatives are uninteresting – we have decided to shift focus away from these and, instead, discuss what we believe are more fundamental concepts.

Indeed, the concepts of sparsity and oracle inequalities have already had a significant impact and everything suggests that this impact will last for a very long time. For example, 'sparsity' has become a true paradigm in many fields (not only statistics) including applied mathematics, theoretical computer science, signal and image processing, inverse problems, scientific computing and so on. While the potential for sparsity has been understood for a while now, there were relatively very few papers on this subject twenty years ago. In contrast, it is startling to see that the number of research papers and talks with 'sparsity' as a central theme has been exploding over the last few years. An oracle inequality, on the other hand, is a decision-theoretic tool and its use has thus far been confined to the field of statistical estimation. There are many forms of oracle inequalities and, as we will see, they have proved extremely successful in addressing the performance of many new estimation strategies 'post-wavelet shrinkage'. Without a doubt, oracle inequalities will continue to play a vital role in years to come.

1.4. Organization of the paper

We begin our survey with early important ideas in linear estimation, which are presented in Section 2. What is interesting here is that these ideas make explicit the connection between the estimation problem and the representation problem (the subject of applied harmonic analysis). Section 3 motivates the need for nonlinear estimation procedures. Section 4 introduces nonlinear estimation (nonlinear shrinkage to be more exact) and the powerful concept of oracle inequality. Section 5 introduces the notion of sparsity and shows that thresholding rules are very accurate for estimating sparse objects, *e.g.*, parameter vectors with only a few significant entries. Section 6 argues that the problems of efficiently estimating, approximating, or compressing a signal (or a function) are all related and all linked to the fundamental problem of finding efficient signal representations. In Section 7, we consider extensions of thresholding ideas when there is no orthobasis (*i.e.*, orthonormal basis) in which the object is sparse. Section 8 revisits some topics in model selection and introduces the Dantzig selector, a new effective and computationally tractable estimation strategy for estimating signals from undersampled data. Section 9 explores the possibility of adaptive basis estimation. Finally, we close the paper by discussing further topics, essentially inverse problems and false discovery rate thresholding rules in Section 10.

Because the intended audience is wide-ranging, we also include a Glossary, on page 68, where the reader will find definitions or explanations of the main statistical terms or concepts. The words or expressions to be found in the Glossary are marked by a superscript star, *.

2. Linear estimation

2.1. The Wiener filter

We start with a classical estimation problem known as 'Wiener filtering' in the electrical engineering literature. This example is primarily of historical significance and the author would otherwise have been guilty of omission. But more importantly, this example will play a pedagogical purpose as it wonderfully introduces some of the key ideas surveyed in this paper.

We wish to recover a Gaussian signal* $X = (X_1, X_2, \ldots, X_n)$ from noisy data Y of the form

$$Y_t = X_t + Z_t, \quad t = 1, \ldots, n; \tag{2.1}$$

here, Y is the observed process, X is the signal, which is assumed to be a Gaussian process with mean zero and covariance matrix Σ, *i.e.*, $X \sim N(0, \Sigma)$, and Z is Gaussian white noise, *i.e.*, $Z \sim N(0, \sigma^2 I)$, and independent of the signal X. One may want to view this as a Bayesian estimation* problem where the prior on the unknown signal is Gaussian. The goal is to reconstruct the signal by producing an estimator $\hat{X} = g(Y)$ which can be computed from the data, and which has small mean-squared error

$$\mathrm{MSE}(X, \hat{X}) = \mathbb{E}\|X - \hat{X}\|_2^2 = \mathbb{E}\sum_{t=1}^{n}(X_t - \hat{X}_t)^2. \tag{2.2}$$

As is well known in Bayesian statistics (*e.g.*, see Lehmann (1997)), the estimator which achieves the minimum MSE is the conditional expectation of X given the observed process Y:

$$\hat{X} = \mathbb{E}(X \mid Y). \tag{2.3}$$

In detail, the tth component is given by

$$\hat{X}_t = \int_{\mathbb{R}^n} z_t \, p_{X|Y}(z) \, \mathrm{d}z,$$

where $p_{X|Y}$ is the conditional density of the random vector X. At first glance, the analytical evaluation of the conditional expectation might seem a little delicate. Having said that, a detour by way of principal components greatly simplifies things.

Recall that the principal components of a process $(X_t)_{1\le t\le n}$ are the orthonormal eigenvectors φ_k, $1 \le k \le n$, which diagonalize the covariance matrix

Σ of X. In matrix notation, the matrix of principal components Φ is the n by n orthonormal matrix obeying

$$\Sigma = \Phi D \Phi^T, \quad D = \operatorname{diag}(d_k^2). \tag{2.4}$$

We will assume that the eigenvalues are arranged in decreasing order of magnitude $d_1^2 \geq d_2^2 \geq \cdots \geq d_n^2$. (We use the notation d_k^2 to emphasize that the eigenvalues of Σ are nonnegative since Σ is positive semidefinite.) The interpretation is that, if X is Gaussian, then the level sets of the joint density of the vector X are concentric ellipsoids, and the principal components are simply the (normalized) principal axes of these ellipsoids. A more general interpretation, which holds for general stochastic processes (not necessarily Gaussian), is that the first principal component is a projection with maximal variance; φ_1 is a unit vector obeying

$$\operatorname{Var}(u^T X) \leq \operatorname{Var}(\varphi_1^T X), \quad \text{for all } \ u \in \mathbb{R}^n : \|u\| = 1.$$

The second principal component φ_2 is then a projection with maximal variance among all projections orthogonal to φ_1

$$\operatorname{Var}(u^T X) \leq \operatorname{Var}(\varphi_2^T X), \quad \text{for all } \ u \in \mathbb{R}^n : \|u\| = 1, \, u \perp \varphi_1,$$

and so on for $\varphi_3, \varphi_4, \ldots, \varphi_n$.

With this in mind, principal component analysis is the action of decomposing a process X as a superposition of its principal components. It consists of two steps.

(1) The analysis step finds the orthonormal eigenvectors φ_k and projects X onto this basis, *i.e.*,

$$X' = \Phi^T X.$$

(2) The synthesis step reconstructs the process from the principal components using the orthonormal eigenvectors by $X = \Phi X'$, *i.e.*,

$$X_t = \sum_{k=1}^{n} X_k' \varphi_k(t). \tag{2.5}$$

This formula is also known as the Karhunen–Loeve decomposition: see Leon-Garcia (1994).

By definition, the coefficients X_k' in the expansion (2.5) are uncorrelated – the covariance matrix of X' is the diagonal matrix D – and are therefore also independent in the case where X is Gaussian since $X' \sim N(0, D)$. Hence, the Karhunen–Loeve decomposition provides a representation of Gaussian stochastic processes as a superposition of *independent* components.

We now return to the estimation problem and 'rotate' the observation vector Y in the orthonormal basis of principal components by applying Φ^T

on both sides of (2.1)

$$\begin{aligned}\langle Y, \phi_k \rangle &= \langle X, \phi_k \rangle + \langle Z, \phi_k \rangle, \\ Y'_k &= X'_k + Z'_k.\end{aligned}$$

The coordinates $X'_k \sim N(0, d_k^2)$ are independent; the Z'_k are i.i.d.* $N(0, \sigma^2)$ and independent of X'. Obviously, the problem has not changed and we are merely looking at it from a different perspective . In particular, to estimate X, we may just as well estimate its coefficient sequence X' with $\widehat{X'}$: that is, with any estimator with minimum mean-squared error. The synthesis step would then provide the reconstruction $\hat{X} = \Phi\widehat{X'}$,

$$\hat{X}_t = \sum_{k=1}^{n} \widehat{X'_k}\, \varphi_k(t),$$

and owing to the isometry

$$\|X - \hat{X}\|^2 = \|X' - \widehat{X'}\|^2,$$

this would be exactly the estimator with minimum MSE: $\hat{X} = \mathbb{E}(X \mid Y)$. The point of all this is that $\widehat{X'}$ is now easy to compute since

$$\widehat{X'_k} = \mathbb{E}(X'_k \mid Y') = \mathbb{E}(X'_k \mid Y'_k),$$

where the second equality uses the fact that X'_k is independent of all the components Y'_j with $j \neq k$. Now the pair (X'_k, Y'_k) follows a bivariate normal distribution with mean zero and covariance matrix $\mathrm{Var}(X'_k) = d_k^2 = \mathrm{Cov}(X'_K, Y'_k)$, and $\mathrm{Var}(Y'_k) = d_k^2 + \sigma^2$. It is a classical exercise in regression analysis to show that the conditional distribution of X'_k is Gaussian with conditional mean

$$\mathbb{E}(X'_k \mid Y'_k) = \frac{\lambda_k^2}{\lambda_k^2 + \sigma^2} Y'_k, \tag{2.6}$$

so that the Wiener estimator is given by

$$\hat{X}_t = \sum_{k=1}^{n} w_k \langle Y, \phi_k \rangle \varphi_k(t), \quad w_k = \frac{\lambda_k^2}{\lambda_k^2 + \sigma^2}. \tag{2.7}$$

In short, the Wiener filter transforms the data with respect to the orthobasis of principal components, and downweights each coefficient as a function of the signal-to-noise ratio since one can think of the coordinates of w as the ratio between the expected signal power and the expected signal + noise power. Note that downweighting and the whole estimation procedure are linear, and that one can write $\hat{X}$ as

$$\hat{X} = \Phi W \Phi^T Y,$$

where $W = \mathrm{diag}(w_k)$.

It is interesting to consider special instances of Wiener filtering. Suppose for example that the process process X is stationary (and periodic) in the sense that the covariance between X_s and X_t only depends on the time lag

$$\Sigma_{s,t} = \mathrm{Cov}(X_s, X_t) = \gamma(s-t), \quad 1 \le s, t \le n,$$

where it is understood that subtraction operates modulo n. This property says that the statistical properties of the signal are invariant with respect to time shifts, which conveys the idea that the process is spatially homogeneous. Because Σ is a circulant matrix, the basis of principal components is the Fourier basis which, for even sample sizes n, takes the form

$$\begin{aligned} \varphi_1(t) &= 1/\sqrt{n}, \\ \varphi_{2k}(t) &= \sqrt{2/n}\cos(2\pi kt/n), \quad k = 1, 2, \ldots, n/2-1, \\ \phi_{2k+1}(t) &= \sqrt{2/n}\sin(2\pi kt/n), \quad k = 1, 2, \ldots, n/2-1, \\ \varphi_n(t) &= (-1)^t/\sqrt{n}, \end{aligned}$$

and the eigenvalues are the Fourier coefficients of the vector $(\gamma(0), \ldots, \gamma(n-1))$. Hence, Bayes' rule or Wiener's solution exhibit the following key structure:

(1) Bayes' rule transforms the data in the frequency domain,

(2) Bayes' rule shrinks the noisy Fourier coefficients towards zero using a specially selected frequency-dependent factor,

(3) finally, Bayes' rule reconstructs the signal by inverting the Fourier transform.

As we shall see, this transformation–shrinkage–inverse transformation structure is a recurrent theme in modern statistical estimation. What is interesting here is that the estimation problem makes no reference to any particular basis, nor to any particular shrinkage rule, and yet this structure naturally emerges as the optimal strategy.

In conclusion, the Wiener filter is optimal for Gaussian signal priors. In the case where X is non-Gaussian, however, the estimator (2.7) is only guaranteed to have minimum mean-squared error among all *linear* estimators; see Leon-Garcia (1994).

2.2. Kernel methods

In contemporary nonparametric statistics, there are other models which do not assume a prior distribution on the signals or functions of interest. The so-called frequentist viewpoint assumes a model of the form

$$y_i = f(t_i) + z_i, \quad 1 \le i \le n, \tag{2.8}$$

where again y is a vector of observations, the function $f(t)$ is the object we wish to recover, and z is a vector of stochastic and independent errors. In nonparametrics, the object f is completely unknown and does not depend upon a few parameters. The goal is to estimate f from the data y. Note that to develop a fruitful methodology, one would need to restrict the classes of objects f of interest, since to extract the object, one would need to be able to distinguish it from noise. Examples of common assumptions include imposing a bounded total variation, a bounded curvature, or bounded higher-order derivatives.

One of the first developed and most frequently discussed approaches for estimating the regression function f is the kernel method: see Silverman (1986) and Scott (1992) for an introduction. The idea is to estimate the response $f(t)$ by a local averaging of the data y_i with 'time indices near' the point t under consideration. To do this, one selects a kernel K, usually a symmetric density function, which is nonnegative and integrates up to one. Typical examples include the boxcar kernel $K(t) = 1$ if $-1/2 \le t \le 1/2$ and zero otherwise, the Gaussian kernel $K(t) = (2\pi)^{-1/2}e^{-t^2/2}$, and the 'spline' kernel or Epanechnikov kernel equal to $\frac{3}{4}(1-t^2)_+$, where here and below x_+ is the positive part of the scalar x. With such a kernel, the kernel regression sets

$$\hat{f}(t) = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}, \tag{2.9}$$

where the weights are given by the formula

$$w_i = K(h^{-1}(t - t_i)). \tag{2.10}$$

Hence, the estimator is a weighted average and closer points naturally receive larger weights since typical kernels $K(t)$ decay as $|t|$ increases. The parameter h is the window width, or the bandwidth, and essentially determines which observations are averaged together. A small bandwidth averages over very few points, while a very large bandwidth may average over a significant fraction of the data set.

To connect kernel regression with our earlier discussion, suppose that the t_is are equispaced in $[0,1]$, *e.g.*, $t_i = i/n$ with $1 \le i \le n$ and that the estimand $f(t)$ is periodic. These assumptions are only useful for getting simple results. In the equispaced design, the Priestley–Chao kernel smoother is of the form

$$\hat{f}(t) = \frac{1}{nh}\sum_{i=1}^n K(h^{-1}(t - t_i))y_i, \tag{2.11}$$

where the subtraction is understood modulo $[0,1]$. The estimator is then a convolution in the time domain or, equivalently, a multiplication in the Fourier domain. Let $(w_k(h))_{k\in\mathbb{Z}}$ be the sequence of Fourier coefficients of the density $h^{-1}K(\cdot/h)$

$$w_k(h) = \int_0^1 h^{-1}K(h^{-1}t)e^{-i2\pi kt}\,\mathrm{d}t$$

and let $(\tilde{y}_k)_{k\in\mathbb{Z}}$ be those of the vector y

$$\tilde{y}_k = \int_0^1 n^{-1}\sum_{i=1}^n y_i\delta(t-t_i)\,e^{-i2\pi kt}\,\mathrm{d}t = \frac{1}{n}\sum_{j=1}^n e^{-i2\pi kt_j}y_j$$

(note that $\tilde{y}_{k+n} = \tilde{y}_k$). We also denote the coefficient sequence of f by $(\theta_k)_{k\in Z}$. In the frequency domain, the estimator (2.11) obeys

$$\hat{\theta}_k = w_k(h)\cdot\tilde{y}_k, \tag{2.12}$$

where we observe that $0\le|w_k(h)|\le 1$. In short, the kernel method estimates the Fourier coefficients of f by shrinking those of the observations y, and hence the structure of this procedure is similar to that of the Wiener filter: the estimation combines the transformation of the data in the Fourier domain with frequency-by-frequency dumping. If W is the Fourier transform of K,

$$W(\omega) = \int K(t)e^{-i2\pi\omega t}\,\mathrm{d}t,$$

then $w_k(h)\approx W(kh)$ and $|w_k(h)|$ typically decreases as the frequency index $|k|$ increases. For example, if K is the Gaussian kernel, $W(kh) = e^{-(kh)^2/2}$. The bandwidth h controls the decay of the weights $w_k(h)$; the larger h, the faster the decay and hence the greater the amount of smoothing.

Whereas the Wiener filter gives an explicit formula for the weights, here the sequence $w_k(h)$ depends upon the kernel and above all upon the bandwidth. Automatic selection of the bandwidth h – *i.e.*, how much to smooth – is the topic of an immense literature. There are theoretical rules based on asymptotics which guarantee good MSEs for estimating smooth functions together with more practical rules for finite samples, *e.g.*, based on cross-validation: see Green and Silverman (1994).

2.3. Smoothing splines

Another popular approach for estimating the regression function is based on smoothing splines. The idea is to find an estimator $\hat{f}$ which minimizes the trade-off between the goodness of fit and the complexity of the estimator, as measured by the size of the second derivative of the fitted function. Quantitatively, we wish to find the function $\hat{f}(t)$ which minimizes the variational problem

$$\hat{f} = \operatorname{argmin}_g \sum_{i=1}^n (y_i - g(t_i))^2 + \lambda\int_0^1 |g''(u)|^2\,\mathrm{d}u. \tag{2.13}$$

Like the bandwidth, the parameter $\lambda > 0$ controls the smoothness of the fit. The larger λ, the smoother the fit (in the limit where λ goes to infinity, the regression function is the regression line). It is not difficult to show that the solution $\hat{f}(t)$ to (2.13) is a cubic spline with knots at the sampled points t_i – hence the name of the method. The problem of fitting the data is then a finite-dimensional problem, which can be solved efficiently on a computer.

As before, we wish to develop an understanding of the structure of the solution by making some useful simplifying assumptions. Suppose that the points $t_i = i/n$, $1 \le i \le n$ are equispaced and that the estimand f is periodic. We approximate the second term of (2.13) by finite differences so that one is interested in finding the vector $g \in \mathbb{R}^n$ minimizing

$$\min \sum_{1\le i\le n} (y_i - g_i)^2 + \lambda \sum_{1\le i\le n} |(D^2 g)_i|^2, \tag{2.14}$$

with

$$(D^2 g)_i = \frac{g_{i+1} - 2g_i + g_{i-1}}{n^2}.$$

(Because f is assumed periodic, we set $g_0 = g_n$ in the above formula so that the matrix D^2 is circulant.) Let $\tilde{y}_k$ (resp. $\tilde{g}_k$) be the discrete Fourier coefficients of y (resp. g)

$$\tilde{y}_k = \sum_{1\le i\le n} y_i \phi_k(i/n),$$

where $(\phi_k(t))_{1\le k\le n}$ is the sequence of sines and cosines introduced in Section 2.1. Since D^2 is diagonal with eigenvalues $d_1 = 0$, $d_{2k} = d_{2k+1} = 4n^{-2}\sin^2(\pi k/n)$ for $1 \le k \le n/2 - 1$ and $d_n = 4n^{-2}$, then owing to the Fourier isometry, the minimization problem is equivalent to

$$\min \sum_{1\le k\le n} [(\tilde{y}_k - \tilde{g}_k)^2 + \lambda \cdot d_k^2\, \tilde{g}_k^2]. \tag{2.15}$$

The solution is now readily available; namely, the discrete Fourier coefficients $(\hat{\theta}_k)$ of the fitted vector $\hat{f}(i/n)$ are given by

$$\hat{\theta}_k = \frac{\tilde{y}_k}{1 + \lambda d_k^2}. \tag{2.16}$$

Once again, a familiar structure emerges. Spline smoothing rotates the data in the frequency domain and linearly shrinks the high-frequency components towards zero, *i.e.*,

$$\hat{f}_\lambda := \Phi W_\lambda \Phi^T y, \qquad W_\lambda = \mathrm{diag}((1 + \lambda d_k^2)^{-1}).$$

The larger λ, the greater the shrinkage. A small value of λ does not imply a lot of smoothing and yields a low bias* but a large variance. Conversely, a large value of λ gives a fit with large bias and small variance. An important topic in spline smoothing is then how to select the parameter λ. In other words, how best to trade off between bias and variance.

To understand the trade-off, we examine the mean-squared error of the fit

$$\mathbb{E}\|f - \hat{f}_\lambda\|^2 = \sum_{i=1}^{n} \mathbb{E}(f(t_i) - \hat{f}_\lambda(t_i))^2,$$

where $\|f - f_\lambda\|^2$ is the Euclidean norm $\sum_{i=1}^{n}(f(t_i) - f(t))^2$ and $\hat{f}_\lambda$ is the solution to (2.14); that is, $\hat{f}_\lambda = S_\lambda y$ where we put $S_\lambda = \Phi W_\lambda \Phi^T$ for short. The classical bias variance decomposition gives

$$\mathbb{E}\|f - \hat{f}_\lambda\|^2 = \|f - \mathbb{E}\hat{f}_\lambda\|^2 + \mathbb{E}\|\hat{f}_\lambda - \mathbb{E}\hat{f}_\lambda\|^2;$$

the bias term obeys $f_\lambda - \mathbb{E}\hat{f}_\lambda = (I - S_\lambda)f$ while the 'variance term' is given by

$$\mathbb{E}\|\hat{f}_\lambda - \mathbb{E}\hat{f}_\lambda\|^2 = \mathbb{E}\|S_\lambda z\|^2 = \sigma^2 \cdot \mathrm{Tr}(S_\lambda^T S_\lambda) = \sigma^2 \cdot \sum_k w_k^2(\lambda),$$

where $w_k(\lambda) = (1 + \lambda d_k^2)^{-1}$. The squared bias increases as λ increases whereas the variance decreases so that the optimal value of λ trades off between the source of errors. Suppose that the sequence $(\theta_k)_{1\le k\le n}$ is the discrete Fourier coefficient sequence of $(f(t_i))_{1\le i\le n}$; then the MSE obeys

$$\mathbb{E}\|f - \hat{f}_\lambda\|^2 = \sum_{1\le k\le n} [(1 - w_k(\lambda))^2\theta_k^2 + \sigma^2 w_k^2(\lambda)]. \tag{2.17}$$

The best value of the smoothing parameter is that value λ^* which minimizes the above mean-squared error. Expressed in a different way, an 'omniscient' procedure knowing in advance λ^* would automatically answer the fundamental question: how much to smooth? This information is, of course, not available in practice, and this is why we used the word 'omniscient' to qualify the procedure. In practice, the best one can hope for is to select a smoothing parameter $\hat{\lambda}$ – based on the data – close to the optimal one. An interesting question is then whether it is possible to find $\hat{\lambda}$ such that the performance of the resulting estimator is close to that of the ideal one. As we will see, such issues will form a recurring theme of this paper.

We conclude this short overview of smoothing splines by pointing out that the solution to (2.13) has the exact same structure as that discussed above even in the case where the design points t_i are unequispaced. In short, there is an orthonormal basis $(\varphi_k)_{1\le k\le n}$ known as the Demmler–Reinsch system (Wahba 1990) which – like the discrete Fourier basis – diagonalizes the minimization problem (2.13) so that the solution in that basis is given by

$$\hat{f}(t) = \sum_{1\le k\le n} \hat{\theta}_k \phi_k(t),$$

where the coefficients $\hat{\theta}_k$ are given by the same relation as (2.16) with, of course, slightly different eigenvalues. The Demmler–Reinsch functions are boundary-adapted sinusoidal waveforms.

2.4. Statistical theory

On the theoretical side, there is a large literature showing that if the shrinkage parameters are chosen appropriately, the corresponding linear estimators are, in an asymptotic sense, optimal for recovering objects assumed to belong to certain types of functional classes. These results are perhaps best presented in the so-called 'white noise model', that is,

$$Y(\mathrm{d}t) = f(t)\,\mathrm{d}t + \varepsilon W(\mathrm{d}t), \quad t \in [0,1]. \tag{2.18}$$

Here $W(t)$ denotes a Wiener process (*i.e.*, the primitive of white noise); ε is a noise level; and f is the object to be recovered. Formally, this model says that if we take a finite numbers of projections of the data Y and define

$$y_k := \langle Y, \varphi_k \rangle = \langle f, \varphi_k \rangle + \varepsilon\, z_k, \quad 1 \le k \le n$$

where the $\varphi_k(t)$s are any functions bounded in L_2, then $z = (z_1, \dots, z_n)$ is a Gaussian vector with mean 0 and covariance matrix $\mathrm{Cov}(z_k, z_\ell) = \langle \varphi_k, \varphi_\ell \rangle$, the Gram matrix of the waveforms φ_k. In particular, if the φ_ks are orthogonal, the coordinates of z are independent. This explains why the white noise model should be understood as the large sample limit of the discrete model (2.8) where the errors z_i are i.i.d. $N(0, \sigma^2)$ under the calibration $\varepsilon = \sigma/\sqrt{n}$. To see why this is so, consider averaging (2.18) over intervals of the form $I_i := [(i-1)/n, i/n]$. This gives

$$y_i := n\,\langle Y, 1_{I_i} \rangle = \bar{f}_i + \varepsilon\sqrt{n}\, z_i,$$

where $\bar{f}_i = \mathrm{Ave}_{I_i} f$ and the z_is are i.i.d. $N(0,1)$. For sufficiently nice functions, $\bar{f}_i$ is close to $f(i/n)$ when n is large, which justifies the claim. In summary, the asymptotics in the continuous white noise model as $\varepsilon \to 0$ have similar characteristics to the asymptotics in the discrete model as $n \to \infty$. In fact, although the model is continuous and real data are typically discretely sampled, the asymptotic theory deriving from the white noise model has typically been found to lead directly to comparable asymptotic theory in a sampled data model. We do not wish to elaborate on this point, and refer the reader to Brown and Low (1996), Nussbaum (1996) for general theory, and to Efroĭmovich and Pinsker (1981, 1982), Nussbaum (1996), Donoho and Nussbaum (1990), Donoho and Liu (1991) and Donoho and Johnstone (1999) for examples of translations of optimal solutions in the white noise model to corresponding solutions in the sampled data model. The advantage is that the white noise model is more homogeneous than sampled data models, and since estimation in the white noise model is in general neither easier nor harder than in sampled models, it has proved to be a fruitful theoretical tool.

Decision theory develops a mathematical theory for making decisions in the face of uncertainty. In the theory of estimation, for example, suppose we wish to estimate a function θ on the basis of a sample $Y = (Y_1, \dots, Y_n)$, where the distribution of the Y_is depend on θ. Then, by choosing an estimator $\hat{\theta} = g(Y)$, the decision maker incurs a loss $\ell(\theta, \hat{\theta})$ whose expected value is called the risk* function

$$R(\theta, \hat{\theta}) = \mathbb{E}\ell(\theta, \hat{\theta}).$$

In the set-up of interest here, the parameter f is the unknown regression function and the observations follow the white noise model (2.18). If we take as a loss the L_2-squared error $\ell(f, \hat{f}) = \|f - \hat{f}\|_{L_2}^2$, the risk is the integrated mean-squared error

$$\mathrm{MSE}(f, \hat{f}) = \mathbb{E}\|f - \hat{f}\|_{L_2}^2.$$

Decision theory is concerned with finding good decisions, *i.e.*, decision functions with small risk. Note that the risk depends on f which is not known. Some decisions may be good for certain values of the parameters and poor for others. Consider for instance, two estimators $\hat{f}_i$, $i = 1, 2$, which are constant and equal to f_i. Suppose f_1 and f_2 are wildly different. When the true state of nature is f_1 the first estimator has vanishing risk, but a very large risk when the true state is f_2, and *vice versa* for the second estimator. The two dominating viewpoints for getting around this difficulty are the minimax and Bayesian paradigms.

(1) The minimax* point of view defines a functional class $\mathcal{F}$ and searches for an estimator $\hat{f}$ which exactly or approximately attains the minimax risk (here the minimax mean-squared error):

$$M^*(\varepsilon, \mathcal{F}) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{MSE}(f, \hat{f}).$$

In other words, one is interested in the estimator with minimum worst-case error. The minimax approach puts no restriction on the estimator; all measurable procedures – *i.e.*, all measurable functions of Y – are allowed.

(2) The Bayesian point of view assumes a prior process π about f (so that $\pi(A)$ is the probability that the object f belongs to the set A) and searches for the estimator achieving the minimum average mean-squared error, the so-called Bayes risk

$$B(\pi) = \mathbb{E}_\pi \mathrm{MSE}(f, \hat{f}).$$

Here one averages the MSE over the prior distribution π. This is the viewpoint of the Wiener filter which assumes a Gaussian prior process.

> If one is given a functional class, as in the minimax framework, then a possible approach is to select a prior on $\mathcal{F}$, a probability distribution on the elements $f \in \mathcal{F}$ obeying $\pi(\mathcal{F}) = 1$.

A key result of statistical decision theory is that the minimax risk is lower-bounded by the Bayes risk for any choice of prior π obeying $\pi(\mathcal{F}) = 1$.

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}} \mathrm{MSE}(f, \hat{f}) \geq B(\pi). \tag{2.19}$$

Under mild conditions, a famous result due to Wald proves the existence of prior distributions satisfying inequality (2.19); such distributions are called *least favourable priors*.

A splendid result in the minimax theory of linear estimation is due to Pinsker. We wish to recover an object f which is assumed to lie in a Sobolev ball

$$\mathcal{F} = \{f : \|f\|_{W_2^m} \leq R\},$$

where $\|\cdot\|_{W_2^m}$ is the Sobolev norm

$$\|f\|_{W_2^m}^2 := \int_{[0,1]} |f(t)|^2 + |f^{(m)}(t)|^2 \, \mathrm{d}t, \tag{2.20}$$

in which $f^{(m)}$ is the mth derivative of the function f. In short, the mth derivative of f is assumed to be bounded in an L_2-sense. Pinsker's solution applies linear shrinkage in the Fourier domain, and is given by

$$\hat{f}(t) = \sum_{k \geq 0} w_{k,\varepsilon} \langle Y, \varphi_k \rangle \varphi_k(t). \tag{2.21}$$

Because we are now studying continuous-time models, $(\varphi_k(t))_{k \geq 0}$ is the continuous-time orthonormal Fourier basis of $L_2(0, 1)$

$$\begin{aligned} \phi_0(t) &= 1, \\ \phi_{2k}(t) &= \sqrt{2}\cos(2\pi k t), \quad k \geq 1, \\ \phi_{2k-1}(t) &= \sqrt{2}\sin(2\pi k t), \quad k \geq 1, \end{aligned}$$

and the weights are given by

$$W_{k,\epsilon} = (1 - \lambda k^m)_+;$$

in the above expression, the scalar λ actually depends on ε and R: see (2.24). It is important to take note that the weights depend on the parameters that define the functional class: the degree of smoothness m and the modulus of smoothness R. The result is that $\hat{f}(t)$ is asymptotically minimax.

Theorem 2.1. (Pinsker's theorem) The estimator (2.21) is asymptotically minimax

$$\sup_{\mathcal{F}} \mathrm{MSE}(f, \hat{f}) = M^*(\varepsilon, \mathcal{F})(1 + o(1)),$$

where $o(1)$ is a term tending to zero as ε tends to zero.

To give a geometric interpretation of Pinsker's theorem, introduce the empirical Fourier coefficients

$$\begin{aligned} \langle Y, \varphi_k \rangle &= \langle f, \varphi_k \rangle + \varepsilon \langle W, \varphi_k \rangle, \\ y_k &= \theta_k + \varepsilon z_k. \end{aligned}$$

By the Parseval theorem, the condition imposing a size constraint on the size of the mth derivative is equivalent to a weighted-ℓ_2 size estimate on the Fourier coefficient sequence of f:

$$f \in \mathcal{F} \quad \Leftrightarrow \quad \theta \in \Theta,$$

where Θ is the infinite-dimensional ellipsoid

$$\Theta := \left\{ \theta : \sum_{k \geq 0} (1 + k^{2m})(|\theta_{2k-1}|^2 + |\theta_{2k}|^2) \leq R^2 \right\}. \tag{2.22}$$

The problem is then to recover $\theta \in \Theta$ from the infinite Gaussian sequence model $y \sim N(\theta, \varepsilon^2 I)$. The idea is that for ellipsoids, least favourable priors are essentially Gaussian. Consider a general ellipsoid

$$\Theta(R) := \left\{ \theta : \sum_k a_k^2 \theta_k^2 \leq R^2 \right\}$$

in which $a_k > 0$ tends to infinity as k tends to infinity. Note that in the case of the Sobolev ball, $a_{2k-1} = a_{2k} = k^m$, or $(1 + |k|^{2m})^{1/2}$ to be more exact. The least favourable prior over the ellipsoid nearly has Gaussian independent components given by

$$\theta_k \sim N(0, \tau_k^2), \quad \tau_k^2 = \varepsilon^2 \lambda^{-1} (a_k^{-1} - \lambda)_+, \tag{2.23}$$

where the scalar λ is that appearing in Pinsker's weights. This scalar is chosen as the smallest real number with $\sum_k a_k^2 \tau_k^2 \leq R^2$, *i.e.*, λ is the solution to

$$\varepsilon^2 \lambda^{-1} \sum_k a_k (1 - \lambda a_k)_+ = R^2. \tag{2.24}$$

The careful reader will notice that $\pi(\Theta(R)) < 1$ but it is possible to consider small perturbations of this prior which asymptotically concentrate on $\Theta(R)$. We leave out the details and refer to Johnstone (2002). For Gaussian priors, one can calculate Bayes' rule, which takes the form

$$\hat{\theta}_k := (1 - \lambda a_k)_+ \, y_k.$$

This is none other than Pinsker's estimate with weights $w_k = (1 - \lambda a_k)_+$, and the MSE of this estimator obeys

$$\mathrm{MSE}(\theta, \hat{\theta}) = \sum_k (1 - w_k)^2 \theta_k^2 + w_k^2 \varepsilon^2,$$

which actually simplifies to $\varepsilon^2 \sum_k w_k$.

3. Why nonlinear estimation?

Linear estimation is well suited for estimating Gaussian processes, or objects taken from functional classes which are ellipsoids when viewed in the right basis. The problem is that many stochastic processes of scientific interest are not Gaussian and that many functional classes are not ellipsoids. Unfortunately, linear estimation is very often of poor quality in such circumstances. We give a few examples.

3.1. Non-Gaussian processes

We follow Yves Meyer and introduce the *Ramp process* $X(t)$, $t \in [0, 1)$, with periodic sample paths defined by

$$X(t) = t - 1(t \geq \tau), \tag{3.1}$$

where τ is drawn uniformly at random in $[0, 1)$. The sample path increases linearly from 0 to τ in the interval $[0, \tau)$, is decreased by 1 at $t = \tau$, and increases linearly from $\tau - 1$ to 0 in the interval $[\tau, 1)$. This process is very simple, and estimating X from noisy data is an exercise in parametric statistics. Without calculating Bayes' rule, one could recover X by simply estimating the location of the discontinuity.

The best linear estimator is given by the Wiener filter. To calculate the Karhunen–Loeve decomposition of X, Meyer observes that the covariance matrix is given by

$$\mathrm{Cov}(X(s)X(t)) = \min(s, t) - st,$$

and is the same as that of the Brownian bridge $B(t) = W(t) - tW(1)$ where W is a Brownian motion. Since $(\sqrt{2}\sin(\pi k t))_{k \geq 1}$ are the eigenfunctions of the covariance matrix of the Brownian bridge with eigenvalues $d_k^2 = (\pi k)^{-2}$, the best linear estimator would operate by linearly shrinking the Fourier coefficients of $Y(\mathrm{d}t) = X(t)\,\mathrm{d}t + \varepsilon W(\mathrm{d}t)$. Obviously, this is a poor estimation strategy since, to achieve a small MSE, partial Fourier series would need to give very good approximations of the sample paths of the process X with just a few terms (we will elaborate on this later). But the slow decay of the eigenvalues of the covariance matrix says that this is not the case. This is an instance of the well-known Gibbs phenomenon, which asserts that partial Fourier series provide poor reconstructions of otherwise smooth signals with isolated singularities. Quantitatively, the MSE of the Wiener filter is given by

$$\mathrm{MSE}(X, \hat{X}) = \sum_{k \geq 1} \frac{d_k^2 \varepsilon^2}{d_k^2 + \varepsilon^2} \geq \frac{1}{2} \sum_{k \geq 1} \min(d_k^2, \varepsilon^2), \tag{3.2}$$

since $a^2b^2/(a^2 + b^2) \geq \frac{1}{2}\min(a^2, b^2)$ for all $a, b \in \mathbb{R}$, with equality when $a = b$. With $d_k^2 = (\pi k)^{-2}$, this gives

$$\mathrm{MSE}(X, \hat{X}) = \sum_{k \geq 1} \frac{d_k^2 \varepsilon^2}{d_k^2 + \varepsilon^2} \geq \varepsilon/\pi.$$

To drive the point home, recall the asymptotic calibration $\varepsilon = 1/\sqrt{n}$, which says that if we were to think about this estimation in the sampled data model, the MSE would scale like $1/\sqrt{n}$, where n is the sample size. This is substandard since we are dealing with a parametric problem for which there are estimators converging at the parametric rate of about $1/n$ (or about ε^2).

3.2. Other functional classes

Suppose now that we are interested in estimating objects with bounded variations. A function with finite bounded variations is a function whose first derivative is a signed measure with finite mass. Then it turns out that, for this functional class, any estimator which asymptotically achieves or nearly achieves the minimax risk *must be nonlinear.* There are many such examples. Suppose the functional class is defined via

$$\mathcal{F} = \{f : \|f\|_{W_p^m} \leq R\},$$

where $\|\cdot\|_{W_p^m}$ is the L_p-Sobolev norm

$$\|f\|_{W_p^m}^2 := \int_{[0,1]} |f(t)|^p + |f^{(m)}(t)|^p \, \mathrm{d}t. \tag{3.3}$$

When $m = 1$ and $p = 1$, this definition is close to the bounded variation norm (with the proviso that the first derivative may not be an integrable function). In Section 2.4, we have seen that if $p = 2$, there is a clean solution which achieves the minimax risk and that this solution is linear. When $p < 2$, however, any estimator whose risk scales like the minimax risk as $\varepsilon \to 0$ *must be nonlinear.* In other words, linear estimators achieve markedly suboptimal rates of convergence.

Geometrically suppose that one is interested in estimating the mean vector θ from the data $y_k = \theta_k + \sigma z_k$, where the z_ks are i.i.d. $N(0, 1)$. Then, if Θ

is an ellipsoid, linear estimation is all-powerful! But suppose Θ is the body

$$\Theta := \left\{\theta : \sum_{k\geq 1} |\theta_k| \leq R\right\}.$$

This is a convex body – an octahedron to be precise – but not an ellipsoid, and this causes a substantial modification in what constitutes an optimal or near-optimal estimation strategy.

3.3. Spatial adaptivity

Suppose that the function we wish to recover has a few isolated singularities but is otherwise smooth, and that we employ a linear kernel smoother. Suppose, further, that we have available an *oracle* which supplies the best bandwidth, in the sense that it tells us which h yields the smallest MSE. This optimal choice of the bandwidth comes from the classical bias/variance trade-off: the smaller the bandwidth, the smaller the bias but the greater the variance. On the one hand, to keep the bias low we would need to use a small bandwidth, as otherwise the estimation error would be large, since one would smooth away the discontinuities. But on the other hand, to keep the variance low we would need to use a large bandwidth, as otherwise the error would be large, since one would undersmooth the flat part of the object f.

To get out of this dead end, one would like to use, instead, a spatially varying bandwidth. That is, one would like to be able to use a small bandwidth when the estimand is rough or discontinuous and a larger bandwidth when it is smooth or flat. That is, one could imagine using a spatially adaptive bandwidth which we would estimate from the data. This would turn the overall estimation strategy into a nonlinear procedure. And if we could somehow find the right bandwidth at every point, we could in principle obtain much better MSEs.

3.4. Adaptive estimation

The asymptotically optimal estimator (2.21) is sensitive to the parameters m and R which define the class $\mathcal{F} := \{f : \|f\|_{W_2^m} \leq R\}$. Should these parameters be mis-specified, statistical optimality would no longer hold. In practice, however, one must confess that we would rarely know in advance the exact degree of smoothness or the object we wish to estimate. And even if we did, we would not know the exact size of the radius of the ball. Such practical considerations suggest abandoning the idea of an asymptotically exact estimator for a particular class in favour of estimators with nearly optimal asymptotic properties *simultaneously* over a wide range of classes of interest. Admittedly, this may seem like an overly ambitious goal. Perhaps surprisingly, this is, however, possible in many interesting cases. The upshot is that such estimators are nonlinear.

4. Shrinkage estimators and oracle inequalities

In this section and the next, we consider the problem of estimating a (possibly infinite) vector $\theta \in \mathbb{R}^d$ from observations $y \sim N(\theta, 1)$, and focus on the statistical underpinnings of this problem. Only much later shall we identify θ with the coefficient sequence of a function f in an appropriate basis, and translate some of the decision-theoretic results in the language of nonparametric function estimation. The importance of this section relies upon the fact that it introduces the idea of an oracle inequality.

4.1. The James–Stein estimator

We wish to estimate $\theta \in \mathbb{R}^d$ from $y \sim N(\theta, I)$, and use the mean-squared error to measure performance

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\|\hat{\theta} - \theta\|^2$$

(here and below $\|\cdot\|$ denotes the Euclidean norm). The maximum-likelihood estimate (MLE) is of course given by $\hat{\theta}^{\mathrm{MLE}} = y$ and obeys

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{MLE}}, \theta) = d.$$

Everybody would agree that the MLE is a good estimator. After all, what other estimator could we use in the absence of any additional information about the parameter θ? The surprising discovery of James and Stein (1961) is that when $d > 2$, the MLE is not admissible. That is, there exist estimators which are more accurate than the MLE (or better than the sample mean in the case where one gets independent copies of y). Consider, for example, the estimator

$$\hat{\theta}^{\mathrm{JS}} = w(y) \cdot y, \quad w(y) = \left(1 - \frac{d-2}{\|y\|^2}\right)_+ \tag{4.1}$$

which shrinks the data y towards the origin. James and Stein proved that $\hat{\theta}^{\mathrm{JS}}$ obeys

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{JS}}, \theta) < \mathrm{MSE}(\hat{\theta}^{\mathrm{MLE}}, \theta), \quad \text{for all } \ \theta \in \mathbb{R}^d.$$

In words, the performance of the shrinkage estimator is superior to that of the sample mean *for all values of the parameter* θ. This is surprising, because y may measure seemingly unrelated quantities such as the taste of clams and the age of the universe, to paraphrase Le Cam (2000). It is therefore surprising that by mixing information about completely disconnected problems, one can obtain an estimator with a total mean-squared error that is smaller than that one would obtain by considering each problem separately.

This result has had an enormous influence on the field and is still difficult to comprehend, although, by now, there are many papers that provide some explanations for this strange phenomenon: see, for example, the empirical Bayes interpretation of Efron and Morris (1971). We will not attempt to summarize this literature and, instead, merely note that nonlinear shrinkage improves performance.

4.2. Ideal linear shrinkage estimator and oracle inequalities

It is time to revisit the main issue discussed thus far – although in an abstract setting: how much should we smooth or, rather, how much should we shrink? To estimate $\theta \in \mathbb{R}^d$ from $y \sim N(\theta, I)$, consider the family of diagonal estimators

$$\hat{\theta}^c = c \cdot y$$

where c is a scalar. For each coordinate, recall that the bias $\hat{\theta}^c_k$ is given by $\theta_k - \mathbb{E}\hat{\theta}^c_k = (1-c)\theta_k$ and the variance obeys $\operatorname{Var}(\hat{\theta}^c_k) = c^2$ so that $\mathbb{E}(\theta_k - \hat{\theta}^c_k)^2 = (1-c^2)\theta_k^2 + c^2$. Summing over coordinates gives

$$\operatorname{MSE}(\hat{\theta}^c, \theta) = (1-c)^2\|\theta\|^2 + c^2 d.$$

We now search for an *ideal estimator* which selects that estimator $\hat{\theta}^{c^*}$ from the family $(\hat{\theta}^c)_{c\in\mathbb{R}}$ with minimal MSE: that is, c^* is the solution to

$$\min_{c\in\mathbb{R}} (1-c)^2\|\theta\|^2 + c^2 d.$$

Analytically, c^* is given by

$$c^* = \frac{\|\theta\|^2}{\|\theta\|^2 + d},$$

and the ideal MSE obeys

$$\operatorname{MSE}(\hat{\theta}^{c^*}, \theta) = \frac{\|\theta\|^2 d}{\|\theta\|^2 + d}.$$

This estimator is ideal because we would of course not know which estimator $\hat{\theta}^c$ is best; that is, to achieve the ideal MSE, one would need an *oracle* that would tell us which shrinkage factor to choose. The difference from the James–Stein estimate is that $\hat{\theta}^{\mathrm{JS}}$ is estimating the shrinkage factor from the data y, while in the ideal scenario, the ideal shrinkage factor which depends on $\|\theta\|$ is simply given to us. Obviously,

$$\inf_c \operatorname{MSE}(\hat{\theta}^c, \theta) \le \operatorname{MSE}(\hat{\theta}^{\mathrm{JS}}, \theta).$$

But the interesting fact is that there is an equality in the other direction.

Theorem 4.1. The James–Stein estimate obeys

$$\operatorname{MSE}(\hat{\theta}^{\mathrm{JS}}, \theta) \le 2 + \inf_c \operatorname{MSE}(\hat{\theta}^c, \theta). \tag{4.2}$$

In other words, the James–Stein estimator is almost as good as the ideal estimator in a mean-squared error sense. When the dimension d is large, the additive factor is small compared to the MSE of the MLE, which is equal to d. The inequality (4.2) is an *oracle inequality.* An oracle inequality relates the performance of a real estimator with that of an ideal estimator which relies on perfect information supplied by an oracle, and which is not available in practice. Oracle inequalities are a powerful concept that we shall use extensively in the remainder of this paper.

To prove (4.2), one needs to come up with a formula, or at least with an estimate for the MSE of the James–Stein estimate. Perhaps the most elegant derivation is based on the *Stein unbiased risk estimate*, due to Stein (1981), which goes as follows. Let $Y \sim N(\theta, I)$ and consider the estimator $\theta = Y + g(Y)$ where $g : \mathbb{R}^d \to \mathbb{R}^d$ is a weakly differentiable function. Then, under mild integrability assumptions,

$$\mathbb{E}\|Y + g(Y) - \theta\|^2 = \mathbb{E}[d + 2\nabla \cdot g(Y) + \|g(Y)\|^2], \tag{4.3}$$

where $\nabla \cdot g(Y)$ is the divergence of g, $\nabla \cdot g(Y) := \sum_{k=1}^d \partial_k g_k(Y)$. To see why this is so, observe that

$$\mathbb{E}\|Y + g(Y) - \theta\|^2 = \mathbb{E}\|Y - \theta\|^2 + 2\mathbb{E}(Y - \theta)^T g(Y) + \mathbb{E}\|g(Y)\|^2.$$

Since $\mathbb{E}\|Y - \theta\|^2 = d$, we only need to argue that

$$\mathbb{E}(Y - \theta)^T g(Y) = \mathbb{E}\nabla \cdot g(Y).$$

This follows from an integration by parts. Let $\phi(y)$ be the density function of the standard multivariate normal distribution $\phi(y) = (2\pi)^{-n/2} e^{-\|y\|^2/2}$, and recall that $\partial_k \phi(y - \theta) = -(y_k - \theta)\phi(y - \theta)$. Then, assuming that g is sufficiently smooth,

$$\begin{aligned}\mathbb{E}(Y_k - \theta_k) g_k(Y) &= \int_{\mathbb{R}^d} (y_k - \theta_k) g_k(y) \phi(y - \theta)\, \mathrm{d}y \\ &= \int_{\mathbb{R}^d} \partial_k g_k(y) \phi(y - \theta)\, \mathrm{d}y.\end{aligned}$$

The idea is now to use the relation (4.3) to compute the MSE of the James–Stein estimate. To avoid unnecessary technicalities due to the non-differentiability of $\hat{\theta}^{\mathrm{JS}}$, we prove (4.2) with the slightly modified estimator $\hat{\theta} = \tilde{w}(y)y$, where $\tilde{w}(y) = (1 - (d-2)/\|y\|^2)$; that is, we remove the positive part. It seems intuitively clear that $\mathrm{MSE}(\hat{\theta}^{\mathrm{JS}}, \theta) \le \mathrm{MSE}(\hat{\theta}, \theta)$, which is true. With this notation, $\hat{\theta} = Y + g(Y)$, where

$$g(Y) = -\frac{d-2}{\|Y\|^2} Y.$$

Since

$$\nabla \cdot g(Y) = -\frac{(d-2)^2}{\|Y\|^2},$$

the Stein unbiased risk formula reads

$$\mathbb{E}\|Y + g(Y) - \theta\|^2 = d - (d-2)^2 \cdot \mathbb{E}\frac{1}{\|Y\|^2}.$$

Set $X = \|Y\|^2$, then $\mathbb{E}X = \|\theta\|^2 + d$, and since the function $1/x$ is convex, Jensen's inequality yields

$$\mathbb{E}\frac{1}{X} \geq \frac{1}{\mathbb{E}X} = \frac{1}{\|\theta\|^2 + d}.$$

In other words, this would give

$$E\|\hat{\theta}^{\mathrm{JS}} - \theta\|^2 \leq d - \frac{(d-2)^2}{\|\theta\|^2 + d} \leq 4 + \inf_c \mathbb{E}\|\theta^c - \theta\|^2.$$

This is not exactly the content of (4.2) since we have an additive factor of 4 instead of 2. To improve on this, we need a sharper lower bound on $\mathbb{E}\|Y\|^{-2}$. More work would show that

$$\mathbb{E}\frac{1}{\|Y\|^2} \geq \frac{1}{d - 2 + \|\theta\|^2},$$

where the equality holds if $\theta = 0$. This sharper estimate would give (4.2). We refer the reader to Johnstone (2002) for details.

4.3. Ideal shrinkage and adaptive estimation

Returning to the theme of nonparametric estimation, there is a beautiful application of such oracle inequalities. We have seen that one can find asymptotically minimax estimators for L_2-Sobolev balls of the form $\mathcal{F}^m(R) = \{f : \|f\|_{W_2^m} \leq R\}$. Pinsker's solution requires knowledge of m and R, but in practice these are unknown. Is it possible to achieve asymptotic minimaxity over $\mathcal{F}^m(R)$, simultaneously for each value of m and $R > 0$?

Taking the sequence space viewpoint, the problem is equivalent to that of estimating the Fourier coefficients (θ_k) of f from the Gaussian sequence model

$$y_k = \theta_k + \varepsilon z_k, \tag{4.4}$$

where the infinite-dimensional vector θ belongs to the ellipsoid

$$\Theta =: \left\{\theta : \sum_{j\geq 0} \sum_{k \in B_j} (1 + k^{2m})|\theta_k|^2 \leq R^2\right\}. \tag{4.5}$$

In the above expansion, we have partitioned the sum into blocks which we assume are dyadic sub-bands

$$B_j := \{k \geq 0 : 2^j \leq k < 2^{j+1}\}.$$

That is, the block B_j is the family of all those Fourier coefficients with frequency indices in the dyadic interval $[2^j, 2^{j+1})$. This partitioning goes back a long way in harmonic analysis and was first introduced by Littlewood and Paley (see Frazier, Jawerth and Weiss (1991)) to study the property of functions and of their Fourier series.

Let $d_j = 2^j$ be the size of the jth block B_j. With this notation, we introduce the block James–Stein estimator defined by

$$\hat{\theta}_j^{\mathrm{BJS}}(y) = \begin{cases} y_j, & j < J_0, \\ \left(1 - \frac{(d_j - 2)\varepsilon^2}{\|y_j\|^2}\right)_+ y_j, & J_0 \leq j < J_\varepsilon, \\ 0, & j \geq J_\varepsilon. \end{cases} \tag{4.6}$$

For example, one can set $J_0 = 2$, and J_ε to be the nearest integer to $\log_2(1/\varepsilon^2)$. The interpretation is that the very low-frequency components are untouched, the intermediate-frequency components are shrunk towards zero, and the high-frequency components are thrown away. In summary, the function $f(t)$ is estimated by (1) taking the data in the frequency domain, (2) applying the James–Stein estimator to each dyadic sub-band B_j, and (3) returning to the original time domain.

A remarkable result due to Efroĭmovich and Pinsker (1984) shows that the block James–Stein estimator is asymptotically minimax over all Sobolev ellipsoids.

Theorem 4.2. For all ellipsoids of the form (4.5), the MSE of the block James–Stein estimator (4.6) obeys

$$\sup_{\theta \in \Theta} \mathrm{MSE}(\hat{\theta}^{\mathrm{BJS}}, \theta) \leq 2^{2m} M^*(\varepsilon, \Theta)(1 + o(1)), \tag{4.7}$$

where $o(1)$ is a term tending to zero as $\varepsilon \to 0$. In fact it is possible to get asymptotic minimaxity, namely,

$$\sup_{\theta \in \Theta} \mathrm{MSE}(\hat{\theta}^{\mathrm{BJS}}, \theta) = M^*(\varepsilon, \Theta)(1 + o(1)),$$

by choosing shorter (but not too short) blocks $B_j = \{k : \ell_j \leq k \leq \ell_{j+1}\}$ obeying $\ell_{j+1}/\ell_j \to 1$.

The intuition is as follows. Suppose that we have a block $B_j = \{k : \ell_j \leq k \leq \ell_{j+1}\}$ obeying $\ell_{j+1}/\ell_j \to 1$, and let θ^j be the vector $(\theta_k)_{k \in B_j}$. The key point is that to estimate the coordinates of θ_j, an estimator of the form

$$\hat{\theta}_k^j = c_j \cdot y_k,$$

with weights depending on the block index, but not on the individual coefficients within a block, is almost as efficient as any other estimator. To understand this, one can check that Pinsker's (optimal) weights are nearly constant on each block for sufficiently large j. With the notation of Section 2.5, this is indeed a consequence of $\sup_{k,k' \in B_j} a_k/a_{k'} = (\ell_{j+1}/\ell_j)^m \to 1$. Continuing at this informal level of discussion, it follows that if we could find the best block-dependent shrinkage factor, then we would do very well. But we have seen that this is precisely what the James–Stein estimate does (Theorem 4.1). Thus $\hat{\theta}^{\mathrm{BJS}}$ is efficient and provably asymptotically minimax: see Johnstone (2002) for a rigorous argument. When one uses dyadic blocks, $\ell_{j+1}/\ell_j \to 2$ and the weights are not nearly constant but vary within a factor 2^m. Replacing these variable weights with a constant weight is responsible for the slight loss in precision; compare (4.7).

5. Ideal shrinkage and thresholding rules

All of the estimators we have encountered so far are based on the belief that large coefficients occur at low frequencies. As a consequence, high-frequency components are systematically shrunk toward zero. We remarked earlier that signals of interest may exhibit significant high-frequency components because of singularities or otherwise. Why should we then enforce shrinkage if the data provide evidence that some special high-frequency components are statistically significant or unlikely to be noise?

To makes things concrete, consider an extreme example, where $\theta \in \mathbb{R}^n$ is of the form

$$\theta = (0, \ldots, 0, \mu, 0, \ldots, 0),$$

where $\mu \neq 0$ and the location of the nonzero coordinate is not known in advance. Then it is clear that linear estimators would be highly ineffective in this setting. The James–Stein estimator, which is essentially a linear estimator – albeit with a nonlinear data-dependent shrinkage factor – would also be very ineffective. This section introduces thresholding rules which are true nonlinear estimation procedures, and which perform very well in this setting and, of course, in much more complicated settings as well.

5.1. Ideal shrinkage

We consider the same Gaussian sequence model (4.4), where we think of $(\theta_k)_{1 \le k \le n}$ as the coefficient sequence of f in a fixed basis $(\psi_k(t))_{1 \le k \le n}$. To recover $\theta \in \mathbb{R}^n$ from $y \sim N(0, \varepsilon^2 I)$, we now consider the family of diagonal shrinkage estimators

$$\hat{\theta}^w = Wy \quad \Leftrightarrow \quad \hat{\theta}_k = w_k y_k$$

where $W = \operatorname{diag}(w_k)$. Just as before, we consider the ideal estimator θ^* which minimizes the MSE among all diagonal shrinkage estimators

$$\theta^* = \operatorname{argmin}_{w \in \mathbb{R}^n} \mathbb{E}\|\hat{\theta}^w - \theta\|^2.$$

Note that we have already computed θ^*, since for each coordinate k, the optimal weight w_k^* minimizes the trade-off between the squared bias and the variance

$$\mathbb{E}(w_k y_k - \theta_k)^2 = (1 - w_k)^2 \theta_k^2 + w_k^2 \varepsilon^2$$

whose solution is given by

$$w_k^* = \frac{\theta_k^2}{\theta_k^2 + \varepsilon^2}, \quad \text{and} \quad E(\hat{\theta}_k^* - \theta_k) = \frac{\theta_k^2 \varepsilon^2}{\theta_k^2 + \varepsilon^2}.$$

Closely related is the ideal projection estimator θ^I, where we additionally require that W be a projection matrix. This condition simply says that the weights w_k are either 0 or 1,

$$\theta^I = \operatorname{argmin}_{w \in \{0,1\}^n} \mathbb{E}\|\hat{\theta}^w - \theta\|^2.$$

A simple calculation then shows that

$$\theta_k^I = w_k y_k, \quad w_k = \begin{cases} 0, & |\theta_k| < \varepsilon, \\ 1, & |\theta_k| \geq \varepsilon. \end{cases}$$

This is a keep-or-kill estimate. The interpretation is that, for $w_k = 1$, $w_k y_k$ has vanishing bias and a variance equal to ε^2, while for $w_k = 0$, $w_k y_k$ has bias θ_k and vanishing variance. The optimal choice then minimizes between the squared bias and the variance and, therefore, the risk of the ideal projection is given by

$$\mathbb{E}(\theta_k^I - \theta_k)^2 = \min(\theta_k^2, \varepsilon^2).$$

We have already seen that for $a, b \geq 0$, $ab/(a + b) \leq 2 \min(a, b)$ and thus

$$\mathbb{E}(\theta_k^I - \theta_k)^2 \leq 2 \min(\theta_k^2, \varepsilon^2),$$

which gives

$$\operatorname{MSE}(\theta^*, \theta) \leq \operatorname{MSE}(\theta^I, \theta) \leq 2 \operatorname{MSE}(\theta^*, \theta).$$

In short, the risk of the ideal projection comes within a factor of 2 of that of the ideal shrinkage estimator. From now on, it will be convenient to compare the risk of any real estimator with that of the ideal projection which obeys

$$\operatorname{MSE}(\theta^I, \theta) = \sum_k \min(\theta_k^2, \varepsilon^2). \tag{5.1}$$

We then ask the question: is it possible to find estimators whose risk comes close to that of the ideal projection?

5.2. *Thresholding rules*

In the spirit of the ideal projection, we consider thresholding rules for estimating the mean of a Gaussian distribution. There are many such rules, and we focus on the most commonly studied rules, namely the so-called hard-thresholding and soft-thresholding rules. For other types of thresholding rules, consider the garrote method of Gao (1998), for example. A hard-thresholding rule is of the form

$$\hat{\theta}_k = \begin{cases} y_k, & |y_k| \geq \lambda, \\ 0, & |y_k| < \lambda, \end{cases} \tag{5.2}$$

where λ is a some positive scalar parameter. A hard-thresholding rule yields a keep-or-kill estimate. Observations which pass the threshold are considered significant and untouched, while all observations below the threshold are set to zero. A soft-thresholding rule is similar but performs additional shrinkage:

$$\hat{\theta}_k = \begin{cases} y_k - \lambda, & y_k \geq \lambda, \\ 0, & |y_k| < \lambda, \\ y_k + \lambda, & y_k < -\lambda. \end{cases} \tag{5.3}$$

That is, the significant observations are also pulled towards zero by an amount equal to λ. We note that a soft-thresholding $\hat{\theta}(y)$ rule is a continuous function of y while the hard-thresholding rule is not. In this sense, the soft-thresholding rule is a smoother rule, hence the name.

The hard- and soft-thresholding rules also have an interpretation as minimum complexity estimates for complexity penalties which are not quadratic. For example, the hard thresholding rule at level λ is the solution to

$$\min_{\tau \in \mathbb{R}} \quad (y_k - \tau)^2 + \lambda^2 \cdot 1(\tau \neq 0),$$

while the soft-thresholding rule solves

$$\min_{\tau \in \mathbb{R}} \quad (y_k - \tau)^2 + 2\lambda \cdot |\tau|.$$

For n-dimensional problems, hard-thresholding each coordinate at level λ solves the variational problem

$$\min_{\tau \in \mathbb{R}^n} \quad \|y - \theta\|^2 + \lambda^2 \cdot \|\tau\|_{\ell_0},$$

where $\|\tau\|_{\ell_0} := \sum_{1 \leq k \leq n} 1(\tau_k \neq 0)$ is the number of nonzero components of τ. Similarly, soft-thresholding each coordinate at level λ solves the variational problem

$$\min_{\tau \in \mathbb{R}^n} \quad \|y - \theta\|^2 + 2\lambda \cdot \|\tau\|_{\ell_1},$$

where $\|\tau\|_{\ell_1} := \sum_{1 \leq k \leq n} |\tau_k|$. Hence, thresholding rules may be thought of as a complexity-penalized estimation procedure where the complexity of the fit is nonquadratic and given either by the ℓ_0 or the ℓ_1-norm.

5.3. Oracle inequalities

A foundational result in modern estimation is that correctly tuned thresholding rules nearly achieve the risk of ideal projections.

Theorem 5.1. (Donoho and Johnstone) Suppose that $n \geq 2$ and set $\lambda = \epsilon\sqrt{2 \log n}$. Assume that $y \sim N(\theta, \varepsilon^2 I_n)$ and let $\hat{\theta}$ be either a hard- or soft-thresholding estimate with parameter λ. Then

$$\mathbb{E}\|\theta - \hat{\theta}\|^2 \leq (2 \log n + 1) \cdot \left(\varepsilon^2 + \sum_{k=1}^{n} \min(\theta_k^2, \varepsilon^2)\right). \tag{5.4}$$

To sum up, the risk of a thresholding estimator is at most $2 \log n$ times larger than the ideal mean-squared error. Further, what is interesting here is that the oracle inequality (5.4) is nonasymptotic and holds for any finite sample size $n \geq 2$. Finally, we have seen somewhat sharper oracle inequalities where the multiplicative factor is actually equal to one (see (4.2)), and it is therefore legitimate to ask whether the logarithmic factor is sharp. It turns out that without any further assumptions on the parameter θ, the logarithmic factor is optimal – in an asymptotic sense.

Theorem 5.2. (Donoho and Johnstone) Consider the class of diagonal estimators obeying $\hat{\theta}_k = \hat{\theta}_k(y_k)$. Under the same assumptions as before,

$$\inf_{\hat{\theta} \text{ diagonal}} \sup_{\theta \in \mathbb{R}^n} \frac{\mathbb{E}\|\theta - \hat{\theta}\|^2}{\varepsilon^2 + \sum_k \min(\theta_k^2, \varepsilon^2)} \to 2 \log n \quad \text{as} \quad n \to \infty. \tag{5.5}$$

The above result says that when the parameter space of interest is $\mathbb{R}^n$, then from a minimax point of view, no diagonal estimator can essentially do better, at least asymptotically.

5.4. Risk of thresholding rules

This section gives a proof of Theorem 5.1 for the soft-thresholding rule. The proof for the hard-thresholding rule is similar and is only more technical. We may also just assume that $\varepsilon = 1$ as the general case follows from a simple rescaling argument.

We need to develop a formula for the risk of a scalar soft-thresholding rule and introduce some notation. We let η_S be the scalar nonlinearity $\eta_S(y) = \text{sgn}(y)(y - \lambda)_+$ and let $r_S(\lambda, \mu)$ be the risk of the soft-thresholding rule η_S, *i.e.*,

$$r_S(\lambda, \mu) = \mathbb{E}(\eta_S(y) - \mu)^2, \qquad y \sim N(\mu, 1).$$

Because soft-thresholding rules treat each coordinate separately, the idea of the proof is to develop an upper bound on the accuracy of scalar thresholding rules for $\mu = 0$ in a first step, and to use the bound to deduce a bound for all values of $\mu \in \mathbb{R}$ in a second step. This strategy uses the following lemma.

Lemma 5.3. The risk of the soft-thresholding rule obeys

$$r_S(\lambda, \mu) \le r_S(\lambda, 0) + \min(\mu^2, 1 + \lambda^2). \tag{5.6}$$

Proof. The proof is an exercise in calculus. By symmetry, we may just as well assume that $\mu \ge 0$. Note that

$$\begin{aligned} r_S(\lambda, \mu) &= \int (\eta_S(y) - \mu)^2 \, \phi(y - \mu) \, \mathrm{d}y \\ &= \mu^2 \int_{|y| \le \lambda} \phi(y - \mu) \, \mathrm{d}y + \int_{y > \lambda} (y - \lambda - \mu)^2 \, \phi(y - \mu) \, \mathrm{d}y \\ &\qquad + \int_{y < -\lambda} (y + \lambda - \mu)^2 \, \phi(y - \mu) \, \mathrm{d}y, \end{aligned}$$

where $\phi(y) = (2\pi)^{-1/2} e^{-y^2/2}$. A change of variables then gives

$$r_S(\lambda, \mu) = \mu^2 \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z) \, \mathrm{d}z + \int_{\lambda-\mu}^{\infty} (z - \lambda)^2 \, \phi(z) \, \mathrm{d}z + \int_{-\infty}^{-\lambda-\mu} (z + \lambda)^2 \, \phi(z) \, \mathrm{d}z,$$

which shows that the derivative with respect to μ obeys

$$\partial_\mu r_S(\lambda, \mu) = 2\mu \int_{-\lambda-\mu}^{\lambda-\mu} \phi(z) \, \mathrm{d}z \le 2\mu.$$

Therefore, $r_S(\lambda, \mu)$ is increasing in μ, and on the one hand

$$r_S(\lambda, \mu) \le \lim_{\mu \to \infty} r_S(\lambda, \mu) = 1 + \lambda^2.$$

On the other hand,

$$r_S(\lambda, \mu) - r_S(\lambda, 0) \le \int_0^\mu 2u \, \mathrm{d}u = \mu^2,$$

and we conclude that

$$r_S(\lambda, \mu) \le \min(r_S(\lambda, 0) + \mu^2, 1 + \lambda^2),$$

which proves the lemma. □

It is interesting to note that we established an estimate which is slightly better than (5.6). The quantity $\min(r(\lambda, 0) + \mu^2, 1 + \lambda^2)$ is of interest because one can prove that this is a proxy for the risk of the soft-thresholding rule since there is an inequality in the other direction:

$$r_S(\lambda, \mu) \ge \frac{1}{2} \min(r_S(\lambda, 0) + \mu^2, 1 + \lambda^2). \tag{5.7}$$

In other words, the risk of soft-thresholding is just about $\min(r_S(\lambda, 0) + \mu^2, 1 + \lambda^2)$.

The second lemma develops a bound on $r_S(\lambda, 0)$.

Lemma 5.4. The risk of the soft-thresholding rule obeys

$$r_S(\lambda, 0) \le \frac{2\phi(\lambda)}{\lambda}. \tag{5.8}$$

Proof. By symmetry of the Gaussian distribution, the risk $r_S(\lambda, 0)$ obeys

$$r_S(\lambda, 0) = 2 \int_{y>\lambda} (y - \lambda)^2 \, \phi(y) \, \mathrm{d}y,$$

and an integration by parts shows that

$$\int_{y>\lambda} (y - \lambda)^2 \, \phi(y) \, \mathrm{d}y = -\lambda\phi(\lambda) + (1 + \lambda^2)\Phi([\lambda, \infty)),$$

where $\Phi([\lambda, \infty)) = \int_{y\in[\lambda,\infty)} \phi(y) \, \mathrm{d}y$. The claim then follows from

$$\Phi([\lambda, \infty)) \le \int_\lambda^\infty \phi(y) \, \mathrm{d}y \le \int_\lambda^\infty \frac{y}{\lambda}\phi(y) \, \mathrm{d}y = \frac{\phi(\lambda)}{\lambda}. \qquad \square$$

We now specialize (5.6) and (5.8) to $\lambda = \sqrt{2 \log n}$, which gives

$$r_S(\sqrt{2 \log n}, 0) \le \frac{1}{n \sqrt{\pi \cdot \log n}} \le \frac{2 \log n + 1}{n},$$

as soon as $n \ge 2$. This proves Theorem 5.1 since

$$\begin{aligned} \mathbb{E}\|\theta - \hat{\theta}\|^2 &\le n \cdot r_S(\sqrt{2 \log n}, 0) + \sum_k \min(\theta_k^2, 1 + 2 \log n) \\ &\le (1 + 2 \log n) + \sum_k \min(\theta_k^2, 1 + 2 \log n) \\ &\le (2 \log n + 1)\Big(1 + \sum_k \min(\theta_k^2, 1)\Big), \end{aligned}$$

as claimed.

5.5. Choice of threshold

Besides the fact that $\lambda = \sqrt{2 \log n}$ allows proving sharp estimation results, there is a large literature arguing why this is intuitively the correct threshold for the Gaussian model. One explanation is as follows. Suppose that θ is identically equal to zero, *i.e.*, $\theta_i = 0$ for all is. In the language of signal estimation, this assumption states that there is no signal and that y is just white noise, $y \sim N(0, I_n)$. Then one would like to declare that there is no signal, *i.e.*, we would like to have an estimator obeying $\hat{\theta}_i = 0$ for all *is* with large probability. In the language of tests of hypotheses, we would like to accept the null hypothesis (which postulates that there is no signal) with large probability whenever the null is true. From this standpoint, one should select a threshold λ so that

$$P(\max_i |z_i| > \lambda) \le \alpha, \qquad z_i \text{ i.i.d. } N(0,1),$$

where α is a tolerance set in advance. In other words, λ should be a quantile of the distribution of the maximum absolute value of n i.i.d. standard normal random variables. It is well known (Williams 1991), however, that

$$\lim_{n\to\infty} \frac{\max_{1\le i\le n} |z_i|}{\sqrt{2\log n}} = 1 \text{ almost surely,}$$

which justifies the choice of threshold in an asymptotic sense.

This can be made a little more quantitative. In fact, it is possible to show that

$$\lim_{n\to\infty} \mathbb{P}(\max_{1\le i\le n} |z_i| > \sqrt{2\log n}) = 0,$$

which shows that asymptotically $\mathbb{P}(\hat{\theta} = 0) \to 1$ as $n \to \infty$ whenever $\theta = 0$. Introduce the indicator variables

$$I_k(\lambda) = \begin{cases} 1, & |z_k| \ge \lambda, \\ 0, & |z_k| < \lambda. \end{cases}$$

Then

$$\mathbb{P}(\max_k |z_k| > \lambda) \le \sum_k \mathbb{E}[I_k(\lambda)] = n \cdot \mathbb{P}(|z_1| > \lambda) \le 2n\,\frac{\phi(\lambda)}{\lambda},$$

which gives

$$\mathbb{P}(\max_k |z_k| > \sqrt{2\log n}) \le \frac{1}{\sqrt{\pi \cdot \log n}},$$

and the right-hand side tends to zero as n tends to infinity. Conversely, for a fixed threshold λ, the expected number of observations above λ in absolute value obeys

$$\sum_k \mathbb{E}[I_k(\lambda)] = n \cdot \mathbb{E}[I_1(\lambda)] = n \cdot \Phi([\lambda,\infty)) \ge 2n \cdot \frac{\phi(\lambda)}{\lambda} \cdot \left(1 - \frac{1}{\lambda^2}\right).$$

This shows that for λ slightly smaller than $\sqrt{2\log n}$, *i.e.*, $\lambda = (1-\delta)\cdot\sqrt{2\log n}$ for some $\delta > 0$, the number of expected white noise coordinates above threshold tends to infinity as n increases.

Having said all this, one still needs to keep in mind that the $\sqrt{2\log n}$ threshold is driven by asymptotic considerations. In practice, this choice tends to be a little too conservative, in the sense that its bias has a tendency to be a little too large. That is, many coordinates in which the value of θ_k is potentially large are set to zero. In statistical terms, the burden of proof to be deemed 'estimable' is perhaps not as reasonable as one would want. We shall later discuss more flexible and adaptive choices of threshold.

5.6. Example: estimating a very sparse vector

Thresholding is very effective for estimating sparse vectors $\theta \in \mathbb{R}^n$, *i.e.*, vectors which only have a few significant coordinates with unknown *a priori* locations. We illustrate this with a simple toy example. We observe

$$y_k = \theta_k + z_k, \quad z_k \text{ i.i.d. } N(0,1), \quad k = 1, \dots, n,$$

and suppose that all the coefficients are zero except for two spikes, each of size $\mu = \sqrt{n/2}$. (We have adjusted the heights of the spikes so that $\|\theta\|^2 = n = \mathbb{E}\|z\|^2$, so that the signal to noise ratio is one.) The James–Stein estimate is highly ineffective in this setting since the risk of the ideal shrinkage estimator $\hat{\theta}^* = c^* y$ studied in Section 4 obeys

$$\mathbb{E}\|\theta - \theta^*\|^2 \geq n/2. \tag{5.9}$$

Note that the risk of the MLE is n.

In contrast, consider the risk of a hard-thresholding rule with $\lambda = \sqrt{2 \log n}$.

(1) The two observations corresponding to the spikes pass the threshold with overwhelming probability; for each coordinate, the risk is thus about equal to the variance which is one. Formally, for any such coordinate, the risk is equal to

$$\mu^2 \mathbb{E} 1\{|Z + \mu| < \lambda\} + \mathbb{E}[Z^2 1\{|Z + \mu| > \lambda\}] \leq \mu^2 \mathbb{E} 1\{|Z + \mu| < \lambda\} + 1,$$

where Z is a standard normal random variable. Now, because $\mu = \sqrt{n/2}$ and $\mathbb{E} 1\{|Z + \mu| < \lambda\}$ is ridiculously small, *i.e.*, exponentially decaying in n, the risk is about 1.

(2) In all other coordinates, the estimator sets all the data to zero except for a possibly minuscule fraction of noise realizations exceeding the threshold. For each such coordinate, the risk obeys

$$\mathbb{E}[Z^2 1\{|Z| > \lambda\}] \leq 2(\lambda + \lambda^{-1})\phi(\lambda) = \frac{2}{\sqrt{\pi}} \cdot \frac{\sqrt{\log n}}{n}.$$

In conclusion, the risk of the hard-thresholding rule is about

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 \lessapprox 2 + (n-2)\frac{1.13\sqrt{\log n}}{n} \approx 2 + 1.13\sqrt{\log n},$$

which is far better than (5.9).

More generally, the oracle inequality guarantees that if the mean vector θ is sparse in the sense that it has S nonzero and 'significant coordinates', then the mean-squared error of the thresholding rule obeys

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 \leq (2\log n + 1) \cdot (S + 1),$$

which, ignoring the log-factor, is the MSE one would obtain if one had an oracle supplying perfect information about the location of those significant coordinates. In conclusion, thresholding is very effective when the mean vector is sparse – when there is a comparably small number of large coefficients at unpredictable locations so that one cannot say *a priori* where the 'significant coefficients' will be.

6. Interactions with modern harmonic analysis

We have seen that thresholding comes close to the ideal risk (5.1) so that one can think of the ideal risk as a proxy for the performance of thresholding estimators in the white noise model.

6.1. Interpretation of the ideal risk

We now give an interpretation of the ideal risk which links statistical estimation to other contemporary topics. We rearrange the coefficient sequence $(\theta_1, \ldots, \theta_n)$ in decreasing order of magnitude $|\theta|_{(1)} \geq |\theta|_{(2)} \geq \cdots \geq |\theta|_{(n)}$ and let $N(\varepsilon)$ be the number of those coefficients whose absolute value exceeds the noise level ε:

$$N(\varepsilon) = \#\{k : |\theta_k| \geq \varepsilon\}.$$

With this notation, one can express the ideal risk as

$$\begin{aligned}\sum_k \min(\theta_k^2, \varepsilon^2) &= N(\varepsilon) \cdot \varepsilon^2 + \sum_{k > N(\varepsilon)} |\theta|^2_{(k)} \\ &= N(\varepsilon) \cdot \varepsilon^2 + e^2_{N(\varepsilon)}(\theta),\end{aligned}$$

where for a fixed number B, $e^2_B(\theta)$ is the approximation obtained by keeping the B largest coefficients of θ:

$$e_B(\theta)^2 = \|\theta - \theta_B\|^2;$$

θ_B is the truncated vector equal to the B-largest value of θ and zero otherwise. In other words, the proxy for the risk is simply equal to the number of terms above the noise level times the squared noise level plus the approximation error.

The interpretation is now self-evident. Suppose we are interested in estimating an object f and that θ is the coefficient sequence of f in an orthobasis $\mathcal{B}$. Then the mean-squared error of the thresholding estimator in this basis is small if and if the signal f is *compressible* in this basis. That is, if and only if it is possible to obtain an accurate approximation of the signal f with a superposition of just a few selected elements from the basis $\mathcal{B}$. This links nonparametric estimation with nonlinear approximation theory, a subject concerned with methods for finding good approximations to various classes of functions.

It is also interesting to compare the ideal risk with the risk of a *linear* projection

$$\hat{\theta}_k^L = \begin{cases} y_k, & k \in \mathcal{M}, \\ 0, & \text{otherwise}, \end{cases}$$

where the set $\mathcal{M}$ would be set in advance (for example, a set corresponding to low-frequency waveforms). The MSE of this projection obeys

$$\mathbb{E}\|\theta^L - \theta\| \le \#\mathcal{M}\varepsilon^2 + \sum_{k \notin \mathcal{M}} |\theta_k|^2,$$

where the second term of the right-hand side is of course the linear approximation error. The performance of linear projection procedure depends on the precision of linear approximation, while that of thresholding depends on that of nonlinear approximation. Because nonlinear approximation is in general much more precise than linear approximation, thresholding rules are usually far more accurate than the linear estimation strategies we discussed earlier.

There is also a connection to the problem of data compression in information theory. Consider encoding a function $f \in \mathbb{R}^n$ (a digital signal or a digital image) by the method of wavelet transform coding. First, one quantizes its wavelet coefficients $\theta_k = \langle f, \psi_k \rangle$ into integers n_k using a uniform quantum q: for example, one rounds up the coefficients to the nearest multiple of $2q$. One encodes the positions and values of the nonzero coefficients as bit strings by standard devices (run-length coding and so forth). Later, an approximate reconstruction of f can be obtained from $f^q = 2q \sum_k n_k \, \psi_k$. Here we retain the index q to remind us that the quantization stepsize q controls the behaviour of the algorithm. This coding method has distortion $\delta(q)$ obeying

$$\delta(q) \le N(q)q^2 + \sum_{k > N(q)} |\theta|^2_{(k)} = N(q) \cdot q^2 + e^2_{N(q)}(\theta), \tag{6.1}$$

and is the ideal risk with the quantum playing the role of the noise level.

6.2. Sparsity

From a certain viewpoint, statistical estimation, nonlinear approximation, and data compression are closely related. For example, the quality of estimation by thresholding rules depends on the sparsity of the coefficient sequence $(\theta_k)_{k\geq 1}$. One measure of sparsity is the Marcinkiewicz weak-ℓ_p norm defined by

$$\|\theta\|_{w\ell_p} := \sup_{k\geq 1} k^{1/p}\,|\theta|_{(k)}. \tag{6.2}$$

(In all rigour, $\|\cdot\|_{w\ell_p}$ is only a quasi-norm in the sense that it does not obey the triangle inequality, but only $\|\theta^0+\theta^1\|_{w\ell_p} \leq c_p \cdot (\|\theta^0\|_{w\ell_p} + \|\theta^1\|_{w\ell_p})$ where c_p is a constant which can be calculated explicitly.) Suppose that $\|\theta\|_{w\ell_p} < \infty$, then the reordered entries of the possibly infinite sequence $(\theta_k)_{k\geq 1}$ decay at least as fast as $k^{-1/p}$; the smaller p, the faster the decay. We will be interested in bounded sequences in the weak-ℓ_p norm

$$w\ell_p(R) = \{(\theta_k) : |\theta|_{(k)} \leq R \cdot k^{-1/p}, \quad \text{for all} \quad k \geq 1\},$$

which are those sequences that exhibit a special power law decay. Note that weak-ℓ_p balls are slightly larger than corresponding ℓ_p balls

$$\ell_p(R) \subset w\ell_p(R), \qquad \ell_p(R) := \left\{(\theta_k),\ \sum_k |\theta_k|^p \leq R^p\right\}.$$

Weak-ℓ_p norms are useful because the decay of the ideal risk, as $\varepsilon \to 0$, or of the approximation error $e_B(\theta)$, as $B \to \infty$, are simply deduced from membership of $w\ell_p(R)$. We follow Donoho (1993), and introduce norms which measure the precision of nonlinear approximation and the size of the ideal risk. To measure the asymptotics of approximation/compression, define the quasi-norm

$$\|\theta\|_{c,m} = \sup_{k\geq 1} k^m \cdot e_k(\theta),$$

which says that $\|\theta\|_{c,m}$ is finite if and only if the approximation error $e_k(\theta)$ obeys $e_k(\theta) = O(k^{-m})$. In a similar fashion, we introduce a quasi-norm to measure the scaling of the ideal risk

$$\|\theta\|_{e,r} = \sup_{\varepsilon>0}\left(\varepsilon^{-2r} \cdot \sum_k \min(\theta_k^2, \varepsilon^2)\right)^{1/2},$$

which says that $\|\theta\|_{e,r}$ is finite if and only the ideal risk is $O(\varepsilon^{2r})$.

Lemma 6.1. (Donoho 1993) Let $p > 0$ and set $m = 1/p - 1/2$ and $r = \frac{2m}{2m+1}$. Then all these quasi-norms are equivalent: there exist positive finite constants $c_i(p)$ such that

$$c_0(p)\|\theta\|_{c,m} \leq \|\theta\|_{w\ell_p} \leq c_1(p)\|\theta\|_{c,m},$$
$$c_2(p)\|\theta\|_{e,r} \leq \|\theta\|_{w\ell_p} \leq c_3(p)\|\theta\|_{e,r}.$$

The assertions that $|\theta_{(k)}| = O(k^{-1/p})$, or $e_k(\theta) = O(k^{-m})$, or the ideal risk is $O(\varepsilon^{2r})$ are, therefore, all roughly equivalent. Sparsity implies good compressibility, which in turn implies good estimation.

6.3. Minimax estimation of weak-ℓ_p balls

Consider the infinite Gaussian model (4.4) and suppose $\theta \in \Theta \subset w\ell_p(R)$. Lemma 6.1 shows that the ideal risk obeys

$$\sum_k \min(\theta_k^2, \epsilon^2) = O((\epsilon^2)^{\frac{2m}{2m+1}}), \quad 1/p =: m + 1/2.$$

If one further makes an extra assumption on Θ, which roughly says that the large coefficients of $\theta \in \Theta$ do not occur at infinity, thresholding achieves the ideal risk up to a multiplicative logarithmic factor scaling like $O(\log \varepsilon)$. For example, assume that

$$\sum_{k > n_\varepsilon} |\theta_k|^2 = O(\varepsilon^{2r}), \tag{6.3}$$

where n_ε grows at most polynomially in ε. Then set

$$\hat{\theta}_k = \begin{cases} \eta(y_k), & k \le n_\varepsilon, \\ 0, & k \ge n_\varepsilon, \end{cases}$$

where η is a thresholding rule at $\lambda = \varepsilon \cdot \sqrt{2 \log n_\varepsilon}$; we threshold the coefficients in the zone $k \in [1, n_\varepsilon]$ and throw out the others. Then the oracle inequality (5.4) together with (6.3) give

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 \le O(\log \varepsilon) \cdot (\epsilon^2)^{\frac{2m}{2m+1}}. \tag{6.4}$$

To develop lower bounds, we use a standard argument, which consists in embedding large hypercubes or hyper-rectangles in Θ. Suppose that

$$\ell_{p,+}(R) \subset \Theta,$$

where this means that Θ contains n-dimensional hyper-rectangles of the form $[0, R\, n^{-1/p}]^n$ for arbitrary large n. Then the minimax risk obeys

$$\inf_{\hat{\theta}} \sup_{\Theta} \mathbb{E}\|\hat{\theta} - \theta\|^2 \ge \inf_{\hat{\theta}} \sup_{\ell_{p,+}(R)} \mathbb{E}\|\hat{\theta} - \theta\|^2,$$

and we will show that the minimax risk over the hyper-rectangle is bounded below by

$$\inf_{\hat{\theta}} \sup_{\ell_{p,+}(R)} \mathbb{E}\|\hat{\theta} - \theta\|^2 \ge c \cdot R^p \cdot (\epsilon^2)^{\frac{2m}{2m+1}}, \tag{6.5}$$

for some positive constant $c > 0$.

To establish (6.5), we choose a prior π which is supported on the vertices of the hyper-rectangle

$$\mathcal{H} := \prod_k [0, \tau_k] \subset \Theta,$$

and defined by

$$\theta_k = \begin{cases} 0, & \text{with probability } 1/2, \\ \tau_k, & \text{with probability } 1/2, \end{cases}$$

with independent coordinates so that informally $\pi(\theta) = \prod_k \pi(\theta_k)$. Since the coordinates are independent, any given coordinate does not give any information about any other and, therefore, good procedures treat each coordinate individually. In fact, we have already seen that Bayes' rule is indeed given by

$$\hat{\theta}_{\pi,k} = \mathbb{E}(\theta_k \mid y_k).$$

Suppose that the rectangle is tuned so that the sidelength is about equal to the noise level, *i.e.*, we pick n_ε as the largest integer obeying

$$R\, n_\varepsilon^{-1/p} \le \varepsilon,$$

so that $n_\varepsilon \approx R^p \varepsilon^{-p}$. It follows from the choice of parameters that $\theta_k = 0$ with probability 1/2 and $\theta_k \approx \varepsilon$ with probability 1/2. Assume for simplicity that $\theta_k = \varepsilon$ with probability 1/2. A simple rescaling argument shows that

$$\mathbb{E}(\hat{\theta}_{\pi,k} - \theta_k)^2 = B \cdot \varepsilon^2,$$

where B is the Bayes risk of estimating $\theta_k \in \{0, 1\}$ from $y_k \sim N(\theta, 1)$ with a prior which puts equal probability on both outcomes. Therefore, with this choice of prior on the hyper-rectangle, the Bayes risk obeys

$$\begin{aligned} B(\pi) \ge B \cdot n_\varepsilon \cdot \varepsilon^2 &\approx B \cdot R^p \cdot \varepsilon^{2-p} \\ &= B \cdot R^p \cdot (\varepsilon^2)^{\frac{2m}{2m+1}}, \end{aligned}$$

as claimed.

In closing, we have thus established that the minimax risk of weak-ℓ_p balls with the tail property (6.3) is at most within a logarithmic factor of the ideal risk, and that thresholding rules are nearly minimax since they are also within a logarithmic factor of the ideal risk.

6.4. Statistical estimation and harmonic analysis

The consequence of these results is that the problem of finding efficient representations becomes central now that the benefits of sparsity are well understood. The goal is then (1) to identify problems and object classes of scientific interest, and (2) to find efficient representations (orthobases) for those classes. Once such orthobases are constructed, one simply transforms the data into those bases, applies thresholding, and inverts the transformation to separate signal from noise. The best basis to use is of course that in which the objects considered have the sparsest representation. Additionally, one might be interested in representations with fast algorithms for computational efficiency. These are the areas of preoccupation of modern harmonic analysis and this is the reason why, over the last decade or so, there has been, and still is, significant interaction between these two communities.

One such important development is that the program outlined above has been perfectly executed when the functional classes under study belong either to the L_2-Sobolev scale, the L_p-Sobolev scale, or the Besov and Triebel–Lizorkin scales. All these spaces admit *unconditional bases* which are especially well adapted to the estimation problem.

6.5. Optimality of unconditional bases

Assume we are given a function space with a norm $\|f\|_{\mathcal{F}}$. Then an orthonormal basis $(\phi_k)_k$ is said to be *unconditional* for the normed space $\mathcal{F}$ if, for all choices of signs,

$$\big\|\sum \pm_k \theta_k(f)\, \varphi_k\big\|_{\mathcal{F}} \leq C \cdot \|f\|_{\mathcal{F}},$$

where $(\theta_k(f))$ are the coefficients of f in the basis (ϕ_k). This says that arbitrary changes of signs in the expansion do not change the norm by much. Another way to put it is that there is an equivalent norm $\|\theta\|_{\mathbf{f}}$ in the sequence space

$$\|f\|_{\mathcal{F}} \sim \|\theta(f)\|_{\mathbf{f}}$$

obeying

$$\|(\pm_i \theta_i)\|_{\mathbf{f}} = \|\theta\|_{\mathbf{f}}$$

for all choices of signs.

Define Θ as the image of the unit ball in the sequence space

$$\Theta = \{\theta(f) : \|f\|_{\mathcal{F}} \leq 1\},$$

and its critical exponent

$$p^*(\Theta) := \inf\{p : \ \Theta \subset w\ell_p\}.$$

Then, for any orthogonal transform U, Donoho (1993) shows that

$$p^*(U\Theta) \geq p^*(\Theta). \tag{6.6}$$

For a fixed U, one should think of $U\Theta$ as the body of coefficients of the unit ball in another basis. With this in mind, the interpretation is that, among all orthobases, the unconditional basis is that which provides the sparsest coefficient sequence. As a consequence, if there is an unconditional basis, this is the best orthonormal basis to use for nonlinear approximation and for diagonal estimation, in the sense that it provides optimal rates of approximation/estimation.

Fortunately, harmonic analysts have constructed unconditional bases for some important cases of function spaces. Some notable examples are as follows (Meyer 1992).

- Fourier bases are unconditional bases for L_2-Sobolev spaces in any dimension.
- Wavelet bases are unconditional bases for L_p-Sobolev spaces in any dimension.
- Wavelet bases are unconditional bases for Besov and Triebel spaces in any dimension. These spaces depend on 3 parameters (m, p, q) and are extensions of L_p-Sobolev spaces which depend on the pair (m, p): see Triebel (1992) for a definition.

6.6. The wavelet shrinkage

Suppose we wish to recover objects taken from a Besov or a Triebel body from the data

$$Y(\mathrm{d}t) = f(t)\,\mathrm{d}t + \varepsilon W(\mathrm{d}t),$$

and seek an estimator $\hat{f}$ which nearly achieves the minimax risk. Then the answer is simply given by the celebrated wavelet shrinkage algorithm of Donoho. We take a nice wavelet basis $\psi_{j,k}(t)$, where $j \geq j_0$ indexes the scale of the wavelet and $k = 0, 1, \ldots, 2^j - 1$ indexes the location of the wavelet, go into the wavelet domain, and estimate the coefficients of f in the wavelet basis via

$$\hat{\theta}_{j,k}(y) = \begin{cases} y_{j,k}, & j = j_0, \\ \eta(y_{j,k}), & j_0 < j < j_\varepsilon, \\ 0, & j \geq j_\varepsilon; \end{cases} \tag{6.7}$$

in the above equation, the $y_{j,k}$s are the noisy coefficients, and η is a hard- or soft-thresholding rule at the level $\lambda = \varepsilon \cdot \sqrt{2 \log n_\varepsilon}$, where n_ε is the number of coefficients to which the scalar nonlinearity applies. For example, one can set j_ε to be the nearest integer to $\log_2(1/\varepsilon^2)$ so that $n_\varepsilon \approx 1/\varepsilon^2$. Inverting the wavelet transforms gives the estimate

$$\hat{f}(t) = \sum_{j,k} \hat{\theta}_{j,k} \psi_{j,k}(t). \tag{6.8}$$

This estimator has a simple structure since we just take the data in the wavelet domain and throw out the small coefficients.

As an example, suppose we are interested in the space of two-dimensional functions on $[0,1]^2$ of bounded variation,

$$\mathcal{F} := \{f : \|f\|_{\mathrm{BV}} \leq 1\}.$$

We recall that the bounded variation norm is given by $\|f\|_{\mathrm{BV}} = \int |\mathrm{d}f|$. Technically speaking, the space of functions of bounded variations does not admit an unconditional basis, although it is tightly bracketed between two Besov spaces with wavelet orthobases as unconditional bases. Letting $\Theta = \{\theta(f),\ f \in \mathcal{F}\}$ be the coefficient sequence in a sufficiently nice wavelet basis, it is possible to use embeddings of Besov spaces to show that

$$\ell_{1,+}(R) \subset \Theta,$$

for some positive $R > 0$. As we have seen earlier, this immediately gives

$$\inf_{\hat{f}} \sup_{\mathcal{F}} \mathrm{MSE}(f, \hat{f}) \geq c \cdot \varepsilon.$$

The minimax risk of two-dimensional functions with controlled bounded variations goes to zero as least as slowly as ε. In the other direction, a result of Cohen, DeVore, Petrushev and Xu (1999) shows that the wavelet sequence of a function with bounded variations belong to the weak-ℓ_1 ball, which gives that the ideal risk in our wavelet basis obeys

$$\mathbb{E}\|\theta^I - \theta\|^2 \leq C \cdot \varepsilon.$$

Since the wavelet shrinkage estimate $\hat{f}$ (6.7)–(6.8) in a 2-dimensional basis comes within a logarithmic factor of the ideal risk, we have

$$\sup_{\mathcal{F}} \mathbb{E}\|f - \hat{f}\|^2 = O(\log \epsilon^{-1}) \cdot \inf_{\hat{f}} \sup_{\mathcal{F}} \mathrm{MSE}(f, \hat{f})$$

and it is, therefore, asymptotically near-optimal.

6.7. Adaptive minimaxity

The wavelet shrinkage algorithm does not really depend upon the parameters of the functional class one wishes to estimate, which in practice are not known. To guarantee near-optimality, we simply need to work with a basis which is unconditional for the functional class and correctly set the thresholding zone. Seen a little bit differently, suppose first that we settle on a nice wavelet basis. Our basis may not be an unconditional basis for *all* L_p-Sobolev spaces or *all* Besov spaces, but it will be an unconditional basis for many of them, *e.g.*, for all L_p-Sobolev space with $m \leq m_1$ and $p \geq 1$. (For the specialist, the regularity of the wavelet limits the smoothness range over which the fixed wavelet basis is unconditional.) Second, suppose that we ignore small-scale coefficients, *e.g.*, exceeding a fixed scale $j_\varepsilon = \log_2(1/\varepsilon^2)$ which only depends upon the noise level. Then Donoho, Johnstone, Kerkyacharian and Picard (1995) show that the wavelet shrinkage nearly achieves the asymptotic minimax risk for each value of the parameter $m \in [m_0, m_1]$, p, and $R > 0$ (R is the radius of the ball). This is another example of adaption by an oracle inequality.

This universal aspect of wavelet shrinkage should not be understated. The *same* algorithm is near-optimal simultaneously over a wide range of functional classes and the performance automatically adapts to that one would expect if one knew the functional class in advance. The wavelet shrinkage may not be an exact solution to a tightly specified minimax problem but it is an approximate solution for many interesting problems.

6.8. Challenges and limitations

In summary, we have seen that efficient representations lead to efficient estimations, and that certain representations emerge as optimal. In addition, the same representation may very well solve many estimation problems (adaptivity). The challenge is, therefore, to find optimal representations for models of scientific interest. For those models, unconditional bases are, however, unlikely ...

7. Empirical model selection

We have just learned that thresholding in an unconditional basis is statistically near-optimal. Arguably, such results are very satisfying except for the fact that, more often than not, unconditional bases are simply not available. For example, a commonly discussed and interesting model of images without an unconditional is the class of functions $f(x_1, x_2) \in Ł_2([0,1]^2)$, which are twice differentiable away from edges with bounded curvature. To say this slightly differently, our class is composed of objects that are discontinuous along smooth curves, *i.e.*, edges, but otherwise smooth so that one can think about such objects as cartoon-like images. This class and many others do not admit unconditional bases and, therefore, one needs to extend the tools for adaptive estimation to deal with these more common situations. This section has two goals: (1) to develop more flexible estimation strategies which go beyond coefficient estimation in a single basis, and (2) to show that it is possible to deal with classes other than the traditional smoothness classes.

7.1. Estimation with general dictionaries

Instead of being sparse in an orthobasis, a signal $f(t)$ might be sparse in a general dictionary $\mathcal{D}$ of waveforms denoted by $\mathcal{D} = (\varphi_i(t))_{i \in I}$, where I is a finite or countable set. The elements $\varphi_i(t)$ of $\mathcal{D}$ may not be orthogonal or even linearly independent. Given such a dictionary, we will assume that one can write $f(t)$ as the linear combination

$$f(t) = \sum_i \theta_i \varphi_i(t),$$

where this expansion is not unique in the case where the dictionary $\mathcal{D}$ is overcomplete (meaning that the φ_is are linearly dependent). As before, we wish to recover an object from the sampled data model (2.8) or from the continuous white noise model (2.18), and seek an estimator of the form

$$\hat{f}(t) = \sum_i \hat{\theta}_i \varphi_i(t). \tag{7.1}$$

This problem is central in statistics since this is none other than the classical multivariate regression problem, which we discuss next.

7.2. Model selection

To simplify matters, suppose that we have a finite problem and let $\Phi \in \mathbb{R}^{n\times p}$ denote the matrix whose columns are the individual waveforms $\varphi_i(t)$, $t = 1, \ldots, n$, so that the sampled model assumes the form

$$y = \Phi\theta + z,$$

where y is an n-dimensional vector of observations, and $z \sim N(0, \sigma^2 I_n)$ is white noise. Note that when the dictionary is overcomplete, one has $p > n$. We are interested in estimating the object $f = \Phi\theta$ and measure performance with the MSE

$$\mathbb{E}\|\Phi\theta - \Phi\hat{\theta}\|^2 = E\|f - \hat{f}\|^2,$$

where $\hat{f} = \Phi\hat{\theta}$ is our estimate.

We turn our attention to ideas which generalize ideal projection rules. Suppose we are given a subset $\mathcal{M} \subset \{1, \ldots, p\}$ of coordinates, and denote by $V(\mathcal{M})$ the span of $\mathcal{M}$, namely,

$$V(\mathcal{M}) := \{a \in \mathbb{R}^p : a_i = 0 \quad \text{for all} \quad i \notin \mathcal{M}\}.$$

We then consider the least squares estimate which is the solution to

$$\hat{\theta}[\mathcal{M}] = \text{argmin}_{a\in V(\mathcal{M})} \|y - \Phi a\|^2.$$

For example, in the case where Φ is the identity matrix as in Section 5, one would have $\hat{\theta}[\mathcal{M}]_i = y_i$ for $i \in \mathcal{M}$ and $\hat{\theta}[\mathcal{M}]_i = 0$ otherwise. What is the risk of $\hat{\theta}[\mathcal{M}]$? A classical computation which we shall not reproduce here (the reader should really make sure that this is okay!) shows that the MSE obeys

$$\mathbb{E}\|\Phi\theta - \Phi\hat{\theta}[\mathcal{M}]\|^2 = \inf_{a\in V(\mathcal{M})} \|\Phi\theta - \Phi a\|^2 + \sigma^2|\mathcal{M}|. \tag{7.2}$$

Again, this has an interpretation in terms of the classical bias variance decomposition. The first term is the squared bias one gets by using only a subset of columns of Φ to approximate the true object $f = \Phi\theta$. The second term is the variance of the estimator and is simply proportional to the size of the model $\mathcal{M}$.

7.3. Ideal model selection

Just as we selected the ideal projection or keep-or-kill estimate in Section 5, we now introduce the ideal estimator $f^I = \Phi\theta^I$ which automatically selects the best model so that

$$\mathcal{R}^I(\theta, \Phi) := \inf_{\mathcal{M}} \mathbb{E}\|\Phi\theta - \Phi\hat{\theta}[\mathcal{M}]\|^2. \tag{7.3}$$

We will refer to this as the ideal risk. Note that in the case where Φ is the identity or, by extension, any orthonormal matrix, (7.3) is equal to $\sum_i \min(\theta_i^2, \sigma^2)$, which is the risk of the ideal projection we encountered earlier: compare (7.3). In the language of model selection, one would say that we have an oracle which would select for us the best model to use, *i.e.*, the best subset of explanatory variables.

Of course, if the 'true model' $f = \Phi\theta$ has coefficients θ which are very sparse, then the ideal estimator would do very well. For example, since

$$\mathcal{R}^I(\theta, \Phi) \le \mathbb{E}\|\Phi\theta - \Phi\hat{\theta}[\mathcal{M}^*]\|^2,$$

where $\mathcal{M}^*$ is the set of indices corresponding to the nonzero entries of θ, $\mathcal{M}^* := \{i : \theta_i \neq 0\}$, we have

$$\mathcal{R}^I(\theta, \Phi) \le \sigma^2\,|\mathcal{M}^*|$$

(note that the estimator $\hat{\theta}[\mathcal{M}^*]$) is unbiased). In comparison, if one uses the MLE without model selection, the risk would be equal to $n\,\sigma^2$ and hence be much larger. The conclusion is that when there are only a few nonzero parameters and we know which ones they are, we can achieve substantial risk savings.

This extends to situations where most coefficients are nonzero but relatively small, so that there is a small subset $\mathcal{M}^*$ of cardinality much smaller than n with small bias, for instance such that

$$\inf_{a \in V(\mathcal{M}^*)} \|\Phi a - \Phi\theta\|^2 \approx \sigma^2\,|\mathcal{M}^*|.$$

Then the ideal risk is bounded by

$$\inf_{a \in V(\mathcal{M}^*)} \|\Phi a - \Phi\theta\|^2 + \sigma^2\,|\mathcal{M}^*| \ll n\,\sigma^2.$$

In other words, even though there are many parameters to estimate, we can, in principle, ignore the bulk of these to achieve substantial risk savings.

Finally, and just as before, the size of the ideal risk (7.3) quantifies the precision of nonlinear approximation. We let f_m be the best m-term approximation of f, *i.e.*,

$$\|f - f_m\|^2 = \inf_{a\colon \#\{i,\, a_i \neq 0\} \leq m} \|f - \Phi a\|^2;$$

that is, it is that linear combination of at most m columns of Φ which comes closest to the object f of interest. With this notation, one can rewrite the ideal risk as

$$\inf_m \|f - f_m\|^2 + m\sigma^2, \tag{7.4}$$

which is exactly the same trade-off between the approximation error and the number of terms in the partial expansion.

7.4. Oracles and ideal risk

We have seen that one can achieve the ideal risk (7.4) with the help of an oracle and the real issue is how close one can get without. We follow Donoho and Johnstone (1995) and introduce

$$K(\Phi) = \inf_{\hat\theta} \sup_{\theta \in \mathbb{R}^p} \frac{\mathbb{E}\|\Phi\theta - \Phi\hat\theta\|^2}{\sigma^2 + \mathcal{R}^I(\theta, \Phi)}.$$

A value of $K(\Phi)$ close to one would indicate that one could mimic an oracle, while if $K(\Phi)$ were much greater than one, then one could not.

For orthonormal matrices Φ, we argued that $K(\Phi)$ obeys

$$K(\Phi) \approx 2 \log n,$$

as shown by Donoho and Johnstone (1994*a*) and Foster and George (1994). For general $n \times p$ matrices ($p \geq n$), and not necessarily orthonormal, Foster and George (1994) and Donoho and Johnstone (1995) show that $K(\Phi)$ obeys

$$K(\Phi) = O(\log p). \tag{7.5}$$

We also refer to Barron and Cover (1991), Barron (1994) and Birgé and Massart (1997) for similar results in a slightly different context. Equation (7.5) is important because it asserts that it is possible to do nearly as well as someone using an oracle.

Which estimators then mimic the oracle up to at most a logarithmic multiplicative factor? To answer this question, we take a complexity-penalized fitting approach and consider an estimator $\hat\theta$ which minimizes the functional

$$\|y - \Phi a\|^2 + \lambda^2\sigma^2 \cdot \|a\|_{\ell_0}, \tag{7.6}$$

where we recall that $\|a\|_{\ell_0} = \#\{i : a_i \neq 0\}$. In other words, our estimator $\hat\theta$ is the solution of the complexity-penalized residual sum of squares

$$\min_{\mathcal{M}} \|y - \Phi\hat\theta[\mathcal{M}]\|^2 + \lambda^2\sigma^2 \cdot |\mathcal{M}|.$$

Note that this a valid estimator since it can, at least in principle, be computed from the data y. This is the 'canonical selection procedure', to quote Foster and George (1994), and the estimator achieves the best trade-off between the goodness of fit and the complexity of the model. Popular selection procedures such as AIC, C_p, BIC and RIC are all of this form, with different values of the parameter: $\lambda^2 = 2$ in AIC (Akaike 1974, Mallows 1973), $\lambda^2 = \log n$ in BIC (Schwarz 1978), and $\lambda^2 = 2 \log p$ in RIC (Foster and George 1994).

In an unpublished manuscript, Donoho and Johnstone (1995) proved that the performance of this empirical model selection strategy obeys the following oracle inequality.

Theorem 7.1. (Donoho and Johnstone) Select $\lambda^2 = A\cdot(1+\sqrt{2\log p})^2$ where $A > 8$, and let θ be the solution to (7.6). Then

$$E\|\Phi\theta - \Phi\hat{\theta}\|^2 \leq 6\,(1 - 8/A)^{-1} \cdot \lambda^2 \cdot (\sigma^2 + \mathcal{R}^I(\theta, \Phi)). \tag{7.7}$$

The oracle inequality (7.7) is valid for all $n \times p$ matrices Φ and all θ and, therefore, empirical model selection comes within a log factor of ideal model selection.

Proof. We follow Donoho and Johnstone (1995) and sketch a proof based on complexity functionals. Without loss of generality, we may just assume the noise level $\sigma^2 = 1$ (the general follows by rescaling).

We introduce some notation and will call $K(\tilde{\theta}; y)$ the empirical complexity functional

$$K(\tilde{\theta}; y) = \|\Phi\tilde{\theta} - y\|^2 + \lambda^2\, \|\tilde{\theta}\|_{\ell_0}.$$

We make the following observations.

(1) Consider a vector θ_0, which achieves the minimum *noiseless* complexity

$$\theta_0 = \text{argmin} K(\tilde{\theta}; \Phi\theta).$$

Since $\hat{\theta}$ has minimum *noisy* complexity, $\hat{\theta}$ obeys

$$K(\hat{\theta}; y) \leq K(\theta_0; y). \tag{7.8}$$

(2) It follows from the decomposition $y = \Phi\theta + z$ that

$$\begin{aligned} K(\hat{\theta}; y) &= \|\Phi\theta - \Phi\hat{\theta}\|^2 + 2\langle z, \Phi\theta - \Phi\hat{\theta}\rangle + \|z\|^2 + \lambda^2\, \|\hat{\theta}\|_{\ell_0} \\ &= K(\hat{\theta}; \Phi\theta) + 2\langle z, \Phi\theta - \Phi\hat{\theta}\rangle + \|z\|^2. \end{aligned}$$

(3) We may develop a similar expression for $K(\theta_0; y)$, and plugging these equalities on both sides of (7.8) gives

$$K(\hat{\theta}; \Phi\theta) \leq K(\theta_0; \Phi\theta) + 2\langle z, \Phi\hat{\theta} - \Phi\theta_0\rangle. \tag{7.9}$$

Put $\hat{K} = K(\hat{\theta}; \Phi\theta)$ and $K_0 = K(\theta_0; \Phi\theta)$ for convenience. We have

$$\|\Phi\theta - \Phi\hat{\theta}\|^2 \le \hat{K}, \tag{7.10}$$

and it will therefore suffice to develop a bound on the expected value of $\hat{K}$. Now check (7.9). If we could somehow argue that the term $2\langle z, \Phi\hat{\theta} - \Phi\theta_0\rangle$ is small compared to $\hat{K}$, *e.g.*, at least a fraction of $\hat{K}$, then we would be done. This is precisely the strategy we will employ.

To achieve this goal, we let $X(k)$ be the random variable defined by

$$X(k) = \sup_{\theta_1, \theta_2} \{\langle z, \Phi\theta_2 - \Phi\theta_1\rangle,\ \|\Phi\theta_j - \Phi\theta\|^2 \le k,\ \lambda^2\|\theta_j\|_{\ell_0} \le k\}. \tag{7.11}$$

The following lemma gives a bound on the size of $X(k)$.

Lemma 7.2. Define $k_j = 2^j\,(1-8/A)^{-1}\max(K_0, \lambda^2)$ for each $j \ge 0$. Then the event

$$B_j = \{X(k) \le 4k/A\} \tag{7.12}$$

has probability at least $1 - 1/(2^j)!$.

Observe that on the event B_j, one cannot have $k \le K_0 + 2X(k)$, which automatically implies that on this event

$$\hat{K} \le k_j.$$

This property gives a bound on the expected value of $\hat{K}$ since

$$\begin{aligned} \mathbb{E}\hat{K} &\le k_0\,\mathbb{P}(\hat{K} \le k_0) + \sum_{j\ge1} k_j\,\mathbb{P}(\hat{K} \ge k_{j-1}) \\ &\le k_0 \cdot \Big(1 + \sum_{j\ge1} 2^j\mathbb{P}(B^c_{j-1})\Big). \end{aligned}$$

It follows from $\mathbb{P}(B_j^c) \le 1/(2^j)!$ that $\sum_{j\ge1} 2^j\mathbb{P}(B^c_{j-1}) \le 5$ and, therefore,

$$\mathbb{E}\hat{K} \le 6k_0.$$

In conclusion,

$$\mathbb{E}\hat{K} \le 6\,(1-8/A)^{-1}\max(\lambda^2, K_0),$$

which proves the claim since K_0 is no greater than λ^2 times the ideal risk. □

We only briefly discuss Lemma 7.2. We consider k in the range $[\ell\lambda^2, (\ell+1)\lambda^2)$ where ℓ is a fixed positive integer. Note that each feasible element for the optimization problem is a linear combination of at most $\ell = \lfloor k/\lambda^2 \rfloor$ nonzero vectors, and therefore the difference $\theta_2 - \theta_1$ is a linear combination of at most 2ℓ distinct vectors from our dictionary; we let V be the linear space of dimension at most 2ℓ spanned by those vectors and denote by P_V the orthogonal projection onto V. The Cauchy–Schwarz inequality gives

$$|\langle z, \Phi\theta_2 - \Phi\theta_1\rangle| \leq \|P_V z\| \cdot \|\Phi\theta_2 - \Phi\theta_1\| \leq 2\sqrt{k} \cdot \|P_V z\|,$$

since $\|\Phi\theta_2 - \Phi\theta_1\| \leq 2\sqrt{k}$ by assumption. The term $\|P_V z\|^2$ is a chi-squared⋆ random variable with 2ℓ degrees of freedom. The claim essentially follows from large deviation bounds for such chi-squares. Because of space limitations, we do not dwell on this issue.

7.5. *Serious limitations*

Theorem 7.1 is of theoretical importance but highly impractical. Solving (7.6) is in general NP-hard (Natarajan 1995). To the best of our knowledge, solving this problem essentially requires exhaustive searches over all subsets of columns of Φ, a procedure which is clearly combinatorial in nature and has exponential complexity since, for p of size about n, there are about 2^p such subsets. (We are of course aware that in the special case where Φ is orthogonal, the solution is simply obtained by hard-thresholding the vector $\Phi^T y$ at the level $\sqrt{\lambda}\sigma$: see Section 5.)

In other words, and quoting from Candès and Tao (2005*a*), 'solving the model selection problem might be possible only when p ranges in the few dozens. This is especially problematic when one considers that we now live in a data-driven era marked by ever larger datasets.'

In some sense, Theorem 7.1 is merely a theoretical gadget. However, it is a very important one, since it shows what is achievable by a real estimator. A crucial issue is whether there are computationally more efficient estimators with similar properties. In Section 8, we will discuss a new breed of complexity-penalized estimators with surprising properties.

7.6. *An example: recovering edges from noisy data*

Despite its computational infeasibility, Theorem 7.1 gives a precise statement about the performance of a real estimator, and Donoho and Johnstone (1995) give an example of how this might be used. We consider an image model where one tries to recover the indicator function of a smooth set (a shape, if you will)

$$f(x) = 1_B(x), \tag{7.13}$$

where we assume that the second derivative or the edge curvature ∂B is bounded by some constant R, so that one can loosely express the class of objects of interest by

$$\mathcal{F}_2(R) := \{f = 1_B : \ \|\partial B\|_{C^2} \leq R\}.$$

Such models, also known as *boundary fragment* models, have been studied extensively by Korostelëv and Tsybakov (1993) and others. Note that this class of images is neither convex nor orthosymmetric, and does not admit an unconditional basis.

We will suppose that the observations come from the two-dimensional model

$$Y(\mathrm{d}x) = f(x)\,\mathrm{d}x + \varepsilon W(\mathrm{d}x),$$

where W is a two-dimensional Wiener sheet. The problem is to recover the edges of the unknown object from the noisy data and there are many known results about this: see Korostelëv and Tsybakov (1993) and Donoho (1999) and references therein.

It is well known that a good dictionary to represent elements in $\mathcal{F}_2(R)$ is the triangle dictionary

$$\mathcal{D} = \{1_T : (x, y, z) \in [0, 1]^6\},$$

where T denotes the triangle T with vertices x, y, z. The dictionary $\mathcal{D}$ is not countable and, in fact, we shall consider a finite version $\mathcal{D}_\varepsilon$ of $\mathcal{D}$ where one restricts the vertices to belong to a two-dimensional lattice with vertical and horizontal spacing equal to ε^2 so that the cardinality of $\mathcal{D}_\varepsilon$ is polynomial in ε.

It is not really difficult to show that, for objects $f = 1_B$ in the class of interest, there is a superposition of triangles, *i.e.*,

$$f_m = \sum_{i=1}^{m} 1_{T_i}, \qquad 1_{T_i} \in \mathcal{D}_\varepsilon,$$

whose approximation error obeys

$$\|f - f_m\|^2 \le C \cdot m^{-2},$$

at least in the range where the approximation error dominates the quantization error, *i.e.*, $m^{-2} \le \varepsilon^2$. This merely follows from a first-order Taylor approximation and we skip the details. Now it can be shown that there is no dictionary with size growing at most polynomially in m that would yield better rates of convergence: see Donoho (2001) and Candès and Donoho (2000), for example.

The approximation error allows us to derive a bound on the ideal risk in the triangle dictionary since

$$\inf_m \bigl(\|f - f_m\|^2 + m\epsilon^2\bigr) \le \inf_m \bigl(C \cdot m^{-2} + \epsilon^2 m\bigr).$$

Optimizing over m gives that the ideal risk obeys

$$\text{ideal risk} \le C \cdot \epsilon^{4/3}.$$

We can then invoke the oracle inequality (7.7), together with the fact that the size of the dictionary is polynomial in ε, to show that the performance of empirical triangle selection obeys

$$E\|\hat{f} - f\|^2 \leq O(\log 1/\varepsilon) \cdot \varepsilon^{4/3}. \tag{7.14}$$

Now the risk of the empirical triangle selection is nearly optimal since one can show – by embedding appropriate hypercubes – that any estimator must obey

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_2(R)} \mathbb{E}\|f - \hat{f}\|^2 \geq c \cdot \varepsilon^{4/3}$$

and, therefore, (7.14) comes within a logarithmic factor of the minimax risk.

In addition, one could also get similar results for other degrees of smoothness of the edge curve. For example, suppose that the boundary is C^s with $1 \leq s \leq 2$. A function g is bounded in C^s with $1 \leq s \leq 2$ if the first derivative obeys

$$\sup_{t,t'} \frac{|g'(t) - g'(t')|}{|t - t'|^{s-1}} < \infty.$$

(One can then define the modulus of smoothness as the supremum of this ratio.) Then the risk of empirical triangle selection obeys

$$E\|\hat{f} - f\|^2 \leq O(\log 1/\varepsilon) \cdot \varepsilon^{2s/(s+1)}$$

while the lower bound is at least of size $c \cdot \varepsilon^{2s/(s+1)}$. (To deal with smoother edges, one would need to employ dictionaries with higher-order polycurves.)

In conclusion, we have shown that statistical near-optimality and adaptivity can hold even though there are no unconditional bases.

8. The Dantzig selector

Model selection is an especially important topic in statistics in part because of the very large number of users who are routinely fitting large linear models or designing statistical experiments. Therefore, finding computationally feasible strategies whose predictive risk comes close to that of the ideal model selection would be likely to have a large impact. This section presents some new ideas by Candès and Tao which show that this is in fact possible, at least in some special settings.

This work is concerned with a more ambitious goal than that discussed earlier. Indeed, they seek to estimate the parameter vector $\theta \in \mathbb{R}^p$ from the data

$$y = \Phi\theta + z,$$

where Φ is an $n \times p$ matrix with $p \geq n$, and $z \sim N(0, \sigma^2 I_n)$. A typical problem of this nature might be the reconstruction of an image $\theta \in \mathbb{R}^p$ with p pixels from undersampled and noisy data, *e.g.*, from its noisy and incomplete Fourier coefficients – a problem that frequently arises in medical imaging. Now, because $p \geq n$, one might wonder how this is possible. Indeed, suppose that we are in the noiseless case in which $\sigma = 0$; then, to recover θ, one would need to solve a system of linear equations *where there are more unknowns than equations.* Elementary linear algebra tells us that this is problematic. But suppose now that θ is sparse or has entries decaying like a power law, as explained in Section 6. Then this premise radically changes the problem, making the search for solutions feasible.

8.1. The noiseless case

In fact, Candès and Tao (2005*b*) showed that in the noiseless case, one could actually recover θ *exactly* by solving a linear program

$$(P_1) \qquad \min_{\tilde{\theta} \in \mathbb{R}^p} \|\tilde{\theta}\|_{\ell_1} \quad \text{subject to} \quad \Phi\tilde{\theta} = y, \tag{8.1}$$

provided that the matrix $\Phi \in \mathbb{R}^{n\times p}$ obeys a so-called *uniform uncertainty principle* (recall $\|\tilde{\theta}\|_{\ell_1} := \sum_i |\theta_i|$). That is, ℓ_1-minimization finds without error both the location and amplitudes – which we emphasize are *a priori* completely unknown – of the nonzero components of the vector $\theta \in \mathbb{R}^p$.

In detail, Candès and Tao (2005*b*) show that exact reconstruction occurs provided that sparse subsets of columns of the data matrix Φ are approximately orthonormal. For each $\mathcal{M} \subset \{1, \ldots, p\}$, we let $\Phi[\mathcal{M}]$ be the $n \times |\mathcal{M}|$ submatrix obtained by extracting the columns of Φ corresponding to those indices in $\mathcal{M}$; then they define the number δ_S as the smallest quantity obeying

$$(1 - \delta_S)\,\|c\|^2 \leq \|\Phi[\mathcal{M}]c\|^2 \leq (1 + \delta_S)\,\|c\|^2 \tag{8.2}$$

for all subsets $\mathcal{M}$ with $|\mathcal{M}| \leq S$ and coefficient sequences c. Small values of δ_S indicate that every set of columns with cardinality less than S approximately behaves like an orthonormal system. There is a related quantity $\gamma_{S,S'}$, which is the smallest quantity such that

$$|\langle \Phi[\mathcal{M}]c, \Phi[\mathcal{M}']c' \rangle| \leq \gamma_{S,S'}\,\|c\|\,\|c'\| \tag{8.3}$$

holds for all *disjoint* sets $\mathcal{M}, \mathcal{M}' \subseteq \{1, \ldots, p\}$ of cardinality less or equal to S and S', respectively. Small values of γ indicate that disjoint subsets of covariates span nearly orthogonal subspaces.

Theorem 8.1. (Candès and Tao 2005*b*) Let S be the number of entries of $\theta \in \mathbb{R}^p$ that are nonzero, and suppose that $\delta_{2S} + \gamma_{S,2S} < 1$. Then the solution $\theta^\star$ to (8.1) is exact, *i.e.*, $\theta^\star = \theta$.

This theorem is remarkable since it says that one can solve underdetermined systems of linear equations by linear programming. For instance, together with Romberg (Candès and Tao 2004, Candès, Romberg and Tao

2006), they show that one can recover exactly all kinds of sparse signals in some fixed basis from undersampled Fourier data or other types of incomplete measurements, a phenomenon now known as *compressive sampling* and with far-reaching implications. But what is more surprising is that compressive sampling extends to noisy data.

8.2. Ideal model selection

To get a sense of what might be possible, let us consider as before the least squares estimate

$$\hat{\theta}[\mathcal{M}] = \operatorname{argmin}_{a \in V(\mathcal{M})} \|y - \Phi a\|^2.$$

Since $\hat{\theta}[\mathcal{M}]$ vanishes outside $\mathcal{M}$, we have that

$$\mathbb{E}\|\theta - \hat{\theta}[\mathcal{M}]\|^2 = \|P\theta - P\hat{\theta}[\mathcal{M}]\|^2 + \sum_{i \notin \mathcal{M}} |\theta_i|^2,$$

where P is the projection on the coordinate subset $\mathcal{M}$. We then write

$$P\theta - P\hat{\theta}[\mathcal{M}] = H\,(g + z),$$

where $H = (\Phi[\mathcal{M}]^T \Phi[\mathcal{M}])^{-1} \Phi[\mathcal{M}]^T$ and $g = \Phi\theta - \Phi P\theta$. It follows that

$$\mathbb{E}\|P\theta - P\hat{\theta}[\mathcal{M}]\|^2 = \|Hg\|^2 + \sigma^2 \operatorname{Tr}((\Phi[\mathcal{M}]^T \Phi[\mathcal{M}])^{-1}).$$

However, since all the eigenvalues of $\Phi[\mathcal{M}]^T \Phi[\mathcal{M}]$ belong to the interval $[1 - \delta_{|\mathcal{M}|}, 1 + \delta_{|\mathcal{M}|}]$, we have

$$\mathbb{E}\|P\theta - P\hat{\theta}[\mathcal{M}]\|^2 \geq \frac{1}{1 + \delta_{|\mathcal{M}|}} \cdot |\mathcal{M}| \cdot \sigma^2.$$

For each set $\mathcal{M}$ with $|\mathcal{M}| \leq S$ and $\delta_S < 1$, we have

$$\mathbb{E}\|\theta - \hat{\theta}[\mathcal{M}]\|^2 \geq \sum_{i \in \mathcal{M}^c} \theta_i^2 + \frac{1}{2} |\mathcal{M}| \cdot \sigma^2.$$

If we then define the ideal estimator θ^I as

$$\theta^I = \operatorname{argmin}_{\mathcal{M}} \mathbb{E}\|\theta - \hat{\theta}[\mathcal{M}]\|^2,$$

we have shown that the ideal mean-squared error is bounded below by

$$\begin{aligned} \mathbb{E}\|\theta - \theta^I\|^2 &\geq \frac{1}{2} \min_{\mathcal{M}} \|\theta - \hat{\theta}[\mathcal{M}]\|^2 + |\mathcal{M}| \cdot \sigma^2 \\ &= \frac{1}{2} \sum_i \min(\theta_i^2, \sigma^2). \end{aligned}$$

We feel that we do not need to make further comment on the right-hand side! What we would like to know is whether there is a computationally efficient estimator which can mimic the ideal risk.

8.3. The noisy case

Assume for simplicity that the columns of Φ are normalized (there are variations to handle the general case). Then the Dantzig selector estimates θ by solving the convex program

$$\text{(DS)} \qquad \min_{\tilde{\theta} \in \mathbb{R}^p} \|\tilde{\theta}\|_{\ell_1} \quad \text{subject to} \quad \sup_{1 \le i \le p} |(\Phi^T r)_i| \le \lambda \cdot \sigma \tag{8.4}$$

for some $\lambda > 0$, and where r is the vector of residuals

$$r = y - \Phi\tilde{\theta}. \tag{8.5}$$

The solution to this optimization problem is the minimum ℓ_1 vector which is consistent with the observations. The constraints impose that the residual vector is within the noise level and does not correlate too well with the columns of Φ. We would like to mention that there exist related, yet different proposals in the literature, and most notably the lasso introduced by Tibshirani (1996).

The program (DS) is convex and can be recast as a linear program (LP)

$$\min \sum_i u_i \tag{8.6}$$

subject to

$$-u \le \tilde{\theta} \le u \qquad -\lambda\sigma\,\mathbf{1} \le \Phi^T(y - \Phi\tilde{\theta}) \le \lambda\sigma\,\mathbf{1},$$

where the optimization variables are $u, \tilde{\theta} \in R^p$, and $\mathbf{1}$ is a p-dimensional vector of ones. This is nice because linear programming is a very mature field with stable and efficient solvers. As a matter of fact, the paper by Candès and Tao (2005*a*) reports on experiments where p is in the hundreds of thousands.

The Dantzig selector is not only computationally tractable, it is also accurate.

Theorem 8.2. (Candès and Tao 2005*a*) Set $\lambda := (1 + t^{-1})\sqrt{2 \log p}$ in (8.4) and suppose that θ has S nonzero terms with $\delta_{2S} + \gamma_{S,2S} < 1 - t$. Then

$$\mathbb{E}\|\hat{\theta} - \theta\|^2 \le O(\log p) \cdot \left(\sigma^2 + \sum_i \min(\theta_i^2, \sigma^2)\right). \tag{8.7}$$

The slogan is thus that linear programming can mimic the oracle. It is worth mentioning that the oracle inequality (8.7) is not exactly the statement contained in Candès and Tao (2005*a*) where it is only shown that $\|\hat{\theta} - \theta\|^2$ is bounded by the right-hand side of (8.7) with very large probability. A minor modification of their argument, however, gives the bound on the MSE.

The assumptions are here more restrictive than in Theorem 7.1, but this is to be expected since we are looking at a more difficult problem, namely, estimating θ rather than $\Phi\theta$. For example suppose that $\delta_{2S} = 0$, which may indicate that there is a matrix $\Phi[\mathcal{M}_1 \cup \mathcal{M}_2]$ with $2S$ columns ($|\mathcal{M}_1| = S$, $|\mathcal{M}_2| = S$), which is rank-deficient. If this is the case, there is a pair of vectors $\theta_1 \in V(\mathcal{M}_1)$, $\theta_2 \in V(\mathcal{M}_2)$ with the property

$$\Phi(\theta_2 - \theta_1) = 0, \quad \Leftrightarrow \quad \Phi\theta_2 = \Phi\theta_1.$$

This is why we need $\delta_{2S} < 1$. For, otherwise, the model may not be identifiable since both θ_1 and θ_2 have at most S nonzero entries. The condition $\delta_{2S} + \gamma_{S,2S} < 1$ (or less than $1 - t$) is only slightly stronger than the identifiability condition.

There are other versions of Theorem 8.2 which only require θ to be sparse in the sense that many of its entries are small but not necessarily zero, *e.g.*, θ may belong to a weak-ℓ_p ball for some $p > 0$: see Candès and Tao (2005*a*) for details. In addition, the Dantzig selector is a kind of soft-thresholding estimator and therefore has the tendency to underestimate the true value of θ. The aforementioned reference details simple versions which correct for the bias and have better practical performance.

8.4. Comparison with the combinatorial search

For sufficiently sparse vectors the near-orthogonality property (8.2) of the matrix Φ shows that

$$\|\Phi(\theta - \hat{\theta})\| \asymp \|\theta - \hat{\theta}\|$$

where $\asymp$ means that the ratio is bounded above and below. Thus, one can recast (8.7) as

$$\mathbb{E}\|\Phi\theta - \Phi\hat{\theta}\|^2 \le O(\log p) \cdot (\sigma^2 + \mathcal{R}^I(\theta, \Phi)). \tag{8.8}$$

Like the 'combinatorial search estimator' (7.6), the Dantzig selector comes within a logarithmic factor of the ideal risk (7.3).

The catch, however, is that although the hypotheses of Theorem 8.2 are in some sense necessary to estimate θ accurately, they are probably too restrictive when one is 'only' interested in estimating $\Phi\theta$. For instance, Theorem 7.7 does not assume anything about the matrix Φ and about the sparsity of the true vector $\theta \in \mathbb{R}^p$. It is likely that the Dantzig selector would also obey (8.8) under more general conditions. As a matter of fact, we regard as extremely significant the problem of deciding whether or not there is – under mild conditions – a computationally tractable estimator mimicking the oracle.

9. Frames and libraries

Getting back to the familiar framework of thresholding, it is important to realize that thresholding can be successful even outside the specific case where one is given a single orthobasis. In this section we discuss two cases in which thresholding is highly effective even though there is no (single) orthobasis.

9.1. Tight frames

In harmonic analysis, it is generally much easier to construct a tight frame than an orthobasis. In $\mathbb{R}^n$, a tight frame is a collection of vectors (φ_i) with the property

$$\|f\|^2 = \sum_i |\langle f, \varphi_i\rangle|^2. \tag{9.1}$$

If we arrange the vectors φ_i as the columns of a matrix Φ, then this property may be expressed as

$$\|\Phi^T f\|^2 = \|f\|^2,$$

which says that Φ^T is an isometry. The isometry property provides a simple reconstruction formula from the frame coefficients $(\langle f, \varphi_i\rangle)$ since $\Phi\Phi^T = I_n$, or equivalently

$$f = \sum_i \langle f, \varphi_i\rangle \varphi_i. \tag{9.2}$$

The only difference between (9.1)–(9.2) and an orthobasis is that the elements φ_i may not be linearly independent. In particular, we may have more elements than the dimension of the space. In general, a tight frame is a collection of vectors taken from a Hilbert space obeying (9.1). For example, we have tight frames in $L_2(\mathbb{R})$, $L_2(\mathbb{R}^2)$, and so on, where the inner product is of course the usual inner product over square integrable functions.

The exact orthogonality between elements is what can make the construction of orthobases extremely challenging. In contrast, one has more flexibility in constructing tight frames and this is why this is easier. For instance, while tight Gabor frames exist, Balian and Low have shown that it is impossible to find an orthonormal equivalent with nice time-frequency localization properties (there are orthobases of local cosines but this is somewhat different): see Mallat (1999). Also, Candès and Donoho (2004) have constructed nice tight frames of *curvelets* and it is not known whether one can construct an orthonormal equivalent with nice time-frequency localization properties.

Suppose that we observe $y \sim N(f, \sigma^2 I_n)$; then we can define the empirical frame coefficients $\tilde{y} = \Phi^T y$ which obey the Gaussian model

$$\tilde{y}_i = \theta_i + \tilde{z}_i, \tag{9.3}$$

where $\tilde{z}$ is a Gaussian process with zero mean and covariance matrix

$$\mathrm{Cov}(\tilde{z}_i, \tilde{z}_j) = \sigma^2 \langle \varphi_i, \varphi_j \rangle.$$

In particular, the variance of $\tilde{z}_i$ obeys $\mathrm{Var}(z_i) = \sigma^2 \|\varphi_i\|^2$ which we denote by σ_i^2. The situation is analogous in the continuous white-noise model where the empirical coefficients are defined by $\tilde{y}_i = \int \varphi_i(t)\, Y(\mathrm{d}t)$ giving an infinite-dimensional version of the sequence model (9.3) (the covariance is $\varepsilon^2 \langle \varphi_i, \varphi_j \rangle$). Also note that, since $\|\varphi_i\| \le 1$, we have $\sigma_i \le \sigma$ and

$$\sum_i \sigma_i^2 = \mathbb{E}\|\tilde{z}\|^2 = \mathbb{E}\|\Phi^T z\|^2 = \mathbb{E}\|z\|^2 = n\sigma^2.$$

One can of course apply individual thresholding in a tight frame. Suppose we are in the sampled model with n observations. We have seen in Section 5 that the risk of a thresholding rule, with threshold $\sqrt{2\log n} \cdot \sigma_i$, obeys

$$\mathbb{E}\|\theta_i - \hat{\theta}_i\|^2 \le (2\log n + 1) \cdot (\sigma_i^2/n + \min(\theta_i^2, \sigma_i^2))$$

and therefore

$$\mathbb{E}\|\theta - \hat{\theta}\|^2 \le (2\log n + 1) \cdot \left(\sigma^2 + \sum_i \min(\theta_i^2, \sigma_i^2)\right).$$

Returning to the original domain gives an estimator $\hat{f} = \sum_i \hat{\theta}_i \varphi_i$ obeying

$$\mathbb{E}\|f - \hat{f}\|^2 = \mathbb{E}\|\Phi\theta - \Phi\hat{\theta}\|^2 \le \mathbb{E}\|\theta - \hat{\theta}\|^2,$$

where we have used the fact that, for any vector h, $\|\Phi h\| \le h$. It then follows that the performance of the shrinkage estimator is bounded by

$$\mathbb{E}\|f - \hat{f}\|^2 \le (2\log n + 1) \cdot \left(\sigma^2 + \sum_i \min(\theta_i^2, \sigma_i^2)\right). \tag{9.4}$$

The message is of course that, *if the frame coefficient sequence* is sparse, then this strategy is highly effective.

We emphasized the 'frame coefficient sequence' for a reason. There are many ways to expand a signal or a vector in a frame, and depending upon the frame, the frame decomposition may be dense while there may exist other very sparse decompositions. We give an example. Suppose that the frame is composed of two orthobases $\Phi = [\Phi_1, \Phi_2]/\sqrt{2}$ where each Φ_j is an $n \times n$ orthonormal matrix. To make things concrete, suppose Φ is the time-frequency dictionary where Φ_1 is the identity matrix and Φ_2 is the unitary discrete Fourier matrix. Now consider a signal f made out of one spike

$$f = (\mu, 0, \dots, 0),$$

where μ is some large amplitude. Then f is a multiple of a single column of Φ and the ideal risk (Section 7) is simply equal to σ^2. Now, for each i, $|\Phi_2^T f|_i = \mu/\sqrt{n}$ and, if the amplitude of the spike is large enough, then all the Fourier coefficients will exceed the noise level. Applying the thresholding estimator and using the proxy (5.7), we would not expect anything substantially better than

$$\frac{2\log n + 1}{2} \cdot (n+1) \cdot \frac{\sigma^2}{2},$$

which is horrible since there is only one parameter to estimate!

9.2. The curvelet shrinkage

Candès and Donoho recently introduced tight frames of curvelets to overcome inherent limitations of traditional multiscale representations such as wavelets (Candès and Donoho 2000, Candès and Guo 2002, Candès and Donoho 2004). Conceptually, the curvelet transform is a multiscale pyramid with many directions and positions at each length scale, and needle-shaped elements at fine scales. This pyramid is nonstandard, however, as curvelets have useful geometric features that set them apart from wavelets and the like. For instance, curvelets obey a parabolic scaling relation which says that at scale 2^{-j}, each element has an envelope which is aligned along a 'ridge' of length $2^{-j/2}$ and width 2^{-j}. It is beyond the scope of this paper to discuss this new construction and we refer to Candès and Donoho (2004) for mathematical details and to Candès, Demanet, Donoho and Ying (2005) for the description of fast and accurate digital curvelet transform algorithms.

Curvelets are interesting because they efficiently address very important problems where wavelet ideas are far from ideal. Of interest here is that curvelets provide optimally sparse representations of objects which display *curve-punctuated smoothness* – smoothness except for discontinuity along a general curve with bounded curvature. Such representations are nearly as sparse as if the object were not singular and turn out to be far more sparse than the wavelet decomposition of the object.

Quantitatively speaking, let (θ_i) denote the curvelet coefficient sequence of a C^2 function with piecewise C^2 singularities (edges). Then Candès and Donoho (2004) showed that the nth largest entry $|\theta|_{(n)}$ in the sequence obeys

$$|\theta|_{(n)} \le C \cdot n^{-3/2} (\log n)^{3/2}, \quad \text{for all} \;\; n > 0. \tag{9.5}$$

This decay is optimal: among all possible representations of objects with singularities, this is essentially the sparsest one. That is, there is no basis, tight frame, frames and so on in which the coefficients of a function f with piecewise C^2 edges would have a faster decay.

Of course, the enhanced sparsity shows that one can recover such objects from noisy data by simple curvelet shrinkage and obtain an MSE order of magnitude better than that achieved by more traditional methods, *e.g.*, wavelet shrinkage. Omitting details having to do with the definition of the thresholding zone (Candès and Donoho 2002), one can then plug the estimate (9.5) into the oracle inequality and obtain that the risk obeys

$$\mathbb{E}\|f - \hat{f}\|^2 \le O(\log^2 \varepsilon^{-1}) \cdot \varepsilon^{4/3}.$$

(Recall that the minimax lower bound exceeds $c \cdot \varepsilon^{4/3}$.) It goes without saying that we do not need to solve an intractable problem (like empirical triangle selection) to recover a smooth image with edges from noisy data in an optimal fashion. Instead, one can just go into the curvelet domain (by means of the fast digital curvelet transform), throw out the small coefficients and invert the transform.

9.3. Statistical estimation in a library of bases

Suppose now that we are given a library $\mathcal{L}$ of orthonormal bases

$$\mathcal{L} = \{\mathcal{B}_1, \dots, \mathcal{B}_L\},$$

where the $\mathcal{B}_i$s are L distinct orthonormal bases. For example, the library $\mathcal{L}$ might be a concatenation of several orthonormal bases, *e.g.*, the canonical basis (or the spike basis, as it is called in signal processing), the Fourier basis, a wavelet basis, a spline basis, a ridgelet basis (Candès and Donoho 1999) and so on. Or the library $\mathcal{L}$ might be the cosine, the wavelet (Coifman and Meyer 1991) or the ridgelet packet library (Flesia, Hel-Or, Averbuch, Candès, Coifman and Donoho 2003). We would like to emphasize that we consider libraries of orthonormal bases for simplicity but the results extend to libraries of tight frames (see Candès (2002)), so that it is possible to include the aforementioned curvelets, contourlets, and many other recent interesting constructions in computational harmonic analysis.

We wish to explore the possibility of adaptive basis estimation. Suppose that we observe a signal in white noise. Adaptive basis estimation means that we would like to select, based on the data, the best basis in which to estimate the signal; that is, the basis in which the true unknown signal is in some sense the sparsest possible.

We let $y_i[\mathcal{B}]$ be the coordinates of the observations in the basis $\mathcal{B}$ and, likewise, we let $\theta_i[\mathcal{B}]$ and $z_i[\mathcal{B}]$ be the coordinates of the signal f and of the error vector in $\mathcal{B}$. In the basis $\mathcal{B}$, our statistical model is of the form

$$y_i[\mathcal{B}] = \theta_i[\mathcal{B}] + z_i[\mathcal{B}],$$

and the ideal risk in that basis $\mathcal{B}$ is

$$\mathcal{R}^I(\theta, \mathcal{B}) = \sum_i \min(|\theta_i[\mathcal{B}]|^2, \sigma^2).$$

We now introduce the ideal risk in the library as the minimum over all bases in the library

$$\mathcal{R}^I(\theta, \mathcal{L}) = \min_{\mathcal{B} \in \mathcal{L}} \mathcal{R}^I(\theta, \mathcal{B}). \tag{9.6}$$

This ideal risk is achievable with the aid of (1) a *basis* oracle which selects the best basis and (2) a *coordinate* oracle which tells us which coordinates in that basis are worth estimating.

The issue is then whether one can select a basis in a near-ideal fashion from the data alone. In order to do this, Donoho and Johnstone (1994*b*) introduce the entropy functional

$$\mathcal{E}_\lambda(y, \mathcal{B}) := \sum_i \min(|y_i[\mathcal{B}]|^2, \lambda^2 \, \sigma^2),$$

where λ is a parameter. This quantity is not surprising since this is none other than the empirical complexity functional (7.6) in the basis $\mathcal{B}$

$$\mathcal{E}_\lambda(y, \mathcal{B}) := \min_a \|y[\mathcal{B}] - a\|^2 + \lambda^2 \, \sigma^2 \|a\|_{\ell_0}.$$

It then seems sensible to choose the basis for estimation in which $\mathcal{E}_\lambda(y, \mathcal{B})$ is smallest. The estimation strategy consists of two simple stages.

(1) We select $\hat{\mathcal{B}}$ as the best orthobasis $\hat{\mathcal{B}}$ according to the entropy

$$\hat{\mathcal{B}} := \text{argmin} \mathcal{E}_\lambda(y, \mathcal{B}).$$

(2) We then apply hard-thresholding (with level $\lambda \, \sigma$) in that basis so that

$$\hat{\theta}_i[\hat{\mathcal{B}}] = \begin{cases} y_i[\hat{\mathcal{B}}], & |y_i| > \lambda \, \sigma, \\ 0, & \text{otherwise.} \end{cases}$$

The result is that if λ is correctly tuned, empirical basis selection nearly achieves the performance of the ideal estimator.

Theorem 9.1. (Donoho and Johnstone 1994*b*) Let M_n be the number of distinct vectors in the library and set $\lambda^2 = A(1 + \sqrt{2 \log M_n})^2$ for some $A > 8$. Then

$$\mathbb{E}\|\hat{\theta}[\hat{\mathcal{B}}] - \theta[\mathcal{B}]\|^2 \leq 6(1 - 8/A)^{-1} \cdot \lambda^2 \cdot (\sigma^2 + \mathcal{R}^I(\theta, \mathcal{L})). \tag{9.7}$$

If there is an efficient basis for estimation, then empirical basis selection will find it and the error of estimation will be small.

The reader is right to suspect that the proof of Theorem 9.1 is based on minimum complexity functionals and is nearly identical to that of Theorem 7.1 and we will, therefore, not reproduce it.

An interesting example concerns denoising in a packet library such as cosine or wavelet packets. In a cosine packet library, for instance, there are about $n \log_2 n$ distinct elements where n is the number of samples, while the number of orthobases is equal to the number of dyadic trees of depth about $\log_2 n$, which is exponential in n. This looks daunting as one would naively think that one would need to evaluate exponentially many entropy functionals in order to find the best basis. Fortunately, because of the additivity property of the entropy functional and of the tree structure of the library of bases, there is a way to invoke dynamic programming to select the best basis. In particular, Coifman and Wickerhauser (1992) show that one can compute $\hat{\mathcal{B}}$ in $O(n)$. Since all the noisy coefficients in the library (there are about $n \log_2 n$ of them) can be computed in $O(n \log^2 n)$, the empirical best basis estimator can be rapidly computed.

10. Further topics

In this last section, we discuss a selection of other important problems and topics which we hope will give an idea of how broad the field really is.

10.1. From theory to practice

We have not talked much about the practical performance of shrinkage ideas in signal and image processing. Wavelet shrinkage ideas have indeed been deployed with great success in many applications, and are nowadays routinely used by researchers and engineers. We mention here a few topics which enhance the estimation.

Thresholding rules in a wavelet basis are known to produce some artifacts, some of which may be removed by applying a translation-invariant type of shrinkage. For example, a frequently discussed approach consists of applying cycle spinning. Cycle spinning is a kind of translation-invariant thresholding rule: this technique computes several individual reconstructions by applying shifts to the noisy data and averages them out, after applying the reverse shifts, of course. Another popular approach consists in applying thresholding in a redundant wavelet representation, such as the undecimated wavelet transform; see the '*à trous*' algorithm in Starck, Murtagh and Bijaoui (1998). The basic idea underlying these methods is that an average of similar-looking estimators produces visually more pleasing results than any of the individual estimators taken individually.

Researchers have also developed the idea of 'block thresholding', which originates in Efroĭmovich (1985). Instead of treating each coefficient individually, the idea is that the statistical properties of images may be used to group coefficients together to better inform the decision. For example, if a wavelet coefficient is large, it may indicate the presence of an edge and, therefore, some of the neighbouring coefficients are likely to be large as well. There are many variations on this theme and we will not attempt to define these strategies. We shall instead simply mention that block thresholding works well empirically and is also amenable to rigorous analysis. We refer the reader to Cai (1999) and Hall, Kerkyacharian and Picard (1999) for experimental and theoretical results in this direction.

In a different direction, several authors (Candès and Guo 2002, Malgouyres 2002, Durand and Froment 2003) have independently proposed an attractive alternative to single basis thresholding. The idea here is to combine basis function expansions with variational principles for the reconstruction of an image/signal whose coefficients (in some basis) are known only approximately: they might be noisy, quantized, and so on. In the denoising problem where one wishes to recover an object f from $y = f + z$, one could imagine solving the following problem:

$$\min \|g\|_{TV} \quad \text{subject to} \quad |\Phi^T(g-y)|_i \le \lambda\,\sigma \quad \text{for all} \;\; i, \tag{10.1}$$

where Φ^T is the transform of interest (*e.g.*, the wavelet transform), $(\Phi^T f)_i = \langle f, \varphi_i \rangle$. Here, the total variation norm $\|g\|_{\mathrm{BV}}$ measures the complexity of the fit and is roughly equal to the integral of the Euclidean norm of the gradient. The aforementioned references demonstrate that this procedure works extremely well. Thresholding rules tend to produce artificial oscillations near discontinuities even though the original signal/image may be flat on both sides of the discontinuity, a ‘pseudo-Gibbs phenomenon’. Ideas like (10.1) are very effective at removing such artifacts while retaining other nice properties of shrinkage methods.

In closing, shrinkage methods have inspired a lot of activity and new methods have been tuned to achieve the best practical performance.

10.2. Inverse problems

Another interesting problem occurs when one cannot measure the object $f(t)$ directly, but can only make linearly distorted measurements. That is, we are only able to observe data about $g(u) = Kf(u)$, where K is a linear transform. Such problems arise in multiple scientific settings ranging from medical imaging to physical chemistry to extragalactic astronomy. For example, in the case where K is a convolution transform, the signal is blurred as one measures

$$g(t) = (k * f)(t),$$

where k is a convolution kernel. Recovering blurred images from noisy data is ubiquitous in science and engineering: see Bertero and Boccacci (1998) for a nice survey. Another problem which has received a lot of attention concerns the case where K is the Radon transform

$$g(t,\theta) = \int_{\mathcal{L}_{t,\theta}} f(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2,$$

where for $\theta \in [0, 2\pi)$ and $t \in \mathbb{R}$, $\mathcal{L}_{t,\theta}$ is the line

$$\{x_1 \cos\theta + x_2 \sin\theta = t\}.$$

Recovering an image from its two-dimensional noisy projections (line integrals) is the subject of computed tomography, which has been and still is the focus of intense research. Most interesting problems are ill-posed in the sense that the singular values of K tend to zero (think about a deconvolution problem where the convolution kernel k 'blocks' the high-frequency content of the signal).

Suppose then that we observe y of the form

$$y = Kf + z, \tag{10.2}$$

where z is white noise and f is the object we wish to recover. Suppose we are given an orthobasis or a tight frame (φ_i) for functions 'living' in the object space. Then, under certain conditions, one can define dual basis elements (ψ_i), which 'live' in the data space and obey the relation

$$[Kf, \psi_i] = \delta_i \, \langle f, \varphi_i \rangle, \tag{10.3}$$

where, in the above display, we have used the notation $[\cdot, \cdot]$ to distinguish between the data and the object spaces. Here the δ_is are defined by properties of K and called quasi-singular values; if φ_i is an orthobasis, they are set in such way that $\|\psi_i\| = 1$ (if (φ_i) is a tight frame, we could impose $\|\varphi_i\| = \|\psi_i\|$). The quasi-singular value relation (10.3) expresses the idea that one can measure the coefficients of f from Kf. Suppose that the δ_is do not vanish, then a consequence of the identity $f = \sum \langle f, \varphi_i \rangle \varphi_i$ and (10.3) is the reconstruction formula

$$f = \sum_i \delta_i^{-1} [Kf, \psi_i] \varphi_i. \tag{10.4}$$

This formula is what Donoho calls a biorthogonal decomposition of K; see Donoho (1995) or the wavelet–vaguelette decomposition (WVD) in the case when (ϕ_i) is a wavelet basis. It is an extension of the SVD decomposition which reads

$$f = \sum d_i^{-1} [Kf, h_i] e_i, \tag{10.5}$$

where (d_i^2) and (e_i) are the eigenvalues and eigenfunctions of K^*K, $K^*Ke_i = d_i^2 e_i$, and where h_i is the image of e_i under K, $Ke_i = d_i h_i$. (The ill-posedness means that $d_i \to 0$.)

The point is that many of the tools and ideas we have seen before apply. To make this connection, consider the sequence space version of (10.2), namely,

$$[y, \psi_i] = [Kf, \psi_i] + [z, \psi_i],$$

which one can write as

$$y_i = \delta_i \theta_i + [z, \psi_i]$$

(recall that $\theta_i = \langle f, \varphi_i \rangle$ are the coordinates of f we wish to estimate).

Dividing the above display by δ_i shows that we wish to recover the mean of a Gaussian vector

$$\tilde{y}_i = \theta_i + \sigma_i z_i, \tag{10.6}$$

where $\sigma_i = \sigma\|\psi_i\|$ and the z_is are $N(0,1)$ (the covariance matrix is given by $\operatorname{Cov}(z_i, z_j) = [\psi_i, \psi_j]/(\|\psi_i\|\,\|\psi_j\|)$). The only real difference is that the noise is now heteroscedastic* with σ_i increasing as the quasi-singular values are decreasing.

One can thus see that everything should generalize nicely. In particular, if we apply thresholding, the proxy for the mean-squared error will be

$$\sum_i \min(\theta_i^2, \sigma_i^2), \tag{10.7}$$

and this approach will be very effective if the following two conditions hold: (1) the signal is sparse in the basis (φ_i) and (2) the z_is in (10.6) are not too correlated, so that treating each coefficient individually still makes sense. We note that the latter condition is equivalent to saying that the system (φ_i) nearly diagonalizes the Gram matrix K^*K; by near-diagonalization, we mean that the representation of K^*K in the system (φ_i) is sparse.

The challenge for applied harmonic analysts is then to construct representations which sparsely represent objects of scientific interest and, *at the same time*, sparsely represent the operators under study. This is precisely what multiscale systems such as wavelets and curvelets achieve. On the one hand, they provide sparse representations of convolutions, Radon transforms, and many other types of common operators, and on the other, they simultaneously provide sparse representations of objects allowing for point-like singularities (wavelets) and curve-like singularities (curvelets). This is the reason why they have proved to be useful for solving inverse problems (Donoho 1995, Candès and Donoho 2002). In two dimensions, for instance, there is a quantitative theory showing that, for certain kind of interesting models of images, simple algorithms based on the shrinkage of curvelet biorthogonal decompositions achieve near-optimal statistical rates of convergence (Candès and Donoho 2002).

On the other hand – and this is very important – if one employs instead the singular system (e_i) for estimation, as is common, then the MSE may be very large. The proxy (10.7) lets us understand why this is the case. For the MSE to be small, the signal must be concentrated in the coordinates where the eigenvalues are large. But this is not usually the case, and the MSE is large. For example, in deconvolution problems, tomography problems and many others, the eigenvectors e_i are sinusoids, at least roughly speaking. The problem is that sinusoids provide very poor partial reconstructions of the kinds of signals and images in which one is typically interested: *e.g.*, images of the brain or the interior of the earth all have edges and perhaps other types of singularities. As a consequence, SVD-based methods tend to underperform when the object we wish to image is not smooth.

10.3. FDR thresholding rules

The 'universal' threshold of $\sqrt{2\log n}$ is often criticized because it is very conservative; it potentially sets to zero many coordinates where the signal is larger than the noise level. We close this paper by discussing innovative adaptive choices of thresholds which have their origin in the field of hypotheses testing – in multiple comparisons, to be more exact.

Consider the simpler problem of deciding, for each $i = 1, \dots, n$, whether or not $\theta_i = 0$, given the data

$$y_i = \theta_i + z_i, \quad z_i \text{ i.i.d. } N(0, \sigma^2).$$

Formally, we wish to simultaneously test n hypotheses

$$\begin{aligned} H_{0,i}: &\quad \theta_i = 0, \\ H_{1,i}: &\quad \theta_i \neq 0. \end{aligned}$$

Then one could accept the ith null hypothesis if $|y_i| \le \sigma\sqrt{2\log n}$ and reject it otherwise. This would essentially correspond to the Bonferroni procedure which controls the so-called familywise error rate, defined as the probability of rejecting at least one hypothesis $H_{i,0}$ which is true. If we want a familywise error rate below α, the Bonferroni method would ask us to reject $H_{i,0}$ if and only if

$$|y_i| > \sigma\, z(\alpha/2n),$$

where $z(\alpha)$ is the upper quantile of the Gaussian distribution ($z(\alpha)$ is defined by $\mathbb{P}(N(0,1) > z(\alpha)) = \alpha$). For nearly all reasonable levels α and n large, $z(\alpha/2n)$ is nearly equal to $\sqrt{2\log n}$.

In the problem of multiple comparisons, control of the familywise error rate yields very conservative decisions. Ten years ago, Benjamini and Hochberg (1995) introduced an alternative, and instead proposed to control the false discovery rate (FDR). The FDR is the expected ratio between the number of incorrectly rejected null hypotheses and the total number of rejections. The advantage is that FDR controlling procedures have greater power to detect alternatives. In our problem, we order the values by decreasing order of magnitude $|y|_{(1)} \ge |y|_{(2)} \ge \cdots \ge |y|_{(n)}$, and define i_{FDR} to be the largest index for which

$$|y|_{(i)} \ge \sigma\, z(q\, i/2n).$$

Then the procedure which rejects all the hypotheses corresponding to the i_{FDR} largest values of $|y_i|$ controls the FDR at level q (meaning that the expected proportion of false rejections is less than q).

A little later, Abramovich and Benjamini (1996) proposed applying FDR for estimation and introduced a new thresholding rule. The idea is simply to estimate the parameters corresponding to the rejected hypotheses (these are judged estimable) and set the others to zero. With $\lambda_{\rm FDR} = z(q\, i_{\rm FDR}/2n)$, the FDR thresholding rule is thus defined by

$$\hat{\theta}_i = \begin{cases} y_i, & |y_i| > \lambda_{\rm FDR}\, \sigma, \\ 0, & \text{else}. \end{cases} \tag{10.8}$$

This is interesting because (10.8) is a data-driven thresholding rule which adapts to the sparsity of the signal. The threshold is larger for sparser signals and smaller for denser ones.

To understand why FDR thresholding rules are a good thing, suppose that by looking at y we learn that many of the coordinates θ_i are nonzero. Then the FDR threshold will be lower than the universal threshold and the estimator will have a smaller bias. Of course, we will also occasionally estimate some θ_is which are close to zero, hence increasing the variance a little. But the proportion of 'erroneous' estimations is controlled, and in the bias + variance trade-off we will typically draw significantly ahead of universal thresholding rules. There are numerical experiments showing that FDR thresholding rules perform very well: see Abramovich and Benjamini (1996) and Abramovich, Benjamini, Donoho and Johnstone (2000). There is also a beautiful theory showing that, in some special set-ups where θ belongs to a weak-ℓ_p ball, for example, the estimator achieves adaptive asymptotic minimaxity (Abramovich *et al.* 2000).

FDR thresholding rules are a nice new chapter in the history of thresholding and we suspect that they will generate a lot of interest in the near future. There are also challenging questions that do not have satisfactory answers at the moment. For example, how would FDR thresholding rules adapt when the observations are correlated and how would one use them in more sophisticated estimation problems?

10.4. Last words

Near the beginning of this article, we emphasized that we would focus on a couple of key ideas that have had a very significant impact on my professional development and on the field in general. A large fraction of this paper is a write-up of a series of lectures I delivered in 2004, and the whole manuscript was conceived with the goal of teaching this material to nonspecialists. It is not an exhaustive survey of all the research that occurred in the field, and I hope that this personal selection of topics will not be found offensive.

Last but not least, I would like to thank Carl for encouraging me to write this article.

REFERENCES

F. Abramovich and Y. Benjamini (1996), ‘Adaptive thresholding of wavelet coefficients’, *Comput. Statist. Data Anal.* **22**, 351–361.

F. Abramovich, Y. Benjamini, D. L. Donoho and I. M. Johnstone (2000), Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19, Department of Statistics, Stanford University. To appear in *Ann. Statist.*

H. Akaike (1974), ‘A new look at the statistical model identification’, *IEEE Trans. Automatic Control* **AC-19**, 716–723.

A. R. Barron (1994), ‘Approximation and estimation bounds for artificial neural networks’, *Machine Learning* **14**, 113–143.

A. R. Barron and T. M. Cover (1991), ‘Minimum complexity density estimation’, *IEEE Trans. Inform. Theory* **37**, 1034–1054.

Y. Benjamini and Y. Hochberg (1995), ‘Controlling the false discovery rate: A practical and powerful approach to multiple testing’, *J. Roy. Statist. Soc. Ser. B* **57**, 289–300.

M. Bertero and P. Boccacci (1998), *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, Bristol.

L. Birgé and P. Massart (1997), From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*, Springer, New York, pp. 55–87.

L. D. Brown and M. G. Low (1996), ‘Asymptotic equivalence of nonparametric regression and white noise’, *Ann. Statist.* **24**, 2384–2398.

T. T. Cai (1999), ‘Adaptive wavelet estimation: A block thresholding and oracle inequality approach’, *Ann. Statist.* **27**, 898–924.

E. J. Candès (2002), Multiscale chirplets and near-optimal recovery of chirps. Technical report, Stanford University.

E. J. Candès and D. L. Donoho (1999), ‘Ridgelets: The key to higher-dimensional intermittency?’, *Phil. Trans. R. Soc. Lond. A* **357**, 2495–2509.

E. J. Candès and D. L. Donoho (2000), Curvelets: A surprisingly effective nonadaptive representation for objects with edges, in *Curves and Surfaces* (C. R. A. Cohen and L. L. Schumaker, eds), Vanderbilt University Press, Nashville, TN, pp. 105–120.

E. J. Candès and D. L. Donoho (2002), ‘Recovering edges in ill-posed inverse problems: Optimality of curvelet frames’, *Ann. Statist.* **30**, 784 –842.

E. J. Candès and D. L. Donoho (2004), ‘New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities’, *Comm. Pure Appl. Math.* **57**, 219–266.

E. J. Candès and F. Guo (2002), ‘New multiscale transforms, minimum total variation synthesis: Applications to edge-preserving image reconstruction’, *Signal Processing* **82**, 1519–1543.

E. J. Candès and T. Tao (2004) Near-optimal signal recovery from random projections and universal encoding strategies. Available on the ArXiv preprint server: `math.CA/0410542`. To appear in *IEEE Trans. Inform Theory.*

E. J. Candès and T. Tao (2005*a*), The Dantzig selector: Statistical estimation when p is much larger than n. Technical report, California Institute of Technology, available on the ArXiv preprint server: `math.ST/0506081`. To appear in *Ann. Statist.*

E. J. Candès and T. Tao (2005*b*), 'Decoding by linear programming', *IEEE Trans. Inform. Theory* **51**, 4203–4215.

E. J. Candès, L. Demanet, D. L. Donoho and L. Ying (2005), Fast discrete curvelet transforms. Technical report, California Institute of Technology. Submitted to *SIAM J. Multiscale Modeling and Simulations.*

E. J. Candès, J. Romberg and T. Tao (2006) 'Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information', *IEEE Trans. Inform. Theory* **52**, 489–509.

A. Cohen, R. DeVore, P. Petrushev and H. Xu (1999), 'Nonlinear approximation and the space BV($\mathbf{R}^2$)', *Amer. J. Math.* **121**, 587–628.

R. R. Coifman and Y. Meyer (1991), 'Remarques sur l'analyse de Fourier à fenêtre', *C. R. Acad. Sci. Paris Sér. I Math.* **312**, 259–261.

R. R. Coifman and M. V. Wickerhauser (1992), 'Entropy-based algorithms for best basis selection', *IEEE Trans. Inform. Theory* **38**, 713–718.

D. L. Donoho (1993), 'Unconditional bases are optimal bases for data compression and for statistical estimation', *Appl. Comput. Harmon. Anal.* **1**, 100–115.

D. L. Donoho (1995), 'Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition', *Appl. Comput. Harmon. Anal.* **2**, 101–126.

D. L. Donoho (1999), 'Wedgelets: Nearly-minimax estimation of edges', *Ann. Statist.* **27**, 859–897.

D. L. Donoho (2001), 'Sparse components of images and optimal atomic decomposition', *Constr. Approx.* **17**, 353–382.

D. L. Donoho and I. Johnstone (1994*a*), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrika* **81**, 425–455.

D. L. Donoho and I. M. Johnstone (1994*b*), 'Ideal denoising in an orthonormal basis chosen from a library of bases', *CR Acad. Sci. Paris Sér. I Math.* **319**, 1317–1322.

D. L. Donoho and I. M. Johnstone (1995), Empirical atomic decomposition. Manuscript.

D. L. Donoho and I. M. Johnstone (1999), 'Asymptotic minimaxity of wavelet estimators with sampled data', *Statist. Sinica* **9**, 1–32.

D. L. Donoho and R. C. Liu (1991), 'Geometrizing rates of convergence II, III', *Ann. Statist.* **19**, 633–667, 668–701.

D. L. Donoho and M. Nussbaum (1990), 'Minimax quadratic estimation of a quadratic functional', *J. Complexity* **6**, 290–323.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian and D. Picard (1995), 'Wavelet shrinkage: Asymptopia?', *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.

S. Durand and J. Froment (2003), 'Reconstruction of wavelet coefficients using total variation minimization', *SIAM J. Sci. Comput.* **24**, 1754–1767 (electronic).

S. Y. Efroĭmovich (1985), 'Nonparametric estimation of a density of unknown smoothness', *Teor. Veroyatnost. i Primenen.* **30**, 524–534.

S. Y. Efroĭmovich and M. S. Pinsker (1981), 'Estimation of square-integrable density on the basis of a sequence of observations', *Problemy Peredachi Informatsii* **17**, 50–68.

S. Y. Efroĭmovich and M. S. Pinsker (1982), 'Estimation of square-integrable probability density of a random variable', *Problems Inform. Transmission* **18**, 175–189; translated from *Problemy Peredachi Informatsii* **18**, 19–38 (in Russian).

S. Y. Efroĭmovich and M. S. Pinsker (1984), 'A self-training algorithm for nonparametric filtering', *Avtomat. i Telemekh.* (11), 58–65.

B. Efron and C. Morris (1971), 'Limiting the risk of Bayes and empirical Bayes estimators I: The Bayes case', *J. Amer. Statist. Assoc.* **66**, 807–815.

A. G. Flesia, H. Hel-Or, A. Averbuch, E. J. Candès, R. R. Coifman and D. L. Donoho (2003), Digital implementation of ridgelet packets, in *Beyond wavelets*, Vol. 10 of *Stud. Comput. Math.*, Academic Press/Elsevier, San Diego, CA, pp. 31–60.

D. P. Foster and E. I. George (1994), 'The risk inflation criterion for multiple regression', *Ann. Statist.* **22**, 1947–1975.

M. Frazier, B. Jawerth and G. Weiss (1991), *Littlewood–Paley Theory and the Study of Function Spaces*, Vol. 79 of *NSF-CBMS Regional Conf. Ser. in Mathematics*, AMS, Providence, RI.

H.-Y. Gao (1998), 'Wavelet shrinkage denoising using the non-negative garrote', *J. Comput. Graph. Statist.* **7**, 469–488.

P. J. Green and B. W. Silverman (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

P. Hall, G. Kerkyacharian and D. Picard (1999), 'On the minimax optimality of block thresholded wavelet estimators', *Statist. Sinica* **9**, 33–49.

W. James and C. Stein (1961), Estimation with quadratic loss, in *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.*, Vol. I, University of California Press, Berkeley, CA, pp. 361–379.

I. M. Johnstone (2002), Function estimation and Gaussian sequence models. Available at: `http://www-stat.stanford.edu/~imj`.

A. P. Korostelëv and A. B. Tsybakov (1993), *Minimax Theory of Image Reconstruction*, Vol. 82 of *Lecture Notes in Statistics*, Springer, New York.

L. Le Cam (2000), La statistique mathématique depuis 1950, in *Development of Mathematics 1950–2000*, Birkhäuser, Basel, pp. 735–761.

E. L. Lehmann (1997), *Theory of Point Estimation*, Springer, New York. Reprint of the 1983 original.

A. Leon-Garcia (1994), *Probability and Random Processes for Electrical Engineering*, 2nd edn, Addison-Wesley.

F. Malgouyres (2002), 'Minimizing the total variation under a general convex constraint for image restoration', *IEEE Trans. Image Process.* **11**, 1450–1456.

S. Mallat (1999), *A Wavelet Tour of Signal Processing*, 2nd edn, Academic Press, San Diego, CA.

C. L. Mallows (1973), 'Some comments on c_p', *Technometrics* **15**, 661–676.

Y. Meyer (1992), *Wavelets and Operators*, Cambridge University Press.

B. K. Natarajan (1995), 'Sparse approximate solutions to linear systems', *SIAM J. Comput.* **24**, 227–234.

M. Nussbaum (1996), 'Asymptotic equivalence of density estimation and Gaussian white noise', *Ann. Statist.* **24**, 2399–2430.

G. Schwarz (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**, 461–464.

D. W. Scott (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

B. W. Silverman (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

J.-L. Starck, F. Murtagh and A. Bijaoui (1998), *Image Processing and Data Analysis: The Multiscale Approach*, Cambridge University Press, Cambridge.

C. M. Stein (1981), 'Estimation of the mean of a multivariate normal distribution', *Ann. Statist.* **9**, 1135–1151.

S. M. Stigler (1990), *The History of Statistics: The Measurement of Uncertainty Before 1900*, The Belknap Press of Harvard University Press, Cambridge, MA. Reprint of the 1986 original.

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

H. Triebel (1992), *Theory of Function Spaces II*, Vol. 84 of *Monographs in Mathematics*, Birkhäuser, Basel.

G. Wahba (1990), *Spline Models for Observational Data*, Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, PA.

N. Wiener (1949), *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, The Technology Press of the Massachusetts Institute of Technology, Cambridge, MA.

D. Williams (1991), *Probability with Martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge.

Glossary

Bayesian estimation. In this paper, we often use the terms 'Bayesian estimator' or 'Bayes' rule' to denote any estimator which minimizes the so-called Bayes risk defined by

$$B(\pi) = E_\pi R(\theta, \hat{\theta}) = \int R(\theta, \hat{\theta})\, \pi(\mathrm{d}\theta),$$

where π is the prior distribution on the parameter θ and $R(\theta, \hat{\theta})$ is the risk of $\hat{\theta}$; see below for a definition of the risk.

Bias. The bias of an estimator is defined as the difference between the true value of the parameter vector and the expected value of the estimator under the true distribution. Suppose Y is a vector with joint distribution f_θ, where $\theta \in \Theta$ is a parameter of interest, and let $\hat{\theta}$ be a function of Y used to estimate θ. Then the bias of $\hat{\theta}$ is given by

$$\mathrm{bias}(\hat{\theta}) = \theta - E_{f_\theta}\, \hat{\theta},$$

where E_{f_θ} is the expectation of $\hat{\theta}$ under the true distribution f_θ, $E_{f_\theta}\, \hat{\theta} = \int \hat{\theta}(y)\, f_\theta(\mathrm{d}y)$. We say that an estimator is unbiased if $\mathrm{bias}(\hat{\theta}) = 0$. For example, if $Y_1, Y_2, \ldots, Y_n$ are i.i.d. $N(\theta, 1)$, then $\hat{\theta} = (Y_1 + \cdots + Y_n)/n$ is unbiased for θ.

Chi-square distribution. The chi-square distribution is that of the sum of squares of independent standard normal random variables; if we let $Z_1, Z_2, \dots, Z_d$ be i.i.d. $N(0,1)$, the random variable $Y := Z_1^2 + \cdots + Z_d^2$ follows the (central) chi-square distribution with d degrees of freedom.

Gaussian signal. A Gaussian signal is simply a Gaussian process. A Gaussian process $X = (X_1, X_2, \dots, X_n)$ is a family of random variables whose joint distribution is multivariate normal. A random vector is said to be multivariate normal if every linear combination $a_1 X_1 + \cdots + a_n X_n$ (the a_is are nonrandom) is normally distributed. In the case where the covariance matrix is nonsingular, this is equivalent to saying that the joint density of the random vector is given by

$$f(x) = \frac{1}{(2\pi)^{n/2} \, |\Sigma|^{1/2}} \, e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)/2},$$

where $\mu \in \mathbb{R}^n$ is the mean vector and $\Sigma \in \mathbb{R}^{n \times n}$ the covariance matrix.

i.i.d. 'i.i.d.' stands for independently and identically distributed. We say that the random variables $X_1, \dots, X_n$ are i.i.d. when they are all independent and follow the same distribution.

Heteroscedasticity. A sequence or a vector of random variables is heteroscedastic when the variances of the random variables in the sequence are not all the same. The complement is homoscedasticity.

Minimax estimation. A minimax estimator is any estimator whose worst-case risk is minimal. In other words, a minimax estimator is the solution to

$$\inf_{\hat\theta} \sup_{\theta \in \Theta} R(\theta, \hat\theta),$$

where Θ is the parameter space and the infimum is taken over all measurable functions of the data.

Risk of an estimator. In decision theory, we measure the quality of an estimator by the nonnegative loss function $\ell(\theta, \hat\theta)$. For example, the quadratic loss is given by $(\theta - \hat\theta)^2$ for scalar-valued parameters or $\|\theta - \hat\theta\|_{\ell_2}^2$ for vector-valued parameters. The idea is that the loss is small when θ and $\hat\theta$ are close, and increases as they get far apart. The loss is a random variable since $\hat\theta$ is random, and the risk $R(\theta, \hat\theta)$ is the expected value of the loss

$$R(\theta, \hat\theta) := E_{f_\theta} \, \ell(\theta, \hat\theta).$$

Again, E_{f_θ} is the expectation under the distribution f_θ (see the entry for 'bias').