THE ROLE OF EXPERIMENTAL STATISTICS

RUDOLF BERAN[1]
University of California, Berkeley

SUMMARY

Technological advances are driving statistics beyond mathematical philosophy and beyond computer-aided empiricism towards experimentally supported information science. Interplay between falsifiable theory and reproducible experiment is the essence of experimental statistics. A distinguishing goal of statistics is quantification of uncertainty in data-analyses through risk estimation and confidence sets. Numerical experiments with superefficient estimators for the mean vector in the one-way layout—estimators based on adaptive monotone shrinkage and on adaptive soft thresholding—illustrate how experimental statistics conditions minimax risk comparisons from theoretical statistics.

Key Words and Phrases. Computational environment, information science, soft-thresholding, monotone shrinkage, superefficient estimation.

1. INTRODUCTION

Theoretical statistics grew as a mathematical branch of philosophy. This phase of the subject's history came to the fore in the 1950's through the books of Wald (1950) on *Statistical Decision Functions*, of Savage (1954) on *The Foundations of Statistics*, and of Fisher (1956) on *Statistical*

[1] Research supported in part by National Science Foundation Grant DMS99-70266.

Methods and Scientific Inference. Contemporary papers discussed principles such as sufficiency, ancillarity, conditionality, likelihood, Bayes, and fiducial. In retrospect, these expositions may seem to be examples of grand and simple theories that explain all being. It is worth noting that the intellectual tools available to a statistician of that era were logic, mathematics, and mechanical calculators. By its nature, that technological environment directed theoretical statistics toward discussions of abstract principle and toward test statistics or confidence set pivots whose distributions could be tabulated because they did not depend on unknown parameters. Both Fisher's (1930) *Statistical Methods for Research Workers* and Quenouille's (1959) *Rapid Statistical Calculations* illustrate the strong influence of computational environment on methodology.

An analogy may be sketched between the development of theoretical statistics through the 1960's and the incubation of science by scholastic philosophers from the mid-11th to mid-15th century. The statisticians [scholastics] sought to exhibit the harmony between practice [faith] and theory [reason]. They showed great respect for earlier authorities such as Gauss, Fisher, Neyman and Wald [Aristotle and Augustine]. Some of them ultimately applied the requirements for demonstration specified in Wald's *Statistical Decision Functions* [Aristotle's *Organon*] more rigorously than their predecessors. Though they could rarely meet these standards outside the domains of mathematics and logic, the decision theorists [nominalist philosophers among the scholastics] nurtured objectivity, the start of scientific method.

Advances in time-keeping, glass-making, and printing made possible the rise of experimentally supported physics and chemistry in the centuries after the scholastics. Recent advances in computers, networking, and graphical output have created a new environment where applied and computational statistics are flourishing, traditional theoretical statistics seems out of touch, and ideas are emerging on what might be called experimental statistics. For the latter, I offer a working definition:

> Experimental statistics carries out reproducible computational experiments, numerical or symbolic, that test falsifiable predictions from theoretical statistics about the performance on data of specified statistical procedures.

It is the *interplay* of theory with experiment that distinguishes experimental statistics from pure empiricism. The historical transition from alchemy to experimental chemistry illustrates the importance of testing theory step by step. Success of a theory is gauged by its ability to predict pivotal experimental findings.

In 1269, Peter Peregrinus of Maricourt wrote *Epistola de Magnete*, a pathbreaking study of magnetism that was equally remarkable for stressing the importance of experimental skill in science. Peter stated that an investigator "diligent in the use of his own hands ... will then in a short time be able to correct an error which he would never do in eternity by his knowledge of natural philosophy and mathematics alone." At the same time, "there are many things subject to the rule of reason that we cannot completely investigate by the hand." (See Crombie (1953), p. 208). Manual experimentation, a notion shocking to some medieval scholars, lay outside the seven liberal arts of their university curricula (the *trivium* and *quadrivium*).

Section 2 of this paper sketches how the computing revolution at mid-century widened the divide between theoretical statistics and applied statistics, how part of what is called applied statistics is actually experimental statistics, and how the task of the latter is to rebuild our discipline on stronger foundations. Sections 3 and 4 illustrate the role of experimental statistics in assessing and interpreting recent theoretical results on superefficient estimators for the mean vector of the one-way

layout. Designed to reduce risk rather than for smoothing, such estimators can nevertheless behave like good scatterplot smoothers.

2. THE EMERGENCE OF EXPERIMENTAL STATISTICS

Although its existence is still not widely acknowledged, experimental statistics has considerable prehistory. We need only recall the innovative subsampling experiment published by Gosset (Student (1908a, 1908b)) to support his conjectured formulae for the t-distribution and the distribution of the sample correlation coefficient. This section considers how the schism at mid-century between theoretical and applied statistics contributed to the slowness with which experimental statistics differentiated itself from applied statistics.

After 1960, both the emerging computer environment and Prohorov's results on weak convergence of probability measures began to influence statistics. A widening division between theoretical and applied statistics was the immediate result. Some important attempts at rapprochement in the areas of time series analysis, robustness, and the bootstrap occurred through the 1970's but did not bridge the gulf. Instructive was the inability of the Princeton robustness study by Andrews et al. (1972) to resolve through experiment some outstanding conceptual gaps in robustness theory. Initially, lack of a convenient computing environment for statistics and later, lack of computing skills among those most familiar with theoretical statistics hindered sustained development of experimental statistics.

On the theoretical side after 1960, some researchers pursued the use of the new techniques in asymptotic theory to resolve several outstanding problems. A notable achievement was the understanding reached by 1970 of claims for the asymptotic optimality of maximum likelihood estimators. Though these claims are often attributed to Fisher (1925), their history includes earlier work by Bernoulli, Laplace, Gauss, and Edgeworth. For details, see references cited in Beran (1999). It became known, largely through the work of Le Cam, that maximum likelihood estimators behave very badly in certain highly regular parametric models (e.g. the three-parameter lognormal) and that this failure can be patched through suitable one-step estimators that behave in the manner Fisher suggested for MLE's.

Unexpected was the further discovery, first through the Hodges example analyzed by Le Cam (1953), that the information bound does not hold asymptotically for every estimator in regular parametric models. Le Cam's paper constructed estimators in such models that have smaller asymptotic risk on Lebesgue null sets in the parameter space than do maximum likelihood estimators. In this and subsequent work, he established that, for parameters of dimension one or two, such superefficient estimators must have high risk near their points of superefficiency. Several years later, James and Stein (1961) and Stein (1966) showed that this finding for estimation in very low dimensions is an anomalous special case. Superefficient estimators for parameters of dimension three or higher can dominate estimators of maximum likelihood type over the entire parameter space. In recent years, it has become evident that these esoteric theoretical results are early traces of an important phenomenon: the possibility of substantially reducing the risk of least squares or maximum likelihood fits in models with high-dimensional parameter spaces.

A related historical thread was the investigation, from the 1950's onwards, of estimation in

infinite-dimensional parameter spaces. In problems such as density estimation or nonparametric regression, unbiased estimators typically do not exist and unconstrained estimators of maximum likelihood type are not consistent. In essence, these difficulties are a more severe version of the inefficiency of unbiased or maximum likelihood estimators when the parameter has high, finite dimension.

On the applied side after 1960, some researchers pursued the greatly increased ability to carry out data analyses. It was now feasible to use time-series techniques or linear models with many regressors and to plot selected aspects of the data and subsequent fits. Particularly influential was Tukey's (1970, 1977) book on *Exploratory Data Analysis*. His exposition emphasized perceptive interaction between data and analyst, aided at times by custom-designed graphics. His book offset contemporary statistical textbooks where data analysis was a matter of plugging numbers into standard formulae. Although Tukey's account made little overt use of computers, some readers understood the subtext. Because of emerging technology, the medieval separation of theory and practice—the surgeon lecturing from the book while the barber made the incision—no longer sufficed for statistics.

The creation of the S language and statistical computing environment, described in the book of Becker and Chambers (1984), marked a point at which computing technology began unabashedly to shape the empirical development of statistical methodology. The S language was promulgated in software through the commercial S-PLUS and through the open source R of Ihaka and Gentleman (1996). The remarkable influence of S became evident with the success of *Modern Applied Statistics with S-PLUS* by Venables and Ripley (1994, 1997, 1999), with the increasing frequency of S graphics in journal papers, and with publication of other books, such as Davison and Hinkley (1997), that explicitly included S code.

Alternative computing environments, such as MATLAB on the engineering side, also entered the statistical research literature. Buckheit and Donoho (1995) discussed what is required to make computational experiments truly reproducible. Details of the computing environment underlying an experiment are important. Sawitzki (2000) analyzed live documents that integrate active statistical software components with text components. Examples in his paper are linked dynamically to Voyager, an experimental computing environment for statistics. Written for the Oberon operating system, Voyager is fast and portable.

The distinction between applied statistics and experimental statistics is subtle in that both make use of the same computational tools. Theoretical statistics usually models data as a sample from a probability distribution, so as to invoke the well-developed mathematics of probability theory. Experimental statistics uses case study data or generates artificial data from pseudo-random sequences. These are deterministic sequences, defined by a simple rule, for which certain long-term averages mimic probabilistic laws of large numbers (cf. Chapter 3 in Knuth (1969)). Case studies of data sets arising in statistical practice provide an important motivation for models, whether built with probability theory or pseudo-random sequence, and serve as realistic test scenarios for competing statistical methods. However, case studies not linked to insightful theory are akin to alchemy, a highly empirical but mostly unsuccessful subject that lacked connection to an effective theory of chemistry.

The titles of books such as those by Venables and Ripley or by Davison and Hinkley emphasize applied statistics or applications. This is plausible in that the computational tools described by such

authors are valuable in statistical practice and consulting. But what such books actually present are the results of case studies or simulation studies of modern statistical methods. The essence of applied statistics is the interaction between the statistical consultant and the client or client's data, while the essence of experimental statistics is the interplay between computational experiment and statistical theory. In this sense, the examples in Venables and Ripley and the "practicals" in Davison and Hinkley are laboratory exercises in experimental statistics that introduce computational tools useful in applied statistics.

Of course, experimental statistics can flourish only when its role is understood by both theoreticians and experimenters. For want of coordinated support from a generation of statisticians, the development of experimental statistics has been hesitant until recently. Papers such as Buckheit and Donoho (1995), Loader (1999), or Efromovich (1999) illustrate growing recognition of the field, though it still lacks a generally accepted name. In his book on *Understanding Media*, McLuhan (1964) noted that a technological environment is not perceived, simply because it saturates the whole field of attention. Only by studying previous environments, such as those that surrounded statistics in 1950 or science much earlier, can one understand the remarkable social and intellectual effects of environment. However anonymously, the present technological environment has given experimental statistics a revolutionary task: transforming our discipline from the dichotomy of mathematical philosophy and computer-aided empiricism into an experimentally supported information science.

3. COMPARING SUPEREFFICIENT ESTIMATORS IN THE ONE-WAY LAYOUT

Computational experiments may support conclusions drawn from theoretical study of a statistical procedure while revealing unforeseen aspects of performance. The next two sections illustrate the interplay of theory and experiment in comparing superefficient fits to the linear model that are obtained by soft-thresholding and by monotone shrinkage. Asymptotic theory for these biased alternatives to least squares was developed in Donoho and Johnstone (1995) and in Beran (2000), respectively. It is desirable to penetrate the meaning of the theoretical asymptotic minimax properties possessed by these competing methods.

Least squares fits to the Gaussian linear model began to founder intellectually after Stein (1956) proved their inadmissibility whenever the dimension of the regression space exceeds two. Initially, his result was viewed as paradoxical and even unwelcome. Once computing technology made convenient the fitting of linear models with more than a few regressors, it became more widely appreciated that least squares tends to overfit models with many regressors. This flaw is intuitively evident in special cases. Consider, for instance, a digitized signal observed in white noise (i.e. the one-way layout with one observation per factor level) or a digitized image observed in white noise (i.e. the two-way layout with one observation per factor level). In such examples, the least squares estimator of the signal is the raw data itself. It is no surprise that electrical engineers do not rely on the Gauss-Markov or Lehmann-Scheffé theorems to extract signal from observed signal-plus-noise.

We consider the Gaussian linear model in which the response vector y has a $N(X\beta, \sigma^2 I_n)$ distribution, the regression parameters β and the variance σ^2 being both unknown. Suppose that y is $n \times 1$ and the $n \times p$ regression matrix X has full rank $p \leq n$. The least squares estimator of the mean $\eta = \mathrm{E}(y) = X\beta$ is then $\hat{\eta}_{LS} = X(X'X)^{-1}X'y$. Under normalized quadratic loss, the

risk of an estimator $\hat{\eta}$ of η is $p^{-1}\mathrm{E}|\hat{\eta}-\eta|^2$. This risk is precisely σ^2 for the least squares fit $\hat{\eta}_{LS}$. Both thresholding and adaptive shrinkage techniques yield superefficient estimators for η whose asymptotic risk, as p tends to infinity, can greatly undercut that of the least squares estimator.

Our discussion will be for the one-way layout with p factor levels. A customary choice of regression matrix X for this design is the incidence matrix. Each row of this X contains a single 1, the remaining $p-1$ entries being 0. Rows are repeated according to number of replications at each factor level. The column index of the 1 indicates factor level. To define both thresholding and shrinkage estimators of η, we first transform the data to a suitable new orthogonal basis for the regression space.

For any matrix A, let $\mathcal{M}(A)$ denote the subspace spanned by the columns of A. Let U denote a suitably chosen $n \times p$ matrix with orthonormal columns such that $\mathcal{M}(U) = \mathcal{M}(X)$. Considerations that enter into the choice of U will be discussed below. Define

$$z = U'y, \qquad \xi = \mathrm{E}z. \tag{3.1}$$

Evidently, z has a $N(\xi, \sigma^2 I_p)$ distribution. The mapping between ξ, whose range is R^p, and η, whose range is the p-dimensional regression space $\mathcal{M}(X) \subset R^n$, is one-to-one with

$$\xi = U'\eta, \qquad \eta = U\xi. \tag{3.2}$$

Similarly, any estimator $\hat{\eta}$ of η corresponds in one-to-one fashion to the estimator $\hat{\xi} = U'\hat{\eta}$ of ξ, the inverse relation being $\hat{\eta} = U\hat{\xi}$. Quadratic loss is preserved under this correspondence because $U'U = I_p$ entails

$$p^{-1}|\hat{\eta}-\eta|^2 = p^{-1}|\hat{\xi}-\xi|^2. \tag{3.3}$$

3.1. Sparse and economical bases. We say that the orthonormal basis U provides a *sparse* representation of η if all but a few components of ξ are very nearly zero. The basis is *economical* if it is sparse and is ordered so that all but the first few components of ξ are very close to zero. For a sparse or economical basis, we might seek to estimate from the data the few substantially nonzero components of ξ, estimating the other components by zero. The quadratic risk of $\hat{\xi}$ would then accumulate many small squared biases from the nearly zero components of ξ but would not accumulate the many variances that arise from an attempt to estimate those components from z. This is the simple idea that underlies both thresholding and shrinkage estimators of η. Because ξ and σ^2 are unknown, we will select shrinkage factors or threshold to minimize *estimated* risk rather than risk.

The experiments reported in Section 4 use two standard smooth orthonormal bases in which successive vectors are increasingly wiggly. While wiggliness can be quantified through mathematical measures of variation, it is the empirically discovered economy or sparseness of these two bases in a range of applications that matters. Let s denote the $p \times 1$ column vector whose i-th component is i and let $u = Xs$, where X is the incidence matrix defined earlier in this section.

a) *Polynomial contrast basis.* The regression space $\mathcal{M}(X)$ of the one-way layout is spanned by the columns of the matrix $A = (u^0, u, \ldots, u^{p-1})$, where the power operations on u are performed componentwise. The columns of A are linearly independent because a polynomial of degree $p-1$ has at most $p-1$ distinct roots while $n \geq p$. The polynomial contrast basis matrix U_P

is defined as the Gram-Schmidt orthonormalization of the columns of A. Because A is nearly collinear for large p, sophisticated numerical methods are needed to compute this basis. The function `poly()` in S-PLUS offers one way. Using the polynomial contrast basis in fitting a one-way layout is not the same as fitting a polynomial curve to the data.

b) *Discrete cosine basis.* Of special interest is the trend model where $n = p$ and $X = I_p$. The discrete cosine basis U_{DC} has columns

$$\begin{aligned} c_1 &= \{p^{-1/2} : 1 \le j \le p\} \\ c_k &= \{(2/p)^{1/2} \cos[(2j-1)(k-1)\pi/(2p)] : 1 \le j \le p\} \quad \text{for } 2 \le k \le p. \end{aligned} \tag{3.4}$$

The discrete cosine transform is a modification of the discrete Fourier transform (double the range and reflect about an endpoint) that avoids creating Gibbs phenomena at the beginning and end of shrinkage estimators of η. Rao and Yip (1990) discussed properties, algorithms and applications of the discrete cosine transform to digital signal processing. A generalization of the discrete cosine basis to one-way layouts with unequal replications is the smooth basis described in Beran (2000).

Although we will not treat them in this paper, non-classical orthogonal bases may also be used for the regression space.

3.2. Estimating σ^2. Two estimators of σ^2 prove useful in estimating the risk of candidate superefficient estimators of η. To define these, let $\bar{U}$ denote an $n \times (n-p)$ matrix such that the concatenation $(U|\bar{U})$ is an orthonormal matrix. Set $\bar{z} = \bar{U}'y$.

a) *The high-component variance estimator.* The strategy of pooling sums of squares in analysis of variance suggests

$$\hat{\sigma}^2_H = (n-q)^{-1}[\sum_{i=q+1}^{p} z_i^2 + |\bar{z}|^2] = (n-q)^{-1}[\sum_{i=q+1}^{p} z_i^2 + |y - \hat{\eta}_{LS}|^2], \tag{3.5}$$

where $q \le \min\{p, n-1\}$. The bias of $\hat{\sigma}^2_H$ is

$$(n-q)^{-1} \sum_{i=q+1}^{p} \xi_i^2. \tag{3.6}$$

Consistency of $\hat{\sigma}^2_H$ is assured when this bias tends to zero as $n-q$ tends to infinity. Economy of U makes the bias small when q exceeds the number of basis vectors needed to approximate η well. When $q = p < n$, the estimator $\hat{\sigma}^2_H$ reduces to the familiar variance estimator provided by least squares theory: $\hat{\sigma}^2_{LS} = (n-p)^{-1}|y - \hat{\eta}_{LS}|^2$.

b) *The robust high-component variance estimator.* Let w denote the vector obtained by concatenating $\{z_i : q+1 \le i \le p\}$ with $\bar{z}$. Robustness theory suggests the estimator

$$\hat{\sigma}_{RH} = \text{median}\{|w_j| : 1 \le j \le n-q\}/\Phi^{-1}(.75) \tag{3.7}$$

for σ, where Φ^{-1} is the standard normal quantile function. Under normality, $\hat{\sigma}^2_{RH}$ approaches σ^2 in probability when $n-q$ is large and the high order components of ξ are small.

3.3. Monotone shrinkage and soft thresholding. Both of these are adaptive shrinkage techniques for superefficient estimation in the one-way layout. They differ in the specification of

shrinkage vectors to be considered. *Monotone shrinkage* is motivated by thinking about an economical regression basis U. The monotone class of shrinkage vectors is the closed convex set

$$\mathcal{F}_M = \{f \in [0,1]^p \colon f_1 \geq f_2 \geq \ldots \geq f_p\}. \tag{3.8}$$

In the transformed coordinate system, the candidate shrinkage estimators for ξ are $\{fz \colon f \in \mathcal{F}_M\}$, multiplication being performed componentwise as in S code. It makes sense to damp down the higher order components of z in constructing fz precisely when U is an economical basis for η.

For any vector $h \in R^p$, let $\text{ave}(h) = p^{-1}\sum_{i=1}^p h_i$. The risk of fz under normalized quadratic loss is

$$p^{-1}\text{E}|fz - \xi|^2 = \text{ave}[\sigma^2 f^2 + \xi^2(1-f)^2] \equiv \rho(f, \xi^2, \sigma^2). \tag{3.9}$$

Usually σ^2 and ξ^2 are both unknown but can be estimated by $\hat{\sigma}^2$ and $z^2 - \hat{\sigma}^2$ respectively. This yields the risk estimator

$$\hat{\rho}(f) = \text{ave}[\hat{\sigma}^2 f^2 + (z^2 - \hat{\sigma}^2)(1-f)^2]. \tag{3.10}$$

Tacit in the construction of $\hat{\rho}$ is the supposition that the law of large numbers will make $\text{ave}[(1-f)^2(z^2-\hat{\sigma}^2)]$ consistent for $\text{ave}[(1-f)^2\xi^2)]$. The uniform consistency of $\hat{\rho}(f)$ over $\mathcal{F}_M$ and over other shrinkage classes that are not too large in an entropy sense is proved by Beran and Dümbgen (1998):

$$\lim_{p\to\infty} \sup_{\text{ave}(\xi^2)\leq\sigma^2 r} \text{E} \sup_{f\in\mathcal{F}_M} |\hat{\rho}(f) - \rho(f,\xi^2,\sigma^2)| = 0 \tag{3.11}$$

for every $r > 0$.

In view of (3.11), it is reasonable on asymptotic grounds to use $\hat{\rho}(f)$ as a surrogate for the risk $\rho(f, \xi^2, \sigma^2)$ in identifying the best candidate estimator over $f \in \mathcal{F}_M$. The monotone shrinkage estimator of η is defined to be

$$\hat{\eta}_M = U\text{diag}(\hat{f}_M)U'y \quad \text{where} \quad \hat{f}_M = \operatorname*{argmin}_{f\in\mathcal{F}_M} \hat{\rho}(f). \tag{3.12}$$

The value of $\hat{f}_M$ is unique and can be computed by algorithms for weighted isotonic regression as explained in Beran and Dümbgen (1998). Apart from the details of the variance estimator $\hat{\sigma}^2$, the construction of $\hat{f}_M$ amounts to minimizing the Mallows (1973) C_L criterion or minimizing Stein's (1981) unbiased estimator of the risk $\rho(f, \xi^2, \sigma^2)$.

According to Theorem 4 in Beran (2000), the asymptotic maximum risk of $\hat{\eta}_M$ as p increases and $\hat{\sigma}^2$ tends to σ^2 is generally nonzero. Its magnitude directly reflects the economy of U as a basis for expressing η. If the basis is highly economical, this maximum risk is very small relative to the maximum risk of the least squares estimator $\hat{\eta}_{LS}$. It is never greater. For one formal description of economy, $\hat{\eta}_M$ is an asymptotically minimax estimator of η in the sense of Pinsker (1980). In practice, finding a plausibly economical basis may draw on background knowledge about the nature of η and/or on scrutiny of a plot of the signed square root of z_i versus i. Figure 1 in Section 4 exhibits such a diagnostic plot of z.

Soft thresholding is motivated by thinking about a sparse regression basis U. The class of shrinkage vectors to be considered is

$$\mathcal{G}_T = \{\hat{g}(t) \in [0,1]^p \colon \hat{g}_i(t) = [1 - t/|z_i|]_+ \text{ for } t \geq 0 \text{ and } 1 \leq i \leq p\}. \tag{3.13}$$

Unlike the monotone class $\mathcal{F}_M$ defined in (3.8), this class of shrinkage vectors is already data dependent. In the transformed coordinate system, the candidate shrinkage estimators for ξ are $\{\hat{g}(t)z\colon t \geq 0\}$, multiplication being performed componentwise. The algebraic identity

$$\hat{g}_i(t)z_i = \operatorname{sgn}(z_i)[|z_i| - t]_+ \tag{3.14}$$

reveals that these candidate estimators coincide with the soft-thresholding estimators treated by Donoho and Johnstone (1995).

We incorporate estimation of σ^2 into their paper's definitions. Let $\hat{G}$ denote the empirical cumulative distribution function of the $\{|z_i|\colon 1 \leq i \leq p\}$. Stein's (1981) unbiased estimator for the risk of $\hat{g}(t)z$ combined with a variance estimator $\hat{\sigma}^2$ yield the risk estimator

$$\hat{r}(t) = \hat{\sigma}^2[1 - 2\hat{G}(t)] + \int_0^\infty (u \wedge t)^2 d\hat{G}(u), \tag{3.15}$$

where $\wedge$ denotes the minimum operator. Let $t_0 = (2\log(p))^{1/2}\hat{\sigma}$. The Stein threshold estimator of η is then

$$\hat{\eta}_{ST} = U\operatorname{diag}(\hat{g}(\hat{t}))U'y \quad \text{where} \quad \hat{t} = \operatorname*{argmin}_{t\in[0,t_0]} \hat{r}(t). \tag{3.16}$$

Because $\hat{t}$ must be one of the values $\{|z_i|\colon 1 \leq i \leq p\}$, it can be computed readily. Another proposal by Donoho and Johnstone is the asymptotic threshold estimator

$$\hat{\eta}_{AT} = U\operatorname{diag}(\hat{g}(t_0))U'y. \tag{3.17}$$

On the grounds of improving performance for very sparse bases, Donoho and Johnstone (1995) proposed a hybrid estimator. Let $\gamma = p^{-1/2}\log_2^{3/2}(p)$ and let $s^2 = \operatorname{ave}(z^2/\hat{\sigma}^2 - 1)$. The hybrid soft thresholding estimator is

$$\hat{\eta}_{HT} = \begin{cases} \hat{\eta}_{ST} & \text{if } s^2 > \gamma \\ \hat{\eta}_{AT} & \text{if } s^2 \leq \gamma. \end{cases} \tag{3.18}$$

Theorem 1 of their paper showed that the maximum risk of this hybrid soft thresholding estimator, used with wavelet bases, has the same asymptotic order as the minimax risk computed over certain Besov bodies of η values.

4. SOME EXPERIMENTS

For fitting one-way layouts to data, the significance of the various asymptotic minimax results concerning $\hat{\eta}_M$ and $\hat{\eta}_{HT}$ is not entirely clear. This section describes several numerical experiments on these estimators, all performed in S-PLUS 3.4 (Unix). With some numerical differences, the main findings were reproduced in S-PLUS 3.2 (Windows) and R 0.90.0 (Unix). Numerical accuracy of the function `poly()` at high degrees was noticeably less in R than in S-PLUS. Though revealing, the experiments do not constitute a comprehensive empirical study of the two estimators. The intent of the exposition is to illustrate the role of experimental statistics in interpreting and assessing predictions from theoretical statistics.

4.1 The earnings data. The underlying scatterplot in the left column of Figure 1 exhibits log-income versus age of the individual sampled. This Canadian earnings data was introduced

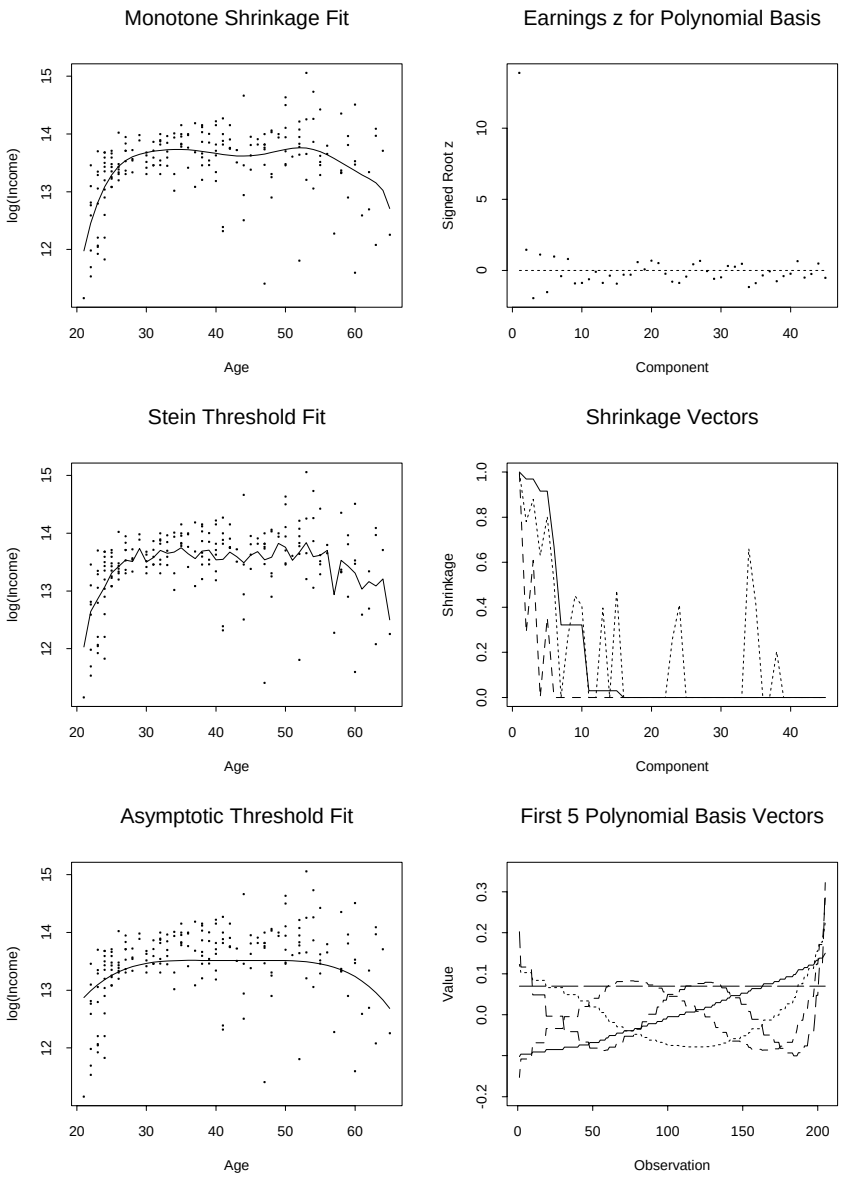

FIGURE 1. Left column: Monotone shrinkage and soft thresholding fits to the Canadian earnings data. Right column: Two diagnostic plots discussed in Section 4.1 and the first five polynomial basis vectors.

by Ullah (1985). Conditioning on the observed ages, we fit an unbalanced one-way layout to the $n = 205$ observed log-incomes, the factor levels being the $p = 45$ distinct ages from 21 to 65, taken in numerical order. The polynomial contrast basis defines the matrix U. The left column of Figure 1 exhibits the competing estimators $\hat{\eta}_M$, $\hat{\eta}_{ST}$ and $\hat{\eta}_{AT}$, using the least squares variance estimator $\hat{\sigma}^2_{LS} = .295$ to compute estimated risks of candidate estimators. In each plot, the components of the estimator have been interpolated linearly. Such interpolation is more than a visual device if we consider mean log(income) to be a continuous function of age.

Visually, the monotone shrinkage estimator fits the data reasonably. On the other hand, the Stein threshold estimator overfits while the asymptotic threshold estimator underfits. For this data,

the hybrid estimator $\hat{\eta}_{HT}$ coincides with the Stein threshold estimator. The risk of $\hat{\eta}_M$ is estimated by $\hat{\rho}(\hat{f}_M) = -.037$. The risks of $\hat{\eta}_{ST}$ and $\hat{\eta}_{AT}$ are estimated respectively by $\hat{r}(\hat{t}) = .025$ and $\hat{r}(t_0) = .199$. These various risk estimates must be taken with a grain of salt because p is only 45 and the negativity of the first risk estimator is a clear warning that the asymptotics may not have taken hold. However all three risk estimates are much smaller than the estimated risk .295 of the least squares fit to the scatterplot and the monotone shrinkage fit has smallest estimated risk. An interpolated plot of the least squares fit (not given here) would join the average log-incomes at successive ages and would greatly overfit the data.

The second column in Figure 1 presents instructive diagnostic plots. The signed square root of z_i is plotted against i in panel (1,2). The square root transformation enhances visibility of those z_i whose absolute value is close to zero. This plot supports the notion that the polynomial contrast basis is economical for mean-log income. Panel (3,2) displays the first five basis vectors in the polynomial contrast basis with linear interpolation between adjacent components. Panel (2,2) compares the shrinkage vectors $\hat{f}_M$ (solid line), $\hat{g}(\hat{t})$ (short dashes), and $\hat{g}(t_0)$ (long dashes). The occasional non-zero value of $\hat{g}_i(\hat{t})$ at high values of i explains the ragged appearance of $\hat{\eta}_{ST}$ in panel (2,1). The almost immediate plunge to zero of $\hat{g}_i(t_0)$ as i increases corresponds to the lack of detail in $\hat{\eta}_{AT}$ in panel (3,1). Neither soft thresholding fit to this earnings data is satisfactory when compared with the monotone shrinkage fit.

4.2 Artificial data. To see how varying basis economy or sparsity affects the performance of monotone shrinkage or soft thresholding estimators, we consider four experiments based on artificial data. In each case, the design is a one-way layout with one observation per factor level and $n = p = 200$. The respective mean vectors are generated from four functions defined on $[0, 1]$:

Smooth: $m_1(t) = 2(6.75x^2(1-x))^3$.

Crash: $m_2(t) = 0$ if $0 \le t \le .25$ and $= \sin(2\pi/t)$ if $.25 < t \le 1$.

Steps: $m_3(t) = 0$ if $0 \le t \le .15$, $= 1.5$ if $.15 < t \le .3$, $= .5$ if $.3 < t \le .475$, $= -.5$ if $.475 < t \le .6$, $= 2$ if $.6 < t \le .8$, and $= 1$ if $.8 < t \le 1$.

Jolt: $m_4(t) = t + \sin(60\pi t)$ if $.25 \le t \le .50$ and $= t$ otherwise.

The jth artificial data set is a pseudo-random sample drawn from the $N(\eta, \sigma^2 I_{200})$ distribution with $\eta = \{m_j(i/201) : 1 \le i \le 200\}$ and $\sigma = .2$. The signal-to-noise ratio $\text{ave}(\eta^2)/\sigma^2$ varies as follows:

	Smooth	Crash	Steps	Jolt
S/N Ratio	27.0	9.3	35.3	8.7

The four scatterplots so obtained, displayed in the successive columns in Figures 2 and 3, each employ the same set of errors generated by `rnorm()` using `set.seed(2)`.

The dashed lines in each column indicate the true mean vector η. We use the discrete cosine basis (3.4) to estimate η by monotone shrinkage and by soft thresholding. This basis is very economical for Smooth, less economical for Crash, and even less economical for Steps. It is fairly sparse for Jolt but not too economical. The variance σ^2 is estimated by the high-component estimator $\hat{\sigma}^2_H$ defined in (3.5), with $q = .75p$. This ad hoc choice of q, suggested by scrutiny of a plot of z_i versus i, provides 50 degrees-of-freedom for the variance estimator and controls its bias because the means of the $\{z_i : 151 \le i \le 200\}$ appear to be relatively close to zero. The solid lines in each column of

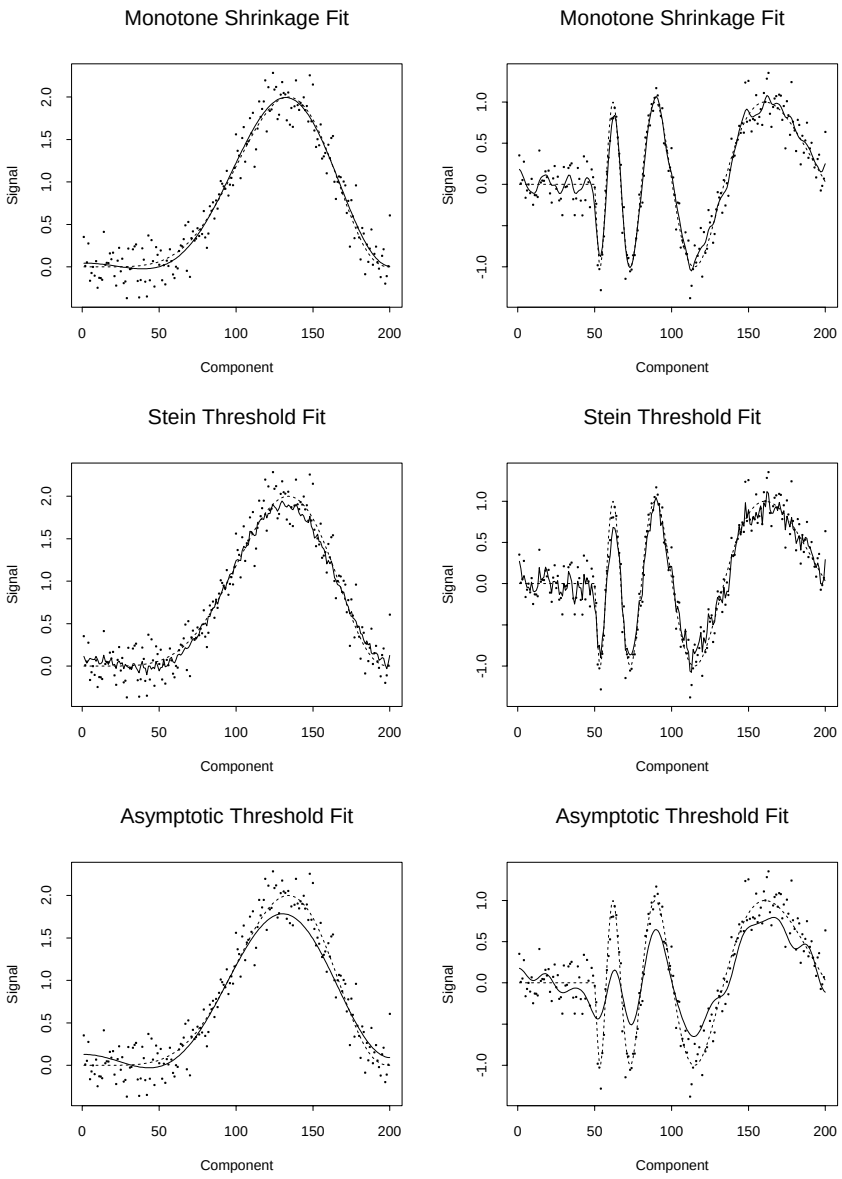

FIGURE 2. The two columns exhibit monotone shrinkage and soft thresholding fits to the Smooth and Crash artificial data, respectively. The dashed line indicates the true η.

Figures 2 and 3 represent, with linear interpolation, the mean vector estimators $\hat{\eta}_M$, $\hat{\eta}_{ST}$ and $\hat{\eta}_{AT}$ for that scatterplot. The fits are not sensitive to modest variation in the choice of q.

In each case, the hybrid estimate $\hat{\eta}_{HT}$ coincides with $\hat{\eta}_{ST}$. Visually, the asymptotic threshold estimator underfits in each instance. The Stein threshold estimate overfits relative to the corresponding monotone shrinkage estimate except in the Jolt example. While the Stein threshold estimate is considerably better than the asymptotic threshold estimate, it does not track η as well as the monotone shrinkage estimate. This is particularly visible at the jumps in the Steps example. In the Jolt example, the discrete cosine basis is reasonably sparse but is not too economical because of the high frequency wiggle. Even there, monotone shrinkage tracks better than Stein thresholding near the upper and lower endpoints.

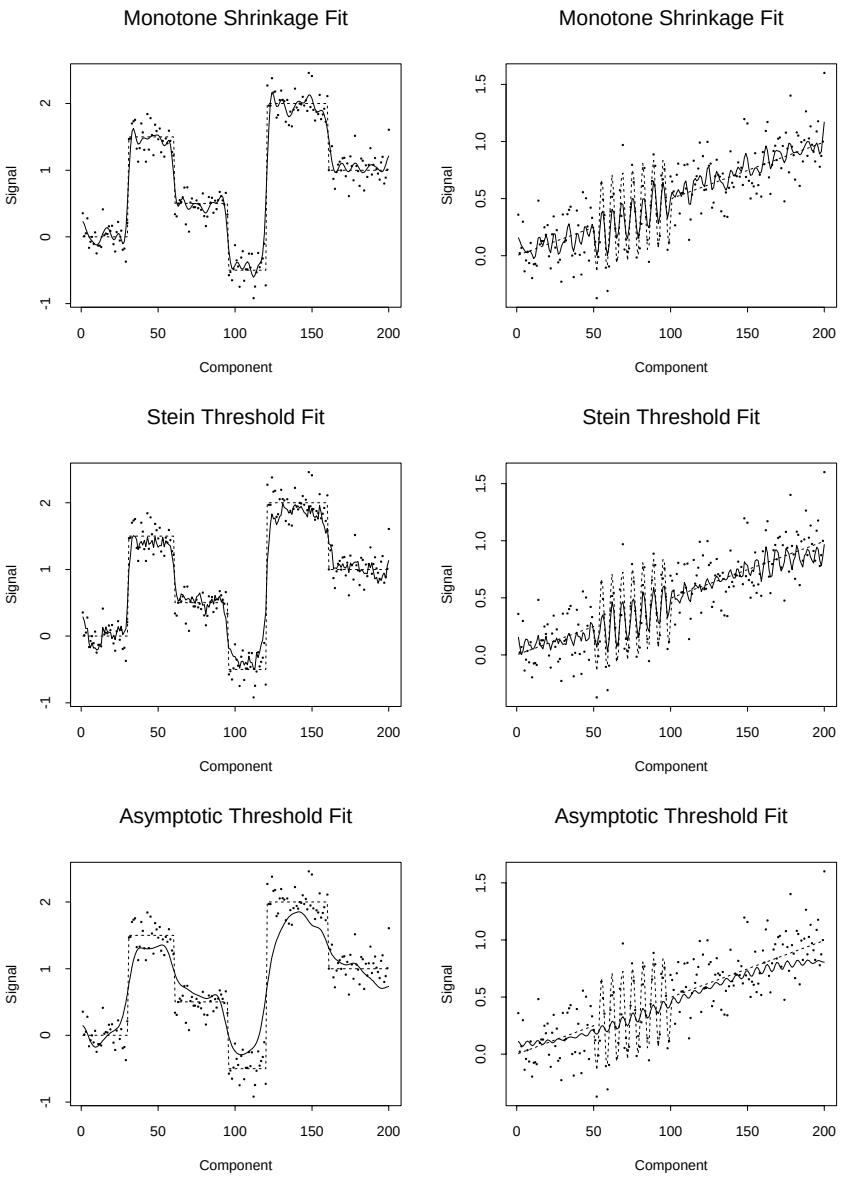

FIGURE 3. The two columns exhibit monotone shrinkage and soft thresholding fits to the Steps and Jolt artificial data, respectively. The dashed line indicates the true η.

The following table records plug-in risk *estimates*, based on $\hat{\rho}$ for $\hat{\eta}_M$, on $\hat{r}$ for $\hat{\eta}_{ST}$ and $\hat{\eta}_{AT}$, and on either function for $\hat{\eta}_{LS}$. Comparing the risk of the least squares estimator, $\sigma^2 = .04$, with the estimated risks in the last column indicates how accurately $\hat{\sigma}^2_H$ estimates σ^2 in each case and reminds us that estimated risk only approximates true risk.

	$\hat{\eta}_M$	$\hat{\eta}_{ST}$	$\hat{\eta}_{AT}$	$\hat{\eta}_{LS}$
Smooth	-.0065	-.0025	.0054	.0448
Crash	.0025	.0158	.0633	.0449
Steps	.0110	.0286	.0846	.0677
Jolt	.0026	.0022	.0123	.0449

The negative risk estimates are of course not distinguishable from zero, though their order might

matter. In general, the visual quality of the fits in Figures 2 and 3 follows the ordering implied by estimated risk. It is interesting to note that the estimated risk of $\hat{\eta}_{AT}$ exceeds that of the least squares fit in both the Crash and Steps examples.

These few experiments reveal wilful blindness in the Gauss-Markov theorem's restriction to unbiased estimators and blinkered vision in recent asymptotic minimax studies of superefficient biased estimators. The near asymptotic minimaxity over Besov bodies of softly thresholded wavelet fits does not ensure strong performance of soft thresholding in general. For economical bases, adaptive monotone shrinkage fits are asymptotically minimax and can dominate noticeably the results of soft thresholding. For sparse but not economical bases, one might investigate further the performance of bounded variation shrinkage (cf. Beran and Dümbgen (1998)).

The foregoing discussion illustrates the role of experimental statistics in interpreting and assessing abstract optimality properties from theoretical statistics. Experiments such as those described are now easily carried out on a personal computer. We can expect that insightful experiments will more frequently complement mathematical arguments for statistical procedures, both in teaching and in research. Active "textbooks" responsive to the user's direction may soon facilitate key experiments through dynamic linkage to a statistical computing environment.

The Introduction to this paper quoted the thirteenth century scientist Peter of Maricourt. His contemporary, Roger Bacon, wrote of Peter, "What others strive to see dimly and blindly, like bats in twilight, he gazes at in the full light of day, because he is a master of experiment" (see Crombie (1953), p. 205). Bacon's words aptly describe the role of experimental statistics, a discipline that technology increasingly supports.

REFERENCES

Andrews, D. F., Bickel, P. J., Hampel, F. R., Rogers, W. H., and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton.

Becker, R. A. and Chambers, J. M. (1984). *S: An Interactive Environment for Data Analysis and Graphics.* Wadsworth, Belmont CA.

Beran, R. (1999). Fisher's program. In *Encyclopedia of Statistical Sciences: Update Volume 3* (S. Kotz, C. B. Read, and D. L. Banks, eds.) 242–246. Wiley, New York.

Beran, R. (2000). REACT scatterplot smoothers: superefficiency through basis economy. *J. Amer. Statist. Assoc.* **95** 155–171.

Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Ann. Statist.* **26** 1826–1856.

Buckheit, J. B. and Donoho, D. L. (1995). WaveLab and reproducible research. Technical Report 474, Dept. of statistics, Stanford University.

Crombie, A. C. (1953). *Robert Grosseteste and the Origins of Experimental Science 1100–1700.* Clarendon Press, Oxford.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press.

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

Efromovich, S. (1999). Quasi-linear wavelet estimation. *J. Amer. Statist. Assoc.* **94** 189–204.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phil. Soc.* **22** 700–725.

Fisher, R. A. (1930). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Hafner, New York.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *J. Computat. Graphical Statist.* **5** 299–314.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* (J. Neyman, ed.) **I** 361–380. University of California Press.

Knuth, D. E. (1969). *The Art of Computer Programming, Vol 2: Seminumerical Algorithms.* Addison Wesley, Reading MA.

Le Cam, L. (1953). On some asymptotic properties of maximum likelihood estimators and related Bayes estimates. *Univ. Calif. Publ. Statist.* **1** 277–330.

Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Statist.* **27** 415–438.

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

McLuhan, H. M. (1964). *Understanding Media: The Extensions of Man.* McGraw Hill, New York.

Pinsker, M. S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems Inform. Transmission* **16** 1200–133.

Quenouille, M. H. (1959). *Rapid Statistical Calculations; a Collection of Distribution-Free and Easy Methods of Estimation and Testing.* Griffin, London.

Rao, K. R. and Yip, P. (1990) *Discrete Cosine Transform. Algorithms, Advantages, Applications.* Academic Press, Boston.

Savage, L. J. (1954). *The Foundations of Statistics.* Wiley, New York.

Sawitzki, G. (2000). Keeping statistics alive in documents. *Computational Statistics* **9** in press.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Third Berkeley Symp. Math. Statist. Probab.* (J. Neyman, ed.) **I** 197–208. University of California Press.

Stein, C. (1966). An approach to the recovery of inter-block information in balanced incomplete block designs. In *Festschrift for Jerzy Neyman* (F. N. David, ed.) 351–366. Wiley, London.

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

Student (1908a). On the probable error of a mean. *Biometrika* **6** 1–25.

Student (1908b). Probable error of a correlation coefficient. *Biometrika* **6** 302–310.

Tukey, J. W. (1970). *Exploratory Data Analysis.* Limited preliminary edition. Regular edition published in 1977. Addison-Wesley, Reading MA.

Ullah, A. (1985). Specification analysis of econometric models. *J. Quantitative Econ.* **2** 187–209.

Venables, W. N. and Ripley, B. D. (1994). *Modern Applied Statistics with S-PLUS.* Springer, New York. Second edition (1994). Third edition (1999).

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.