An Introduction to James-Stein Estimation

John A. Richards

November 5, 1999

Abstract

In 1961, James and Stein introduced an estimator of the mean of a multivariate normal distribution that achieves a smaller mean-squared error than the maximum likelihood estimator in dimensions three and higher. This estimator shrinks the observation vector toward a previously specified point target by an adaptive shrinkage factor. We discuss the properties of the James-Stein estimator and describe how it can be adapted to various settings. We examine its application to the problems of dynamic state estimation and function denoising and discuss an extension of the estimator to the scenario in which specification of a single point shrinkage target is impractical.

Contents

1 Introduction

Consider the standard parameter estimation problem: given a measurement vector $\mathbf{x}$ that depends probabilistically on a parameter vector $\boldsymbol{\theta}$ according to some known probability density function (pdf) $p(\mathbf{x};\boldsymbol{\theta})$, choose from some class $\mathcal{A}$ of functions of $\mathbf{x}$ the estimator $\hat{\boldsymbol{\theta}}$ that minimizes or maximizes a specified criterion function. In maximum likelihood (ML) estimation, for instance, the criterion function to be maximized is the likelihood $p(\mathbf{x};\boldsymbol{\theta})$ and $\mathcal{A}$ is the space of measurable functions of $\mathbf{x}$.

Now consider the problem of estimating the mean of a multivariate normal distribution from a vector of uncorrelated, equal-variance measurements. Specifically, let $\mathbf{x} \in \mathbb{R}^p$ be a measurement drawn from a normal distribution with mean $\boldsymbol{\theta}$ and covariance $\sigma^2 I$, so that

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x}-\boldsymbol{\theta})'(\mathbf{x}-\boldsymbol{\theta})\right], \tag{1}$$

which we abridge to write $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$. Application of the ML criterion to (1) yields the maximum likelihood estimate (MLE)

$$\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{x},$$

which satisfies $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} \sim N(\boldsymbol{\theta}, \sigma^2 I)$.

A common measure of estimator performance is mean-squared error (MSE), defined as

$$J(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}) = E\left(\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\right), \tag{2}$$

where the notation is often shortened simply to $J(\boldsymbol{\theta})$ when the implied estimator is clear from context, or to a similar compact form, *e.g.*, $J_{\mathrm{ML}}(\boldsymbol{\theta})$ for the MSE of an MLE. It is easy to show that the MLE for the pdf in (1) has MSE

$$J_{\mathrm{ML}}(\boldsymbol{\theta}) = p\sigma^2.$$

The MLE was long thought to be the "best" estimator for the multivariate mean estimation problem; although a proof was lacking, it was believed that no estimator existed that achieved a lower MSE for all values of $\boldsymbol{\theta}$.[1]

In 1961, James and Stein [2] introduced an estimator of multivariate normal mean for dimension three and higher that achieves uniformly lower MSE than the MLE for all parameter values $\boldsymbol{\theta}$. That is, they constructed an estimator $\hat{\boldsymbol{\theta}}^{\mathrm{JS}}$ for which

$$J_{\mathrm{JS}}(\boldsymbol{\theta}) < J_{\mathrm{ML}}(\boldsymbol{\theta}), \quad \forall\boldsymbol{\theta} \tag{3}$$

when $p > 2$. (Note that the inequality in (3) is strict.) This was a culmination of work Stein had begun several years earlier [3], which had hinted at an estimator that might achieve better performance than the MLE. The demonstration of the James-Stein estimator (JSE) in 1961 shocked the statistics community [4]. Dennis Lindley, a prominent statistician of the period, summarized his reaction to the JSE in a published review [5] of a later paper by Stein [6]:

> The idea . . . is that the mean of a multivariate normal distribution is not best estimated by the sample mean. When I first heard of this suggestion several years ago I must admit that I dismissed it as the work of one of these mathematical statisticians who are so

[1] Wald [1] attempted a proof along these lines, which was later shown to be flawed [2].

entranced by the symbols that they lose touch with reality. It must, I argued, be due to the unbounded loss function, or it could be an ϵ-improvement, or the sample size was small. But it is none of these things. The estimate proposed by the author is realistic, a great improvement on the sample mean, and makes good practical sense. . . . We have here one of the most important original statistical ideas of the decade, destined, I feel sure, to influence our thinking and our practice.

Lindley's reaction, from initial skepticism to eventual acceptance and enthusiasm, was typical among statisticians [4].

The JSE takes the form

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{\sigma^2(p-2)}{\mathbf{x}'\mathbf{x}}\right)\mathbf{x}$$

for use in the scenario of (1). The JSE does not share many of the "nice" properties of the MLE: it is nonlinear, it is biased, and its pdf cannot be expressed in a compact form. However, the resounding benefit of (3) greatly diminishes the force of these complaints. This inequality states that the JSE will *always* outperform the MLE, regardless of the true value of $\boldsymbol{\theta}$, as long as the dimension of the measurement is at least three. The reduction in MSE achieved by the JSE is attributable solely to the dimensionality of the problem and not to any trick or hidden property of the data. It is a fundamental result.

Resistance to the JSE was significant in the years following its introduction; philosophical arguments against its use and validity were raised [7]. However, the clear fact of its superiority in terms of MSE gradually led to its wider acceptance and use. Today the JSE is well known among statisticians [8, 9] and econometricians [10, 11]. However, the JSE has received little attention within the signal processing and broader engineering communities. Our intention is to describe the JSE and its properties and to examine some applications in which it has found use that are of interest to the signal processing community.

Before proceeding, it will be helpful to establish several definitions. The *loss* of an estimator $\hat{\boldsymbol{\theta}}$ depends on the true parameter value $\boldsymbol{\theta}$ and is defined as $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$. The *risk* of an estimator is simply its expected loss as a function of $\boldsymbol{\theta}$ and is identical to the MSE as we have already defined it in (2).[2] An estimator $\hat{\boldsymbol{\theta}}$ is said to be *minimax* if its largest risk is no greater than that of any other estimator. More precisely, an estimator $\hat{\boldsymbol{\theta}}$ is said to be minimax if $\sup_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) \leq \sup_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ for all other estimators $\tilde{\boldsymbol{\theta}}$. Estimator $\hat{\boldsymbol{\theta}}$ is said to *dominate* estimator $\tilde{\boldsymbol{\theta}}$ if $J(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}) \leq J(\boldsymbol{\theta}; \tilde{\boldsymbol{\theta}})$ for all $\boldsymbol{\theta}$ and if there exists some value of $\boldsymbol{\theta}$ for which the inequality is strict. Finally, an estimator $\hat{\boldsymbol{\theta}}$ is said to be *admissible* if there does not exist another estimator that dominates it. Summarizing these last terms, a minimax estimator is simply one with the best worst-case scenario; a dominating estimator has lower or equal risk for every possible parameter value; an admissible estimator is one that can only be bested for some value of $\boldsymbol{\theta}$ at the expense of a higher risk for another value of $\boldsymbol{\theta}$.

In Section 2 we describe the JSE and its properties in more detail. In the following sections we examine three applications of the estimator. Specifically, in Section 3 we examine the James-Stein state filter proposed by Manton, Krishnamurthy, and Poor [12]; in Section 4 we investigate wavelet shrinkage based partially on James-Stein methods, as proposed by Donoho and Johnstone [13]; in Section 5 we examine a multiple shrinkage JSE proposed by George [14]. Section 6 concludes the paper.

[2] In general, risk can be based on an arbitrary loss function; however, we will consider only squared error loss and thus risk will be the same as MSE.

2 James-Stein Estimation

In Section 1 we introduced the James-Stein estimator (JSE) for the mean of a multivariate normal with covariance $\sigma^2 I$. In this section we will present more general and flexible forms of the JSE and discuss its properties in more detail. The treatment of this section will prove useful for the analysis of the estimator in the contexts of the applications of Sections 3–5.

2.1 James-Stein Estimation for $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$

The canonical scenario examined in Section 1 deals with the estimation of the mean of a multivariate normal with covariance equal to a multiple of the identity matrix. For convenience, we restate the essential result: for $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$, where $\mathbf{x}, \boldsymbol{\theta} \in \mathrm{I\!R}^p$, the JSE is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{\sigma^2(p-2)}{\mathbf{x}'\mathbf{x}}\right)\mathbf{x}. \tag{4}$$

Because $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{x}$, we may also write

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{\sigma^2(p-2)}{\hat{\boldsymbol{\theta}}^{\mathrm{ML}\prime}\hat{\boldsymbol{\theta}}^{\mathrm{ML}}}\right)\hat{\boldsymbol{\theta}}^{\mathrm{ML}},$$

a formulation that will be convenient especially in Section 3. We have already stated the fundamental property of the JSE in (3), *i.e.*, the JSE dominates the MLE when $p > 2$.[3] Because the MLE is minimax [8], the dominance of the JSE implies that the JSE is also minimax.

One way to demonstrate dominance of the JSE is simply to calculate $J_{\mathrm{JS}}(\boldsymbol{\theta})$ for the estimator of (4) and establish that $J_{\mathrm{JS}}(\boldsymbol{\theta}) < p\sigma^2 = J_{\mathrm{ML}}(\boldsymbol{\theta})$. It is also possible to derive the JSE constructively by restricting attention to estimators of the form $(1 - \frac{\alpha}{\mathbf{x}'\mathbf{x}})\mathbf{x}$, choosing α to minimize risk, and demonstrating that this risk is less than $p\sigma^2$ for any value of $\boldsymbol{\theta}$. (This is the tactic used by James and Stein in [2].) The latter approach has the additional benefit of showing that any estimator of the form

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{c\sigma^2}{\mathbf{x}'\mathbf{x}}\right)\mathbf{x} \tag{5}$$

with $0 < c < 2(p-2)$ will in fact dominate the MLE when $p > 2$ [10, 11].[4] We will refer to any instance from the entire class of these estimators as a JSE, assuming the choice of $c = p-2$ as in (4) unless otherwise noted.

It is possible to derive an exact expression for the risk of the JSE [10]. Specifically,

$$J_{\mathrm{JS}}(\boldsymbol{\theta}) = p\sigma^2 - (p-2)\sigma^2 \cdot \exp\left[-\frac{\boldsymbol{\theta}'\boldsymbol{\theta}}{2\sigma^2}\right] \cdot \sum_{n=0}^{\infty} \frac{(\boldsymbol{\theta}'\boldsymbol{\theta})^n}{(2\sigma^2)^n(p-2+2n)n!}. \tag{6}$$

Although the summation on the right-hand side of (6) makes exact calculation of $J_{\mathrm{JS}}(\boldsymbol{\theta})$ unwieldy in practice, we can see qualitatively that $J_{\mathrm{JS}}(\boldsymbol{\theta})$ decreases as $\|\boldsymbol{\theta}\|^2$ approaches zero, and reaches a minimum of $2\sigma^2$ when $\boldsymbol{\theta} = \mathbf{0}$, regardless of the size of p. Comparing to the risk of $\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$, which is $p\sigma^2$ for all $\boldsymbol{\theta}$, we see that if $\boldsymbol{\theta}$ is in the vicinity of $\mathbf{0}$ then the JSE will provide a substantial improvement over the MLE. In fact, the potential savings from use of the JSE grow linearly with p. As $\|\boldsymbol{\theta}\|^2$

[3] It can be shown that for $p \leq 2$, the MLE is an admissible estimator.

[4] Note that for $c = 0$, (5) reverts to the MLE $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{x}$.

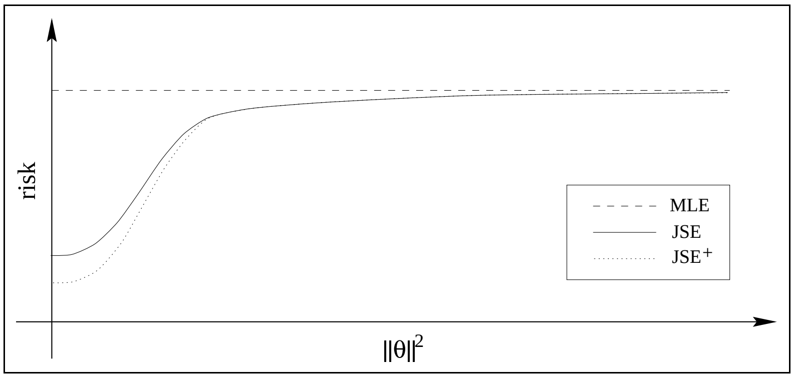

Figure 1: Caricature of Relative Risk of MLE, JSE, and JSE$^+$

increases to infinity, $J_{\rm JS}(\boldsymbol{\theta})$ approaches $J_{\rm ML}(\boldsymbol{\theta}) = p\sigma^2$ and thus the savings diminish to zero. By way of approximation, a useful upper bound on $J_{\rm JS}(\boldsymbol{\theta})$ (derived in [10]) for any value of $\boldsymbol{\theta}$ is

$$J_{\rm JS}(\boldsymbol{\theta}) \leq p\sigma^2 - \frac{\sigma^4(p-2)^2}{\sigma^2(p-2) + \boldsymbol{\theta}'\boldsymbol{\theta}}.$$

Figure 1 depicts the risk of the JSE as a function of $\|\boldsymbol{\theta}\|^2$ in comparison to the risk of the MLE. (The trace labeled "JSE$^+$" corresponds to a modification of the JSE that will be discussed in Section 2.3.3.)

2.2 Heuristic Justifications for the James-Stein Estimator

It is not intuitively clear why the JSE should dominate the MLE. The MLE has the intuitive appeal of "good sense" in many estimation problems: it often takes the form of an average of repeated measurements or a least-squares approximation [15]. Indeed, estimation and statistical decision theory had advanced for over 150 years since the time of Gauss without an indication that what seemed the eminently reasonable and correct way to estimate a mean could be bested. In this section we present two arguments suggesting the superiority of the JSE that stop short of being a mathematical derivation of the JSE.

2.2.1 Stein's Original Argument

Stein's original argument [3] for an estimator of the form in (4) is based on a comparison of $\boldsymbol{\theta}'\boldsymbol{\theta}$ to $\mathbf{x}'\mathbf{x}$ when p is large and proceeds as follows. (A summary of this argument can be found in [9].) Intuitively, a good estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta} \in \mathbb{R}^p$ should satisfy $\hat{\theta}_i \approx \theta_i$ for $i = 1, \ldots, p$. This implies that $\hat{\boldsymbol{\theta}}$ should also satisfy $\hat{\theta}_i^2 \approx \theta_i^2$ for $i = 1, \ldots, p$ and thus

$$\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta}'\boldsymbol{\theta}. \tag{7}$$

We would hope that a chosen estimator satisfies this condition.

Suppose now that we restate the canonical model of (1) as $x_i \sim N(\theta_i, \sigma^2)$ for $i = 1, \ldots, p$ and observe that $\hat{\theta}_i^{\rm ML} = x_i$. Let $y = \frac{1}{p}\sum_{i=1}^p x_i^2 = \frac{1}{p}\mathbf{x}'\mathbf{x}$ so that $E(y) = \sigma^2 + \frac{1}{p}\boldsymbol{\theta}'\boldsymbol{\theta}$. Application of Chebyshev's inequality to y reveals that

$$\frac{1}{p}\mathbf{x}'\mathbf{x} \longrightarrow \sigma^2 + \frac{1}{p}\boldsymbol{\theta}'\boldsymbol{\theta}$$

in probability as $p \to \infty$ [3, 9]. In other words, for large p, it is very likely that $\mathbf{x}'\mathbf{x}$ is larger than $\boldsymbol{\theta}'\boldsymbol{\theta}$. This suggests that to form a good estimate of $\boldsymbol{\theta}$, in the sense of (7), we would need to shrink $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{x}$ toward $\mathbf{0}$, which is exactly what the JSE does. A more intricate argument suggests how the shrinkage factor might be chosen, yielding something very similar to (4) [9].

2.2.2 An Empirical Bayesian Argument

Another argument for the JSE is based on the Bayesian formulation. (This argument is presented in [16], among other places.) The estimation framework up to this point has relied in part on the assumption that $\boldsymbol{\theta}$ is a nonrandom parameter. Suppose for the time being (*i.e.*, within the current section only) that $\boldsymbol{\theta}$ is modeled as a random quantity. Specifically, suppose that we model

$$\boldsymbol{\theta} \sim N(\mathbf{0}, \tau^2 I), \tag{8}$$

where τ^2 is assumed to be known. With the introduction of a prior density, $\boldsymbol{\theta}$ could be estimated in a Bayesian setting. The Bayes least-squares estimate (BLSE) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}}^{\mathrm{BLS}} = \frac{\tau^2}{\sigma^2 + \tau^2}\mathbf{x} = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right)\mathbf{x}. \tag{9}$$

(The right-hand side of the second equality is intended to evoke the form of the JSE.) The BLSE is optimal in the sense that it achieves the smallest expected MSE of any estimator of $\boldsymbol{\theta}$ by construction.

Now, suppose that τ^2 is unknown. We might imagine trying to mimic the BLSE by first estimating τ^2 and then using this estimate to produce an "empirical" BLSE of $\boldsymbol{\theta}$. Such an approach, in which a prior density does not exist or is unknown, but is estimated from the data in order to apply the Bayesian framework, is known as *empirical Bayesian* estimation [8, 9]. The form of (9) suggests that we could forgo estimation of τ^2 and instead estimate $\frac{\sigma^2}{\sigma^2+\tau^2}$ directly. It can be shown [10] that $\frac{\sigma^2(p-2)}{\mathbf{x}'\mathbf{x}}$ is an unbiased estimator of $\frac{\sigma^2}{\sigma^2+\tau^2}$. If this estimator is substituted into (9), the JSE is obtained. Thus one interpretation of the JSE is that it is an attempt to perform Bayes estimation when no prior is present, but one of the form of (8) is assumed.

It can also be shown [8, 10] that assuming the prior of (8), the Bayes risk of the JSE (*i.e.*, the expectation of $J_{\mathrm{JS}}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$) is

$$E(J_{\mathrm{JS}}(\boldsymbol{\theta})) = p\sigma^2 - \frac{(p-2)\sigma^4}{\sigma^2 + \tau^2} = \frac{p\sigma^2\tau^2 + 2\sigma^4}{\sigma^2 + \tau^2},$$

whereas the Bayes risk of the BLSE estimator (with τ^2 known) is

$$E(J_{\mathrm{BLS}}(\boldsymbol{\theta})) = p\sigma^2 - \frac{p\sigma^4}{\sigma^2 + \tau^2} = \frac{p\sigma^2\tau^2}{\sigma^2 + \tau^2},$$

which is always smaller. The JSE can be viewed as an emulator of the BLSE, based on the assumption of a prior of the form in (8) in which τ^2 is unknown and must be estimated from the data. The uncertainty inherent in this estimation necessarily results in a larger Bayes risk.

Before moving on, we note that the assumed prior (8) on $\boldsymbol{\theta}$ could easily have been chosen to have a nonzero mean. The choice of a zero mean resulted in a BLSE that shrinks toward $\mathbf{0}$; choosing a nonzero mean $\bar{\boldsymbol{\theta}}$ would have resulted in shrinkage toward $\bar{\boldsymbol{\theta}}$. In the next section we will examine a modification to the JSE that allows shrinkage toward an arbitrary point. This modified JSE can be defended in the empirical Bayesian context with (8) modified to have a nonzero mean.

2.3 Modifications to the James-Stein Estimator

The JSE defined by (4) produces an estimate by shrinking the observed data $\mathbf{x}$ (or equivalently, the MLE of $\boldsymbol{\theta}$) toward $\mathbf{0}$ by an adaptive shrinkage factor. In this section we examine three modifications to the JSE that can improve its MSE performance. These modified estimators are the form of the JSE most often encountered in practice, *e.g.*, in [12, 13, 14, 17]. The first modification allows application of the JSE to cases where the elements of $\mathbf{x}$ are correlated. The second generalizes the shrinkage, so that the origin $\mathbf{0}$ does not have to be the implicit target. The third modification improves upon the shrinkage coefficient to lower the risk of the already-specified JSE. These modifications are discussed in turn in the next three sections.

2.3.1 James-Stein Estimation for $\mathbf{x} \sim N(\boldsymbol{\theta}, Q)$

The canonical scenario outlined so far is restrictive in that it requires the elements of $\mathbf{x}$ to be uncorrelated and have equal variance σ^2. A more general formulation would be the situation where $\mathbf{x} \sim N(\boldsymbol{\theta}, Q)$, where Q is an arbitrary symmetric $p \times p$ positive definite covariance matrix. In this case the MLE of $\boldsymbol{\theta}$ is still simply $\hat{\boldsymbol{\theta}}^{\mathrm{ML}} = \mathbf{x}$, with risk $J_{\mathrm{ML}}(\boldsymbol{\theta}) = \mathrm{tr}(Q)$. In 1975, Bock derived a JSE (*i.e.*, an estimator with form similar to that of (4) that dominates the MLE under certain conditions described below) for this more general case [18]. Specifically, the JSE in this case can be expressed as

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{\tilde{p} - 2}{\mathbf{x}' Q^{-1} \mathbf{x}}\right) \mathbf{x}, \tag{10}$$

where $\tilde{p}$ is known as the *effective dimension* of Q and is defined as

$$\tilde{p} = \frac{\mathrm{tr}(Q)}{\lambda_{\max}(Q)}, \tag{11}$$

the trace of matrix Q divided by its maximum eigenvalue. Bock showed that when $\tilde{p} > 2$, the estimator of (10) dominates the MLE. More generally, any estimator of the form

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \left(1 - \frac{c}{\mathbf{x}' Q^{-1} \mathbf{x}}\right) \mathbf{x}$$

with $0 < c < 2(\tilde{p} - 2)$ dominates the MLE when $\tilde{p} > 2$.

At first glance, the concept of "effective dimension" might appear to be somewhat unusual. It has a simple intuitive explanation, however. First note that when $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$, we have $\tilde{p} = p$, indicating that the effective dimension corresponds to the usual concept of dimension when $\mathbf{x}$ has covariance $\sigma^2 I$. Now consider the more general case where Q has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, corresponding to eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p$. Suppose that λ_1 is much greater than the sum of the remaining eigenvalues, so that $\tilde{p} \approx 1$. This means that given $\mathbf{x}$, the uncertainty in the location of the mean is primarily along the direction of $\mathbf{v}_1$. Essentially this suggests that we could make the approximation $\mathbf{x} \approx \boldsymbol{\theta} + x\mathbf{v}_1$, effectively reducing a p-dimensional estimation problem to a one-dimensional problem. Because the MLE is admissible for $p \leq 2$, intuitively it would seem that it should also be admissible in a situation like this. Effective dimension is thus an indication of how many significant directions of uncertainty really exist.

2.3.2 Shrinkage Toward an Arbitrary Target

The JSE has a lower risk than the MLE over the entire parameter space, but most markedly when $\boldsymbol{\theta}$ is near $\mathbf{0}$. (This was discussed in Section 2.1, expressed in (6), and depicted in Figure 1.) The reduction in risk when $\|\boldsymbol{\theta}\|^2 \approx 0$ is impressive, but raises the issue of what should be done if there were reason to believe that $\boldsymbol{\theta}$ might lie in some region of parameter space distant from the origin. The risk analysis suggests that in this case the JSE would offer only marginal improvement over the MLE.

Suppose $\bar{\boldsymbol{\theta}} \in \mathbb{R}^p$ is a prior guess or indication of the true value of $\boldsymbol{\theta}$. The JSE can be easily modified to shrink toward $\bar{\boldsymbol{\theta}}$ instead of toward $\mathbf{0}$. All this requires is the utilization of the standard JSE of (10) to estimate $\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}$ from $\mathbf{x} - \bar{\boldsymbol{\theta}}$; this estimate can then be added to $\bar{\boldsymbol{\theta}}$ to yield an estimate of $\boldsymbol{\theta}$. Specifically,

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = \bar{\boldsymbol{\theta}} + \left(1 - \frac{\tilde{p} - 2}{(\mathbf{x} - \bar{\boldsymbol{\theta}})' Q^{-1} (\mathbf{x} - \bar{\boldsymbol{\theta}})}\right)(\mathbf{x} - \bar{\boldsymbol{\theta}}) \tag{12}$$

is a JSE for shrinkage toward an arbitrary point target $\bar{\boldsymbol{\theta}} \in \mathbb{R}^p$. (We will not notationally distinguish the JSE of (12) from that of (10); the shrinkage target will always be specified or clear from context.) If $\bar{\boldsymbol{\theta}}$ is a good guess for the true value of $\boldsymbol{\theta}$, then a significant reduction in MSE will result. Of course, if $\boldsymbol{\theta}$ in fact lies far from $\bar{\boldsymbol{\theta}}$, then we will achieve only a small improvement over the MLE, as was the case when the JSE was applied to shrink toward $\mathbf{0}$. The freedom from a pre-specified frame of reference afforded by (12) is essential to the development of the James-Stein state filter in [12] and the multiple shrinkage theory in [14].

Note that (12) evokes the form of the BLSE of $\boldsymbol{\theta}$ from $\mathbf{x}$ given a Gaussian prior on $\boldsymbol{\theta}$ with mean $\bar{\boldsymbol{\theta}}$. As alluded to in Section 2.2.2, the empirical Bayesian argument can be used to defend (12) on exactly these grounds. This offers the following intuitive explanation of the JSE: it is an attempt to mimic Bayesian estimation in the absence of an explicit prior on $\boldsymbol{\theta}$, but where the prior is assumed Gaussian with mean $\bar{\boldsymbol{\theta}}$ specified in advance by the user and covariance estimated on the fly from $\mathbf{x}$.

2.3.3 The Positive-Part James-Stein Estimator

Just as the JSE dominates the MLE, there are estimators that dominate the JSE.[5] In this section we examine a simple modification to the JSE of (4) and (10) that provides a JSE-dominating estimator that is very difficult to improve upon (*i.e.*, it is very difficult to dominate) [22, 8]. This estimator, known as the positive-part James-Stein estimator (JSE$^+$), was first proposed in 1964 by Baranchik [23].

The motivation for the JSE$^+$ arises from comparison of the JSE of (4) to the optimal Bayes estimator (9) under the assumption of a prior on $\boldsymbol{\theta}$, as in the empirical Bayesian framework examined in Section 2.2.2. Both estimates are of the form $\hat{\boldsymbol{\theta}} = \beta\mathbf{x}$. The β multiplier for the BLSE satisfies $0 < \beta < 1$, so that the elements of $\hat{\boldsymbol{\theta}}^{\mathrm{BLS}}$ always have the same signs as their counterparts in $\mathbf{x}$. The multiplier for the JSE, however, satisfies $-\infty < \beta < 1$. The shrinkage coefficient in the JSE can be an arbitrarily large negative number if $\|\mathbf{x}\|^2$ is small. In this case the JSE essentially *overshoots* its shrinkage target and $\hat{\boldsymbol{\theta}}^{\mathrm{JS}}$ lies across the target from $\mathbf{x}$—something that never happens in Bayesian estimation. This suggests that the JSE might be improved by limiting the shrinkage coefficient to be positive. The JSE$^+$ is obtained in just such a way.

Let us define notation x^+ to signify

$$x^+ = \begin{cases} x, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

[5]Strawderman [19] demonstrated a shrinkage estimator that dominates the JSE; unfortunately, choosing the shrinkage coefficient entails numerical integration or approximation. Shao and Strawderman suggested another dominating estimator in 1994 [20]. Guo and Pal [21] also demonstrated a series of estimators that dominate the JSE. None of these estimators are admissible.

for any scalar x. Then the JSE$^+$ is defined as

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}^+} = \left(1 - \frac{\tilde{p}-2}{\mathbf{x}'Q^{-1}\mathbf{x}}\right)^+ \mathbf{x} \tag{13}$$

for the general case where $\mathbf{x} \sim N(\boldsymbol{\theta}, Q)$. It can be shown [10] that the general-form JSE$^+$ dominates the JSE and thus also the MLE whenever $\tilde{p} > 2$. In fact, the generalized JSE$^+$ of the form

$$\hat{\boldsymbol{\theta}}^{\mathrm{JS}^+} = \left(1 - \frac{c}{\mathbf{x}'Q^{-1}\mathbf{x}}\right)^+ \mathbf{x}$$

dominates the JSE for $0 < c < 2(\tilde{p} - 2)$, again assuming $\tilde{p} > 2$. A heuristic depiction of the improvement of the JSE$^+$ upon the JSE and MLE (for the case where $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$) is given in Figure 1. There do exist estimators that dominate the JSE$^+$ [19, 20, 21], but their improvement is small and they lack the simplicity of the JSE$^+$ [8].

2.4 Stein's Unbiased Risk Estimate

Before proceeding to examine applications of the James-Stein estimators [12, 13, 14], we turn our attention to one final detail that will prove essential to the investigation of wavelet shrinkage in [13] in Section 4 and useful to the investigation of multiple shrinkage estimators [14] in Section 5. Recall the specification of the risk of the JSE in (6). This risk depended on the value of the unknown parameter $\boldsymbol{\theta}$. This presents a difficulty in the analysis of the performance of the estimator in practice: the true value of $\boldsymbol{\theta}$ will necessarily be unknown in any application in which an estimate for $\boldsymbol{\theta}$ is sought! Stein addresses this problem in [24], proposing what has come to be known as Stein's Unbiased Risk Estimate (SURE).

Consider again the canonical problem of (1) in which an estimate of the mean $\boldsymbol{\theta}$ of a multivariate normal with covariance $\sigma^2 I$ is sought from an observation $\mathbf{x}$. Let $\hat{\boldsymbol{\theta}}$ be any estimate that can be expressed as

$$\hat{\boldsymbol{\theta}} = \mathbf{x} + \mathbf{g}(\mathbf{x})$$

where $\mathbf{g} : \mathbb{R}^p \to \mathbb{R}^p$ is an almost differentiable function[6] for which

$$E\left(\sum_{i=1}^{p} \left|\frac{\partial}{\partial x_i} g_i(\mathbf{x})\right|\right) < \infty.$$

Then Stein shows [24] that

$$E\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\mathbf{x})\|^2\right) = p\sigma^2 + E\left(\|\mathbf{g}(\mathbf{x})\|^2 + 2\sigma^2 \nabla \cdot \mathbf{g}(\mathbf{x})\right),$$

where $\nabla \cdot \mathbf{g}(\mathbf{x}) = \sum_{i=1}^{p} \frac{\partial}{\partial x_i} g_i(\mathbf{x})$. In other words, if the required conditions on $\mathbf{g}$ are met, then

$$\mathrm{SURE}_{\hat{\boldsymbol{\theta}}}(\mathbf{x}) = p\sigma^2 + \|\mathbf{g}(\mathbf{x})\|^2 + 2\sigma^2 \nabla \cdot \mathbf{g}(\mathbf{x}) \tag{14}$$

is an unbiased estimate of the risk of $\hat{\boldsymbol{\theta}}$. Additionally, it is possible to generalize (14) to the broader scenario in which the covariance of $\mathbf{x}$ is arbitrary [25], but because the analyses of [13] and [14]

[6] A function $g : \mathbb{R}^p \to \mathbb{R}^p$ is said to be *almost differentiable* if for each of its coordinate functions $g_i(\mathbf{x})$ there exists a function $\nabla g_i(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^p$ such that for all $\mathbf{z} \in \mathbb{R}^p$, we have $g_i(\mathbf{x}+\mathbf{z}) - g_i(\mathbf{x}) = \int_0^1 (\nabla g_i(\mathbf{x} + t\mathbf{z}))\mathbf{z}\, dt$ for almost all $\mathbf{x} \in \mathbb{R}^p$.

only require consideration of an identity-multiple covariance, we will not present the more general result.

Stein's result is quite general; the restrictions on $\hat{\boldsymbol{\theta}}$ for (14) to hold are very minor indeed. It is simple to show that for the JSE of (4), the unbiased risk estimate is given by

$$\text{SURE}_{\text{JS}}(\mathbf{x}) = p\sigma^2 - \frac{\sigma^4(p-2)^2}{\mathbf{x}'\mathbf{x}}. \tag{15}$$

Note that, as expected, $\text{SURE}_{\text{JS}}(\mathbf{x})$ is smaller than the MLE risk $p\sigma^2$ for any value of $\mathbf{x}$. Note also that $\text{SURE}_{\text{JS}}(\mathbf{x})$ might give a negative estimate for risk; SURE only ensures an unbiased estimate, not a sensible one. Finally, note that although SURE is guaranteed to be unbiased, we have specified no bounds on its accuracy. Stein does not present any such analysis in [24].

2.5 James-Stein Estimation Example

To illustrate the utility of the JSE in estimation problems, we present results from an example published by Efron and Morris in [26] (and rehashed in [4]). This example concerns the estimation of batting averages from observations made during the first few weeks of play of the 1970 baseball season. Efron and Morris use a player's batting average (defined as the number of hits divided by the number of times at bat) through the first few weeks of play as measurements from which that player's "true batting ability" may be estimated. Efron and Morris assume that a player's batting average for the remainder of the season (*i.e.*, after the observations) is a good approximation of the player's unobservable true batting ability; thus they pose the problem as one of estimating players' remainder-of-season batting averages from measurements of their averages shortly into the season. To collect a random sample of players throughout the major leagues, Efron and Morris consider only those players who had exactly 45 at-bats at the conclusion of play on April 26, 1970.[7] The resulting sample consists of eighteen players.

Following the notation of earlier sections, let $\boldsymbol{\theta} \in \mathbb{R}^{18}$ be the vector of true batting abilities and let $\mathbf{x} \in \mathbb{R}^{18}$ be the vector of observations made on April 26. Efron and Morris make several reasonable assumptions to facilitate application of the JSE: they model each at-bat as a Bernoulli trial, where the probability of player i getting a hit is θ_i. By a central-limit-theorem argument, each x_i is roughly distributed as $N(\theta_i, \frac{1}{45}\theta_i(1-\theta_i))$. Efron and Morris apply a "variance-stabilizing" transform to equalize the variances of the measurements, defining

$$y_i = \sqrt{45}\arcsin(2x_i - 1). \tag{16}$$

The y_i are then roughly distributed as $N(\psi_i, 1)$, where $\psi_i = \sqrt{45}\arcsin(2\theta_i - 1)$.[8] The problem then has the same form as the canonical problem: estimation of $\boldsymbol{\psi}$ from an observation $\mathbf{y}$ distributed roughly as $N(\boldsymbol{\psi}, I)$. The obtained estimates $\hat{\boldsymbol{\psi}}$ may be transformed back to batting averages using the inverse of the transformation of (16).

Efron and Morris implement the JSE, shrinking toward $\bar{\mathbf{y}} = (\frac{1}{18}\sum_{i=1}^{18} y_i)\mathbf{1}_{18}$ and using a shrinkage coefficient that gives an estimator of the form

$$\hat{\boldsymbol{\psi}}^{\text{JS}} = \bar{\mathbf{y}} + \left(1 - \frac{15}{(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}})}\right)(\mathbf{y}-\bar{\mathbf{y}}). \tag{17}$$

[7] This represents roughly one-tenth of a full season.

[8] Normality is roughly preserved because in the range of typical batting averages, the transformation is nearly linear with slope $\sqrt{45}(\theta_i(1-\theta_i))^{-1/2}$. Efron and Morris show that for $.15 < \theta_i < .85$, a standard deviation of 1 ± 0.036 is obtained.

player	$x_i = \hat{\theta}_i^{\mathrm{ML}}$	$\hat{\theta}_i^{\mathrm{JS}}$	θ_i	$y_i = \hat{\psi}_i^{\mathrm{ML}}$	$\hat{\psi}_i^{\mathrm{JS}}$	ψ_i
Roberto Clemente (Pitt, NL)	.400	.290	.346	-1.35	-2.91	-2.10
Frank Robinson (Balt, AL)	.378	.286	.298	-1.66	-2.97	-2.79
Frank Howard (Wash, AL)	.356	.281	.276	-1.97	-3.04	-3.11
Jay Johnstone (Cal, AL)	.333	.277	.222	-2.28	-3.10	-3.96
Ken Berry (Chi, AL)	.311	.273	.273	-2.60	-3.16	-3.17
Jim Spencer (Cal, AL)	.311	.273	.270	-2.60	-3.16	-3.20
Don Kessinger (Chi, NL)	.289	.268	.263	-2.92	-3.24	-3.32
Luis Alvarado (Bos, AL)	.267	.264	.210	-3.26	-3.30	-4.15
Ron Santo (Chi, NL)	.244	.259	.269	-3.60	-3.37	-3.23
Ron Swoboda (NY, NL)	.244	.259	.230	-3.60	-3.37	-3.83
Del Unser (Wash, AL)	.222	.254	.264	-3.95	-3.45	-3.30
Billy Williams (Chi, NL)	.222	.254	.256	-3.95	-3.45	-3.43
George Scott (Bos, AL)	.222	.254	.303	-3.95	-3.45	-2.71
Rico Petrocelli (Bos, AL)	.222	.254	.264	-3.95	-3.45	-3.30
Ellie Rodriguez (KC, AL)	.222	.254	.226	-3.95	-3.45	-3.89
Bert Campaneris (Oak, AL)	.200	.249	.285	-4.32	-3.53	-2.98
Thurman Munson (NY, AL)	.178	.244	.316	-4.70	-3.61	-2.53
Max Alvis (Mil, NL)	.156	.239	.200	-5.10	-3.68	-4.32

Table 1: James-Stein Estimation Example—1970 Season Batting Averages

These estimates are compared to the MLE of $\boldsymbol{\psi}$, given simply by $\mathbf{y}$. The loss of the MLE and the JSE can be calculated exactly because $\boldsymbol{\psi}$ is available from the players' end-of-season statistics.

The results of the estimation are presented Table 1, in terms of both batting averages and the transformed quantities. The performance of the JSE is superior to that of the MLE. The total squared error of the transformed JSE, $\|\hat{\boldsymbol{\psi}}^{\mathrm{JS}} - \boldsymbol{\psi}\|^2$, is 5.01; this is much smaller than the total squared error of the transformed MLE, $\|\hat{\boldsymbol{\psi}}^{\mathrm{ML}} - \boldsymbol{\psi}\|^2$, which is 17.56. The same performance is observed in terms of batting averages: $\|\hat{\boldsymbol{\theta}}^{\mathrm{JS}} - \boldsymbol{\theta}\|^2$ is 0.022, while $\|\hat{\boldsymbol{\theta}}^{\mathrm{ML}} - \boldsymbol{\theta}\|^2$ is 0.077. Note that although the JSE achieves a lower total squared error than the MLE, it does not produce a reduction in the individual squared error of each player's estimate: the JSE increases the squared error of the estimates of the batting averages of Clemente, Swoboda, and Rodriguez.

In [4], Efron and Morris comment on the range of applicability of the JSE by discussing estimation of an augmented $\boldsymbol{\theta}$ that not only comprises batting averages, but completely unrelated statistics such as the fraction of cars in Chicago that are imported and the incidence of toxoplasmosis in Central American cities. Efron and Morris comment that although the JSE can be applied to this aggregation to dominate the MLE, its use might be ill-advised: there is no guarantee that the MSE of each subcategory (batting averages, importation fraction, toxoplasmosis incidence) will be smaller than that obtained by using the MLE to estimate the mean of that subcategory. In practice it might be better to implement separate JSEs for each subcategory, unless we care about the total ensemble batting average/importation fraction/toxoplasmosis incidence squared error for some arcane reason.

3 James-Stein Dynamic State Estimation

The Kalman filter is one of the pillars of statistical signal processing and has found application in possibly every area in which a dynamic state-space model can be used to describe the behavior of a system. The Kalman filter (KF) is the minimum mean-squared error (MMSE) filter for producing an estimate of the state of a system, given perfect knowledge of the state-space model describing a linear Gaussian system. It produces an MMSE estimate of the state of the system at any time, assuming knowledge of the model parameters and the adherence to that model of the system under consideration. It is well-known, however, that if the state-space model is inaccurate or subject to perturbations, or if the process or measurement noise distributions are not truly Gaussian, then the state estimates produced by the Kalman filter are not only suboptimal, but can be arbitrarily bad. In other words, if the state and measurement dynamics are known perfectly, the Kalman filter produces a MMSE state estimate; if this knowledge is incorrect or imperfect, all bets are off.

In 1998, Manton, Krishnamurthy, and Poor [12] proposed a robust Kalman filter based on the JSE that they called the James-Stein State Filter (JSSF). Many researchers have proposed "locally robust" Kalman filters that assume the model parameters are subject to norm-bounded perturbations [27, 28] or non-Gaussian (but known) noise distributions [29, 30]. The JSSF is globally robust, in the sense that regardless of model perturbations or non-Gaussian noise distributions, the MSE of the state estimate produced by the JSSF is guaranteed to be no larger than a certain upper bound. In this section we examine the JSSF and its close cousin the James-Stein Kalman Filter with Hypothesis Test ($\mathrm{JSKF_H}$), also proposed in [12].

The JSSF is well suited to applications in which the true system dynamics are unknown or poorly understood and where any model for system behavior is likely to be incorrect and imperfect, possibly grossly so. (Applications of this type suggested in [12] include economic analysis and meteorological forecasting.) The JSSF is a essentially a recursive implementation of the JSE that produces a state estimate that is guaranteed to have MSE no greater than the MLE of the state based on the current observation, regardless of how incorrect the state model is or how non-Gaussian the process noise is. When the state-space model assumed by the Kalman filter is accurate, this MSE will likely be significantly greater than that of the Kalman filter, which utilizes all past observations to estimate the current state. When the model is inappropriate, however, the JSSF often produces a much better estimate than the standard Kalman filter. The JSSF is thus a good candidate for application in problems where the available state-space model is ad hoc or crude.

The $\mathrm{JSKF_H}$ is tailored to the analysis of systems in which the nominal dynamics are known and properly modeled, but are subject to occasional hard-to-model deviations. An example of this type of application is target tracking by a sensor or imaging system, in which targets typically behave according to some simple dynamical rules (*e.g.*, straight-line motion), but occasionally maneuver in an abrupt or complicated way. In applications such as these, the JSSF and $\mathrm{JSKF_H}$ provide attractive alternatives to the standard Kalman filter.

The range of applicability of the JSSF and $\mathrm{JSKF_H}$ is limited by the requirement that the dimension of the observation vector be at least as great as that of the state, a condition that is not required for the application of the standard Kalman filter. This requirement arises from the application of the JSE to a general linear regression problem Manton *et al.* use as the foundation for the JSSF and $\mathrm{JSKF_H}$. We will now turn our attention to this regression.

3.1 James-Stein Estimation for Linear Regression

Consider the following problem: we wish to obtain an estimate of the vector $\mathbf{x} \in \mathbb{R}^p$ given an observation $\mathbf{z} \in \mathbb{R}^n$ of the form

$$\mathbf{z} = C\mathbf{x} + D\mathbf{w} \tag{18}$$

where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2 I)$ and where C and D are known $n \times p$ and $n \times n$ real matrices, respectively. If C and D both have full column rank, with $n \geq p$, then the regression yields a unique least-squares solution that corresponds to the MLE of $\mathbf{x}$ based on $\mathbf{z}$:

$$\hat{\mathbf{x}}^{\mathrm{ML}} = (C'(DD')^{-1}C)^{-1}C'(DD')^{-1}\mathbf{z}.$$

This estimate satisfies $\hat{\mathbf{x}}^{\mathrm{ML}} \sim N(\mathbf{x}, \sigma^2(C'(DD')^{-1}C)^{-1})$ [15]. If $n < p$, then (18) does not admit a unique least-squares solution and there is not a unique MLE; in this case uniqueness would require some regularization of the problem, such as the addition of a penalty on the norm of the estimate.

Assuming that $n \geq p$, we can apply the results of Section 2 to produce a JSE of $\mathbf{x}$ based on $\mathbf{z}$. We will employ the positive-part James-Stein estimate (JSE$^+$) of (13) because this estimator dominates the traditional JSE of (4). Premultiplying both sides of (18) by $(C'(DD')^{-1}C)^{-1}C'(DD')^{-1}$ yields

$$\mathbf{y} = \mathbf{x} + \mathbf{v},$$

where $\mathbf{y} = (C'(DD')^{-1}C)^{-1}C'(DD')^{-1}\mathbf{z}$ and $\mathbf{v} = (C'(DD')^{-1}C)^{-1}C'(DD')^{-1}D\mathbf{w}$, so that $\mathbf{v} \sim N(\mathbf{0}, \sigma^2(C'(DD')^{-1}C)^{-1})$. Noting that $\mathbf{y} = \hat{\mathbf{x}}^{\mathrm{ML}}$, (13) yields the following form for the JSE$^+$:

$$\hat{\mathbf{x}}^{\mathrm{JS}} = \left(1 - \frac{\sigma^2 c}{\hat{\mathbf{x}}^{\mathrm{ML}\prime}(C'(DD')^{-1}C)\hat{\mathbf{x}}^{\mathrm{ML}}}\right)^+ \hat{\mathbf{x}}^{\mathrm{ML}},$$

where c can be chosen to be any constant satisfying $0 < c < 2(\tilde{p} - 2)$ so that the JSE$^+$ dominates the MLE. (Here $\tilde{p}$ is the effective dimension of $\sigma^2(C'(DD')^{-1}C)^{-1}$ as defined in (11).) Manton *et al.* set $c = (\min\{(p-2), 2(\tilde{p}-2)\})^+$.[9] This choice of c is clearly nonnegative and bounded above by $2(\tilde{p} - 2)$, so the estimator will in fact dominate the MLE whenever $\tilde{p} > 2$. Additionally, when $\tilde{p} < 2$, $c = 0$ and thus the JSE$^+$ reverts to the MLE. This choice of c results in the specific form of the JSE used in the implementation of the JSSF and JSKF$_{\mathrm{H}}$:

$$\hat{\mathbf{x}}^{\mathrm{JS}} = \left(1 - \frac{\sigma^2(\min\{(p-2), 2(\tilde{p}-2)\})^+}{\hat{\mathbf{x}}^{\mathrm{ML}\prime}(C'(DD')^{-1}C)\hat{\mathbf{x}}^{\mathrm{ML}}}\right)^+ \hat{\mathbf{x}}^{\mathrm{ML}}. \tag{19}$$

This estimator can be trivially modified as in (12) to shrink toward a target other than the origin.

3.2 The Kalman Filter

The Kalman filter has imbued the field of statistical signal processing so deeply that it is scarcely necessary to repeat its assumptions and form. However, for the sake of convenience for the analysis that will follow, we repeat the basic equations and assumptions here.

The discrete-time Kalman filter produces an evolving estimate of the state of a dynamic system based on a state-space model of that system:

$$\mathbf{x}_{k+1} = A_k\mathbf{x}_k + B_k\mathbf{e}_{k+1}, \tag{20}$$
$$\mathbf{z}_k = C_k\mathbf{x}_k + D_k\mathbf{w}_k, \tag{21}$$

where k is a discrete time index, $\mathbf{x}_k \in \mathbb{R}^p$ is the state vector, $\mathbf{z}_k \in \mathbb{R}^n$ is the observation vector, $\mathbf{e}_k \in \mathbb{R}^r$ is the process noise vector, $\mathbf{w}_k \in \mathbb{R}^n$ is the observation noise vector, and $A_k \in \mathbb{R}^{p\times p}, B_k \in \mathbb{R}^{p\times r}, C_k \in \mathbb{R}^{n\times p}$, and $D_k \in \mathbb{R}^{n\times n}$ are matrices describing the behavior of the state and measurement of the system at time k. The time index is usually assumed to begin at 0 and observations $\mathbf{z}_k$ are sequentially available for all $k \geq 0$ as time progresses.

[9]The authors of [12] present a somewhat vague and abstruse argument for the presence of the $2(\tilde{p} - 2)$ term as opposed to the more standard $\tilde{p} - 2$. They admit that this choice is arbitrary and tangential to the conceptualization and implementation of the JSSF and JSKF$_{\mathrm{H}}$.

The MSE optimality of the Kalman filter relies on a number of assumptions. Both noise processes must be Gaussian, temporally white, and independent of each other and $\mathbf{x}_0$, which must also be Gaussian. These quantities must have known distributions $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$, $\mathbf{w}_k \sim N(\mathbf{0}, \sigma^2 I)$, and $\mathbf{x}_0 \sim N(\hat{\mathbf{x}}^{\mathrm{KF}}_{0|-1}, P^{\mathrm{KF}}_{0|-1})$. Finally, the system behavior must be completely described by (20) and (21).

Given the above model and assumptions, the Kalman filter for MMSE estimation of $\mathbf{x}_k$ from $\{\mathbf{z}_0, \ldots, \mathbf{z}_k\}$ can be written in the form

$$
\begin{aligned}
\hat{\mathbf{x}}^{\mathrm{KF}}_{k|k} &= \hat{\mathbf{x}}^{\mathrm{KF}}_{k|k-1} + K_k(\mathbf{z}_k - C_k \hat{\mathbf{x}}^{\mathrm{KF}}_{k|k-1}), && (22)\\
K_k &= P^{\mathrm{KF}}_{k|k-1} C_k' (C_k P^{\mathrm{KF}}_{k|k-1} C_k' + \sigma^2 D_k D_k')^{-1}, && (23)\\
\hat{\mathbf{x}}^{\mathrm{KF}}_{k+1|k} &= A_k \hat{\mathbf{x}}^{\mathrm{KF}}_{k|k}, && (24)\\
P^{\mathrm{KF}}_{k+1|k} &= A_k (I - K_k C_k) P^{\mathrm{KF}}_{k|k-1} A_k' + B_k Q_{k+1} B_k'. && (25)
\end{aligned}
$$

Each $\hat{\mathbf{x}}^{\mathrm{KF}}_{k|k}$ is the MMSE estimate of the state at time k based on observations $\mathbf{z}_0, \ldots, \mathbf{z}_k$; each $\hat{\mathbf{x}}^{\mathrm{KF}}_{k+1|k}$ is the MMSE estimate of the state at time $k+1$ based on the same observations. The matrix $P^{\mathrm{KF}}_{k+1|k}$ is the covariance of the error of the latter estimate. Typically (22) and (23) are known as the *update equations*, and (24) and (25) as the *prediction equations*.[10]

3.3 Derivation of the James-Stein State Filter

Suppose now that although the observation model in (21) is believed to hold, with $\mathbf{w}_k$ Gaussian, temporally white, and independent of any process noise, the state dynamic model of (20) is believed to be inaccurate or incorrect and the process noise $\mathbf{e}_k$ is not necessarily Gaussian or temporally white. Then the Kalman filter is no longer guaranteed to produce a MMSE estimate; indeed, the MSE of the estimate it produces can be arbitrarily large. If the state dynamics are so poorly known that any dependence of the current state on previous states is almost purely speculative, it might make sense to estimate $\mathbf{x}_k$ directly from $\mathbf{z}_k$ and not from the collection of past observations via the incorrect model assumed by (20). Assuming that C_k and D_k both have full column rank (remember, this requires $n \geq p$), the MLE of $\mathbf{x}_k$ based on $\mathbf{z}_k$ is given by

$$\hat{\mathbf{x}}^{\mathrm{ML}}_k = (C_k'(D_k D_k')^{-1} C_k)^{-1} C_k' (D_k D_k)^{-1} \mathbf{z}_k. \tag{26}$$

However, because we know that the JSE dominates the MLE, it might make more sense to use the JSE instead, as prescribed by (19). Now, recall that the JSE can be modified to allow for shrinkage toward an arbitrary target instead of the origin; recall also that the risk of the JSE is minimized when the true value of the parameter being estimated is near the shrinkage target. Thus it would be wise to shrink toward any available prior estimate of $\mathbf{x}_k$. Manton *et al.* suggest that the dynamic state model of (20) offers one such prior estimate.[11] Specifically, given the estimate $\hat{\mathbf{x}}^{\mathrm{JS}}_k$ we could produce an estimate $\hat{\mathbf{x}}^{\mathrm{JS}}_{k+1}$ by shrinking $\hat{\mathbf{x}}^{\mathrm{ML}}_{k+1}$ toward $A_k \hat{\mathbf{x}}^{\mathrm{JS}}_k$. Regardless of how inaccurate a prior estimate of $\mathbf{x}_{k+1}$ the prediction $A_k \hat{\mathbf{x}}^{\mathrm{JS}}_k$ provides, it is guaranteed that the risk of an estimate $\hat{\mathbf{x}}^{\mathrm{JS}}_{k+1}$ obtained by using the JSE with shrinkage toward $A_k \hat{\mathbf{x}}^{\mathrm{JS}}_k$ will not exceed the risk of the MLE. At the same time, if $A_k \hat{\mathbf{x}}^{\mathrm{JS}}_k$ turns out to be a reasonably accurate shrinkage target, then there will be a substantial reduction in MSE.

[10] The often-explicit covariance update equation has been absorbed into the covariance prediction equation (25) for convenience.

[11] Admittedly, the invalidity of this model was the motivation for estimating $\mathbf{x}_k$ based only on $\mathbf{z}_k$ instead of the full set of observations; however, in lieu of any other modeled information about the behavior of the state, application of this model to yield a prior estimate might well be better than arbitrary shrinkage toward $\mathbf{0}$.

The above ideas are formalized in the following recursive specification of the JSSF in [12]:

$$\hat{\mathbf{x}}^{\mathrm{JS}}_{k|k} = \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1} + \left(1 - \frac{\sigma^2(\min\{(p-2), 2(\tilde{p}-2)\})^+}{(\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1})' C'_k (D_k D'_k)^{-1} C_k (\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1})}\right)^+ (\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1}), \quad (27)$$

$$\hat{\mathbf{x}}^{\mathrm{JS}}_{k+1|k} = A_k \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k}, \quad (28)$$

where $\hat{\mathbf{x}}^{\mathrm{ML}}_k$ is specified in (26) and $\tilde{p}$ is the effective dimension of $\sigma^2(C'_k(D_kD'_k)^{-1}C_k)^{-1}$ as defined by (11). The initialization $\hat{\mathbf{x}}^{\mathrm{JS}}_{0|-1}$ implicitly required to start the estimation process of (27) and (28) can be obtained by any prior estimate of $\mathbf{x}_0$, if available, or simply setting $\hat{\mathbf{x}}^{\mathrm{JS}}_{0|-1} = \mathbf{0}$ if none is available. Again, it is important to stress that the implementation of the JSSF requires that the dimension of the observation vector is at least as great as that of the state—if this condition is not met, then the JSSF *cannot* be implemented as specified because $C_k(D_kD'_k)^{-1}C_k$ will not be invertible, *i.e.*, $\hat{\mathbf{x}}^{\mathrm{ML}}_k$ as defined in (26) will not exist.[12]

The JSSF of (27) and (28) is guaranteed to have a risk $J^{\mathrm{JSSF}}_k(\mathbf{x}_k)$ no greater than that of $\hat{\mathbf{x}}^{\mathrm{ML}}_k$, $\sigma^2\mathrm{tr}((C'_k(D_kD'_k)^{-1}C_k)^{-1})$, at each time step k. Additionally, the more accurate a model (20) provides for the true system behavior, the smaller $J^{\mathrm{JSSF}}_k(\mathbf{x}_k)$ will be. $J^{\mathrm{JSSF}}_k(\mathbf{x}_k)$ is *not* guaranteed to be smaller than the MSE of the Kalman filter estimate $\hat{\mathbf{x}}^{\mathrm{KF}}_{k|k}$, although if the state dynamic model is "incorrect enough," this will likely be the case. Note that the JSSF of (27) and (28) can be applied regardless of how incorrect the state dynamic model (20) is, regardless of how non-Gaussian or temporally non-white the process noise $\mathbf{e}_k$ is, and can even be applied in the complete absence of a state dynamic model. Furthermore, the JSSF prediction equation (28) is somewhat arbitrary and can be replaced with any other prediction step that seems to make more sense or bear a closer relationship to the true dynamics of the system. For example, if the modeled state transitions of (20) are a linearized version of some nonlinear state dynamic behavior, then the nonlinear model can be used directly in place of (28), *i.e.*, we can specify $\hat{\mathbf{x}}^{\mathrm{JS}}_{k+1|k} = \phi(\hat{\mathbf{x}}^{\mathrm{JS}}_{k|k})$ for any $\phi : \mathbb{R}^p \to \mathbb{R}^p$. Similarly, if there is some reason to believe that the state at some time instant might be near a certain value, this value as the prediction for that time step in place of (28). Finally, we note that the JSSF and Kalman filter have the same computational order, p^3.

The most severe limitation of the JSSF is that the state dimension of the system under consideration must be no greater than the observation dimension. Another valid criticism is the somewhat hypocritical view the JSSF appears to take of the state dynamic equation (20): the model is seemingly too poor to merit use of the Kalman filter, yet good enough to use as the prediction step of the JSSF in (28). In defense of the JSSF, the MSE of $\hat{\mathbf{x}}^{\mathrm{JS}}_k$ will always be less than that of $\hat{\mathbf{x}}^{\mathrm{ML}}_k$ regardless of how inappropriate or incorrect (20) is; shrinkage toward $A_k\hat{\mathbf{x}}^{\mathrm{JS}}_k$ is merely an attempt to eke out any available indication of a reasonable shrinkage target on the theory that even greatly flawed knowledge might be better than no knowledge at all.

Other grounds for criticism of the JSSF stem from its comparison to the Kalman filter. The JSSF does not explicitly produce an error covariance or MSE estimate, although in the next section we will show how the $\mathrm{JSKF_H}$ utilizes an implicit estimate of $P_{k|k-1}$ produced by the JSSF at each step k.[13] Finally, it could be argued that the update/predict form of (27) and (28) is a hollow gimmick meant to evoke the form of the Kalman filter without being fundamental to the operation of the JSSF as those steps are to the Kalman filter. As stated, there is little the JSSF by itself can do to deflect this criticism; in the context of the $\mathrm{JSKF_H}$, however, this formulation of the JSSF does become fundamental.

[12] Manton *et al.* propose a modification to the JSSF in this case that allows the estimation of an observable substate of $\mathbf{x}_k$ of dimension n.

[13] Alternatively, SURE could provide a MSE estimate at each k at minimal computational cost if one were desired.

3.4 Derivation of the James-Stein Kalman Filter with Hypothesis Test

The JSSF is naturally suited to situations in which the user has little knowledge of the state dynamics of some system, possibly because the physical phenomenon being modeled is imperfectly understood or chaotic (such as might be the case in meteorological forecasting), or based on heuristic generalities (as is often the case of necessity in econometric analysis). In many situations, however, the imperfections in a dynamic state model are not a result of limited understanding, but rather the intractability of capturing certain behavior in a compact linear model. A canonical example is target tracking, in which the target being tracked usually exhibits simple, easy-to-model linear behavior, but occasionally executes a nonlinear or otherwise complicated maneuver. In cases such as these, a specified dynamic model might accurately capture the linear behavior but fail to describe the more complicated dynamics. In these situations it could be beneficial to implement an estimator that uses the power and memory of the Kalman filter during the linear dynamic phases, but falls back on a less model-based estimate when it seems likely that model has been violated. This is the motivation and intention of the JSKF_H.

A basic description of the JSKF_H is as follows: implement the Kalman filter according to a specified state-space and observation model as in (20) and (21). At each time step, calculate a test statistic that attempts to quantify how successful the previous step's prediction was in estimating the current observation. If the previous step's prediction seems to have been accurate, based on a comparison of the predicted and observed observation, use the Kalman filter's updated state estimate as the current state estimate; otherwise, calculate the state estimate using the JSSF. The only assumptions of the JSKF_H are those of the JSSF, with the addition that at least a rudimentary state dynamic model in the form of (20) is now required whereas implementation of the JSSF did not require a state dynamic model at all.

The formulation of the JSKF_H is based on the following supposition: inadequacies in the assumed state dynamics can be modeled by assuming that the matrices A_k and B_k in the model of (20) are known perfectly but that the noise term $\mathbf{e}_k$ is non-Gaussian, so that the occasional deviations from the model appear as outliers in the distribution of $\mathbf{e}_k$. Stated another way, the JSKF_H makes the assumption that at any time step k, either $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$ as modeled, or $\|\mathbf{e}_k\|^2 \gg \operatorname{tr}(Q_k)$ and is drawn from a distribution other than the Gaussian one assumed in the model. This formulation suggests that a hypothesis test based on a statistic reflecting the magnitude of $\mathbf{e}_k$ could provide an indication of whether the model is accurate at time k. When $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$, we will have $\mathbf{x}_k - \hat{\mathbf{x}}^{\text{KF}}_{k|k-1} \sim N(\mathbf{0}, P^{\text{KF}}_{k|k-1})$ and $\mathbf{z}_k - C_k\mathbf{x}_k \sim N(\mathbf{0}, \sigma^2 D_k D_k')$, and thus $\mathbf{z}_k - C_k\hat{\mathbf{x}}^{\text{KF}}_{k|k-1} \sim N(\mathbf{0}, C_k P^{\text{KF}}_{k|k-1} C_k' + \sigma^2 D_k D_k')$. This implies that when $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$,

$$\boldsymbol{\epsilon}_k = (C_k P^{\text{KF}}_{k|k-1} C_k' + \sigma^2 D_k D_k')^{-1/2}(\mathbf{z}_k - C_k\hat{\mathbf{x}}^{\text{KF}}_{k|k-1}) \sim N(\mathbf{0}, I),$$

so that $\boldsymbol{\epsilon}_k$ can be thought of as a collection of n independent, identically distributed zero-mean Gaussians with unity variance. Thus when $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$, the magnitude of the vector is a chi-squared random variable with n degrees of freedom; this magnitude is the test statistic used to decide whether $\mathbf{e}_k$ is drawn from $N(\mathbf{0}, Q_k)$. Specifically, define hypotheses

$$\begin{aligned} H_0 : &\quad \mathbf{e}_k \sim N(\mathbf{0}, Q_k), \\ H_1 : &\quad \mathbf{e}_k \not\sim N(\mathbf{0}, Q_k), \quad \|\mathbf{e}_k\|^2 \gg \operatorname{tr}(Q_k). \end{aligned}$$

Also define the test statistic

$$T_k = \boldsymbol{\epsilon}_k'\boldsymbol{\epsilon}_k = (\mathbf{z}_k - C_k\hat{\mathbf{x}}^{\text{KF}}_{k|k-1})'(C_k P^{\text{KF}}_{k|k-1} C_k' + \sigma^2 D_k D_k')^{-1}(\mathbf{z}_k - C_k\hat{\mathbf{x}}^{\text{KF}}_{k|k-1})$$

for each time k. Then the decision at time k takes the form of a hypothesis test

$$T_k \underset{H_1}{\overset{H_0}{\lesseqgtr}} T_c, \tag{29}$$

where T_c is simply a threshold that must be set to the largest magnitude of $\boldsymbol{\epsilon}_k$ for which it seems reasonable to assume $\mathbf{e}_k \sim N(\mathbf{0}, Q_k)$, possibly by consulting a table of cumulative distributions of the chi-squared random variable with n degrees of freedom in order to implement a Neyman-Pearson criterion decision rule [15]. The result of this hypothesis test indicates whether to use the JSSF (H_1) or the Kalman filter (H_0) to produce a state estimate at each time k.

Suppose an estimate $\hat{\mathbf{x}}^{\mathrm{H}}_{k-1|k-1}$ and prediction $\hat{\mathbf{x}}^{\mathrm{H}}_{k|k-1} = A_k \hat{\mathbf{x}}^{\mathrm{H}}_{k-1|k-1}$ are available at time $k-1$. At time k, another measurement $\mathbf{z}_k$ is obtained, and we can calculate the test statistic T_k and apply the hypothesis test of (29). If H_0 is declared, this is an indication that the assumed dynamic model accurately described the evolution of the system from $k-1$ to k; thus we should use the Kalman filter to calculate an updated state estimate $\hat{\mathbf{x}}^{\mathrm{H}}_{k|k}$ as in (22). If instead H_1 is declared, this is an indication that the dynamic model was invalid; thus we should use the JSSF instead of the Kalman filter to calculate the updated state estimate. More precisely, if we declare H_1, then we use

$$\hat{\mathbf{x}}^{\mathrm{H}}_{k|k} = \hat{\mathbf{x}}^{\mathrm{H}}_{k|k-1} + s_k(\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{H}}_{k|k-1}), \tag{30}$$

where s_k is the shrinkage coefficient we have been using for the JSSF as in (27):

$$s_k = \left(1 - \frac{\sigma^2(\min\{(p-2), 2(\tilde{p}-2)\})^+}{(\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1})' C_k' (D_k D_k')^{-1} C_k (\hat{\mathbf{x}}^{\mathrm{ML}}_k - \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1})}\right)^+.$$

Whether H_0 or H_1 is declared at time k, the updated estimate $\hat{\mathbf{x}}^{\mathrm{H}}_{k|k}$ can be used to produce a prediction $\hat{\mathbf{x}}^{\mathrm{H}}_{k+1|k} = A_k \hat{\mathbf{x}}^{\mathrm{H}}_{k|k}$. Specification of $P^{\mathrm{H}}_{k+1|k}$ is also required under either hypothesis in case H_0 is declared at time $k+1$. Under H_0 at time k, $P^{\mathrm{H}}_{k+1|k}$ is provided by (25). Under H_1 at time k, however, there is no clear way to obtain $P^{\mathrm{H}}_{k+1|k}$. (Indeed, this was one of our criticisms of the JSSF at the conclusion of Section 3.) Manton *et al.* appeal to the empirical Bayesian interpretation of the JSE, as described in Section 2.2.2, in order to obtain an estimate of $P^{\mathrm{H}}_{k+1|k}$ under H_1 at time k. In the empirical Bayesian context, the JSE is interpreted as implicitly estimating $P^{\mathrm{H}}_{k+1|k}$ in order that the BLSE may be emulated. (In Section 2.2.2, this took the form of implicit estimation of τ^2, by way of $\frac{\sigma^2}{\sigma^2+\tau^2}$.) Rewriting (30) after substituting in the values of $\hat{\mathbf{x}}^{\mathrm{ML}}_k$ and $\hat{\mathbf{x}}^{\mathrm{H}}_{k|k-1}$ under H_1 at time k according to (26) and (28) yields

$$\hat{\mathbf{x}}^{\mathrm{H}}_{k|k} = \hat{\mathbf{x}}^{\mathrm{JS}}_{k|k-1} + K^{\mathrm{JS}}_k(\mathbf{z}_k - C_k \hat{\mathbf{x}}^{\mathrm{H}}_{k|k-1}),$$

where

$$K^{\mathrm{JS}}_k = s_k (C_k' (D_k D_k')^{-1} C_k)^{-1} C_k' (D_k D_k')^{-1}.$$

The JSE's implicit estimation of $P_{k|k-1}$ at time k is embodied in specification K^{JS}_k, the obviously Kalman-like gain term describing the ratio in which the prior estimate and the new observation data should be combined. Rearranging (23) to express $P_{k|k-1}$ as a function of the filter gain instead of the other way around, we obtain

$$P^{\mathrm{JS}}_{k|k-1} = \sigma^2 \frac{s_k}{1-s_k} (C_k' (D_k D_k')^{-1} C_k)^{-1}.$$

This equation allows propagation of covariance information so that implementation of the Kalman filter and JSKF_{H} can continue after declaring H_1.

Summarizing and formalizing the discussion of the preceding paragraphs, we now specify the complete JSKF_{H} algorithm of [12].

1. Choose a threshold T_c prior to $k = 0$.

2. Initialize according to Kalman filter rules:

$$\begin{aligned}
\hat{\mathbf{x}}^{\text{H}}_{0|-1} &= E(\mathbf{x}_0), \\
P^{\text{H}}_{0|-1} &= E((\hat{\mathbf{x}}^{\text{H}}_{0|-1} - \mathbf{x}_0)(\hat{\mathbf{x}}^{\text{H}}_{0|-1} - \mathbf{x}_0)').
\end{aligned}$$

3. Compute test statistic

$$T_k = (\mathbf{z}_k - C_k \hat{\mathbf{x}}^{\text{H}}_{k|k-1})'(C_k P^{\text{H}}_{k|k-1} C_k' + \sigma^2 D_k D_k')^{-1}(\mathbf{z}_k - C_k \hat{\mathbf{x}}^{\text{H}}_{k|k-1}).$$

   If $T_k \leq T_c$, declare H_0 and go to step 5; otherwise, declare H_1 and proceed to step 4.

4. Calculate the James-Stein covariance prediction

$$P^{\text{JS}}_{k|k-1} = \sigma^2 \frac{s_k}{1 - s_k}(C_k'(D_k D_k')^{-1} C_k)^{-1}.$$

5. Set

$$P_{k|k-1} = \begin{cases} P^{\text{H}}_{k|k-1}, & \text{if } H_0 \text{ declared}, \\ P^{\text{JS}}_{k|k-1}, & \text{if } H_1 \text{ declared}. \end{cases}$$

   Produce state estimates, state predictions, and error covariance predictions:

$$\begin{aligned}
\hat{\mathbf{x}}^{\text{H}}_{k|k} &= \hat{\mathbf{x}}^{\text{H}}_{k|k-1} + K^{\text{H}}_k(\mathbf{z}_k - C_k \hat{\mathbf{x}}^{\text{H}}_{k|k-1}), \\
K^{\text{H}}_k &= P_{k|k-1} C_k'(C_k P_{k|k-1} C_k' + \sigma^2 D_k D_k')^{-1}, \\
\hat{\mathbf{x}}^{\text{H}}_{k+1|k} &= A_k \hat{\mathbf{x}}^{\text{H}}_{k|k}, \\
P^{\text{H}}_{k+1|k} &= A_k(I - K^{\text{H}}_k C_k) P_{k|k-1} A_k' + B_k Q_{k+1} B_k'.
\end{aligned}$$

6. Increment k, observe $\mathbf{z}_k$, and go to step 3.

The JSKF_{H} has a structure identical to that of the Kalman filter; at time k it relies on a hypothesis test based on T_k to decide whether to use the Kalman filter or JSSF update. The only difference between the two is seen to be the choice of $P_{k|k-1}$ to produce the filter gain and error covariance prediction for time k. The precise choice of T_c will determine how often the JSSF is used in place of the standard Kalman filter: as T_c approaches infinity, the JSKF_{H} reverts to the standard Kalman filter; as T_c approaches zero, the JSKF_{H} becomes the JSSF.

The chief limitation of the JSKF_{H}, like that of the JSSF, is the requirement that the observation vector have dimension at least as great as that of the state vector. Note that although the usage of the JSSF guaranteed an upper bound on the MSE of the state estimate, no such bound can be provided for the JSKF_{H} because we cannot be sure that the JSSF is being used if and only if the state dynamic model (20) breaks down. Whenever H_1 is declared, the MSE of the updated state estimate produced by the JSKF_{H} is necessarily smaller than that of the MLE; it would be hoped that it is also below that of the MLE whenever H_0 is declared, but this cannot be guaranteed. If H_0 is declared incorrectly then the MSE can be arbitrarily large; if H_1 is declared unnecessarily then the MSE might be much larger than it could be (but will still be below that of the MLE).

3.5 James-Stein Dynamic State Estimation Examples

Manton *et al.* present several examples demonstrating the utility and limitations of the JSSF and $\mathrm{JSKF_H}$ in [12], some of which we will duplicate here. The first example compares the performance of the Kalman filter and the JSSF in a system subject to perturbations; the second investigates the behavior of the Kalman filter and $\mathrm{JSKF_H}$ in a simple problem modeled on source intensity estimation by a multisensor imaging system.

For their first example, Manton *et al.* consider a system whose behavior is described by (20) and (21), with

$$A_k = \begin{bmatrix} 1.0 & -0.1 & -0.1 \\ 0.2 & 0.9 & -0.1 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}, \quad B_k = C_k = D_k = Q_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \sigma^2 = 1. \tag{31}$$

The effects of implementing the Kalman filter and JSSF on this system are examined for three cases: first, when the above model is known exactly; second, when the model is known approximately; and third, when the model is essentially unknown and an incorrect model is assumed for the implementation of the filters. In the first case (perfect knowledge), the Kalman filter and JSSF are implemented using the values specified in (31); in the second case (imperfect but approximate knowledge), the filters are implemented using the correct C_k, D_k, Q_k and σ^2, but with the elements of A_k and B_k perturbed by independent noise identically distributed as $N(0, 0.0625)$; in the third case (incorrect knowledge), the filters are implemented using the correct C_k, D_k, Q_k and σ^2, but with A_k and B_k misspecified as

$$A_k = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \qquad B_k = \begin{bmatrix} 9 & 8 & 7 \\ 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix}.$$

For each of these three implementations, 500 Monte Carlo runs were performed for both the Kalman filter and JSSF for the first 500 time steps ($k = 1, \ldots, 500$). The true risk at each time step for each filter was approximated by averaging the losses over the set of Monte Carlo runs. As a benchmark, these risks were also compared to that of the MLE of $\mathbf{x}_k$ from $\mathbf{z}_k$ at each time step. The results of the implementation of each filter for the first 100 time steps in the case of perfect knowledge is depicted in Figure 2, in the case of approximate knowledge in Figure 3 and in the case of misspecification in Figure 4. Table 2 lists the average MSE (*i.e.*, the approximate risk averaged over all 500 time steps over all runs) for the Kalman filter, JSSF, and MLE for each of the three cases described. (Note that the entries in this table are on a dB-scale.) Examining the entries of the table, as well as the figures described above, we see that as expected the Kalman filter outperforms the JSSF and MLE when the model is known exactly. However, when the model is perturbed or incorrect, the JSSF outperforms the Kalman filter, sometimes dramatically so. We also see that as expected, the JSSF always achieves a lower MSE than the MLE; this improvement is more marked when the model is correct or only slightly perturbed and not incorrect, since a correct or slightly perturbed model provides the JSSF with better shrinkage targets to use in (28) than does a completely incorrect model. This example demonstrates the utility of the JSSF in cases where knowledge of the state dynamics are imperfect.

Manton *et al.* present another example based on a model of an imaging system described in [31], a 4×4 sensor array viewing three motionless targets with time-varying intensities. They compare performance of the MLE, Kalman filter, and $\mathrm{JSKF_H}$ for estimating a state vector in $\mathbb{R}^3$ comprising

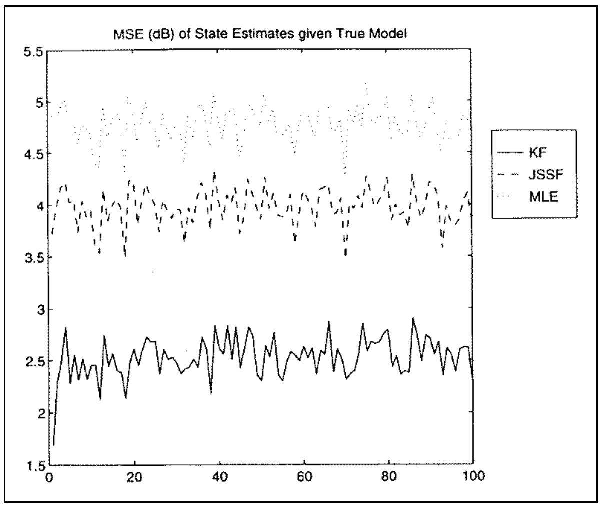

Figure 2: Performance of MLE, JSSF, and KF Given True Model (duplicated from [12])

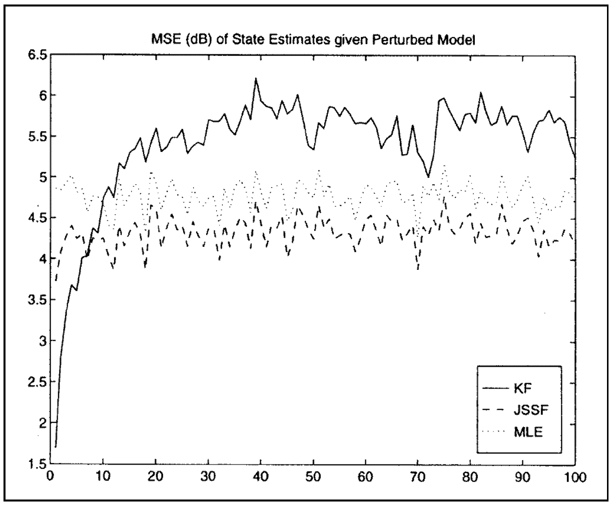

Figure 3: Performance of MLE, JSSF, and KF Given Perturbed Model (duplicated from [12])

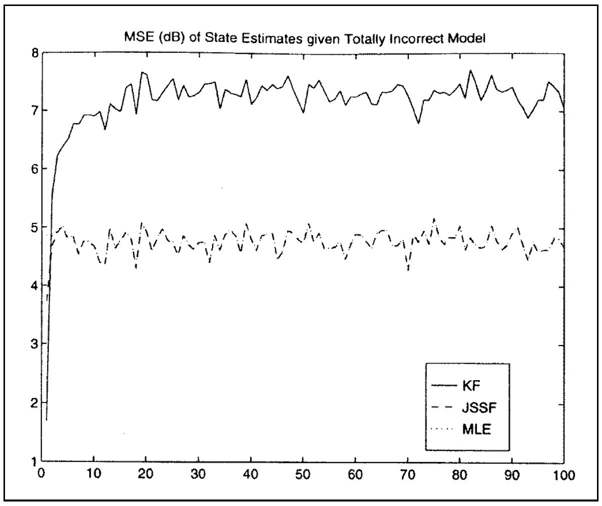

Figure 4: Performance of MLE, JSSF, and KF Given Incorrect Model (duplicated from [12])

	Average MSE (dB)			Improvement to MLE (dB)	
Model	MLE	KF	JSSF	KF	JSSF
correct	4.771	2.519	3.976	2.252	0.795
perturbed	4.771	5.595	4.331	−0.824	0.440
incorrect	4.771	7.336	4.759	−2.565	0.012

Table 2: MSEs of JSSF, KF, and MLE (from [12])

	MLE (dB)	KF (dB)	JSKF$_{\mathrm{H}}$ (dB)							
γ_{FA}	-	-	0.0001	0.0005	0.001	0.005	0.01	0.05	0.1	0.2
$P_e = 0.02$	4.772	3.635	2.785	2.780	2.780	2.817	2.859	3.129	3.349	3.620
$P_e = 0.1$	4.776	3.451	3.048	3.016	3.005	3.008	3.032	3.239	3.425	3.672

Table 3: MSEs of JSKF$_{\mathrm{H}}$, KF, and MLE (from [12])

the intensities of the targets. The true state behavior is described by the model

$$\begin{aligned} \mathbf{x}_{k+1} &= \begin{cases} \mathbf{x}_k + \mathbf{e}_{k+1}, & \text{with probability } 1 - P_e, \\ \mathbf{0}, & \text{with probability } P_e, \end{cases} \\ \mathbf{z}_k &= \mathbf{x}_k + \mathbf{w}_k, \end{aligned}$$

where all vectors are in $\mathbb{R}^3$ and where $0 < P_e < 1$, $Q_k = I$, and $\sigma^2 = 1$. The model for $\mathbf{x}_k$ indicates that the state dynamics behave according to a simple linear model a fraction $1 - P_e$ of the time, but reset to zero in a nonlinear fashion the other P_e of the time. Manton *et al.* analyze the performance of the MLE, Kalman filter, and JSKF$_{\mathrm{H}}$ on this system for the first 1000 time steps, for two different values of P_e, 0.02 and 0.1. Because implementation of the JSKF$_{\mathrm{H}}$ requires the threshold T_c to be set, results are presented for various values of the threshold; these values were set according to a Neyman-Pearson criterion. Specifically, given a probability $0 < \gamma_{FA} < 1$, T_c was set to the highest value that would achieve a probability of false alarm (defined as declaring H_1 and thus using the JSE when in fact the linear model is valid) no greater than γ_{FA}. The average losses over all 1000 data points for 500 Monte Carlo runs for the MLE, Kalman filter, and the JSKF$_{\mathrm{H}}$ with eight different thresholds γ_{FA} are presented in Table 3. We observe that the Kalman filter outperforms the MLE for both P_e tested, but that the JSKF$_{\mathrm{H}}$ outperforms the Kalman filter in almost all cases tested (the exception being for T_c set with $\gamma_{FA} = 0.2$ when $P_e = 0.1$) and outperforms the MLE in all cases tested. Furthermore, the outperformance of the Kalman filter by the JSKF$_{\mathrm{H}}$ appears to be relatively robust to the particular choice of T_c.

4 Denoising by Wavelet Shrinkage

In recent years the analysis and processing of signals using wavelet methods has become more and more common. Wavelet theory began to crystallize in the 1980s and has enjoyed a steady increase in popularity to the present day. The use of wavelets can facilitate the analysis and processing of a wide variety of signals, especially those that are nonstationary or nonlocalized in time or frequency. When properly used, wavelets have the ability to provide a parsimonious representation of a signal in terms of a small number of coefficients corresponding to weights on basis functions. For this reason, one area in which wavelets have found application is signal denoising, or the attempt to recover a signal that has been corrupted by noise [32, 33]. This is the application considered by Donoho and Johnstone in [13] and which we shall examine in this section.

Consider the following denoising problem. We observe N noisy samples of a function f:

$$y_i = f(t_i) + v_i, \quad i = 1, \ldots, N, \tag{32}$$

where t_i are evenly spaced with $t_i = (i-1)/N$, and v_i are independent and identically distributed as $N(0, \sigma^2)$. We seek an estimate of f at the time samples t_i. In vector notation, we observe

$$\mathbf{y} = \mathbf{f} + \mathbf{v}, \tag{33}$$

where all three vectors are in $\mathbb{R}^N$ and $\mathbf{v} \sim N(\mathbf{0}, \sigma^2 I)$. We seek an estimate $\hat{\mathbf{f}}$ of $\mathbf{f}$ based on $\mathbf{y}$. As usual, we want this estimate to have a small MSE, denoted by $J(\mathbf{f}; \hat{\mathbf{f}})$. Because we would in general be willing to accept a greater squared error when the number of samples in $\mathbf{f}$ is large than when it is small, we define a normalized risk

$$J_N(\mathbf{f}; \hat{\mathbf{f}}) = \frac{1}{N} J(\mathbf{f}; \hat{\mathbf{f}}) = \frac{1}{N} E\left(\|\hat{\mathbf{f}} - \mathbf{f}\|_2^2\right).$$

Much of [13] deals with the investigation of asymptotic properties of $J_N(\mathbf{f}, \hat{\mathbf{f}})$ and the search for estimators $\hat{\mathbf{f}}$ that are "nearly minimax" across a wide space of function classes. Most of this investigation is extremely technical and esoteric, and has little to do with James-Stein estimation. Our presentation of [13] will forgo most of these technical results and will be primarily concerned with the description of the two main methodologies proposed in that paper, known as *WaveJS* and *SureShrink*. They will be described in Sections 4.2 and 4.3, respectively.

4.1 Wavelet Analysis and Function Denoising

One-dimensional discrete wavelet analysis is simply the transformation of a sequence by projection onto a basis comprised of dyadic dilations and translations of some *mother wavelet* sequence. The mother wavelet can be chosen to make the basis orthogonal, and the dilated and translated basis elements can be scaled to make the basis orthonormal; the wavelet bases examined in [13] all have this property. Thus for our purposes the wavelet transform of a sequence $\mathbf{y}$ can be expressed as $\mathbf{x} = W\mathbf{y}$, with inverse transform $\mathbf{y} = W'\mathbf{x}$. Each row of W represents a specific (scaled) dilation and translation of the mother wavelet.

Returning to the formulation of (33), suppose that $N = 2^M$. The wavelet transform $\mathbf{x}$ of $\mathbf{y}$ will thus also have $N = 2^M$ terms. The elements of $\mathbf{x}$ can be indexed as $x_{j,k}$ for $j = 0, \ldots, M-1$ and $k = 0, \ldots, 2^j - 1$, where j represents the dilation or scale at which the coefficient was obtained (with $j = 0$ being coarsest) and where k represents the translation within the jth scale corresponding to the wavelet coefficient.[14]

Because of the linearity of the wavelet transform, we can represent $\mathbf{x} = W\mathbf{y}$ as

$$\mathbf{x} = \mathbf{w} + \mathbf{z}, \tag{34}$$

where $\mathbf{w}$ is the wavelet transform of $\mathbf{f}$ and where $\mathbf{z}$ is the wavelet transform of $\mathbf{v}$. Each of these vectors can be indexed using the j, k subscript notation above to denote scale and translation. The problem of estimating $\mathbf{f}$ from $\mathbf{y}$ can be posed in terms of estimating $\mathbf{w}$ from $\mathbf{x}$: any wavelet coefficient estimator $\hat{\mathbf{w}}$ can be transformed from the wavelet domain to yield $\hat{\mathbf{f}} = W'\hat{\mathbf{w}}$. Due to the orthonormality of the wavelet transform, we have $\mathbf{z} \sim N(\mathbf{0}, \sigma^2 I)$ and

$$J(\mathbf{w}; \hat{\mathbf{w}}) = J(\mathbf{f}; \hat{\mathbf{f}}). \tag{35}$$

In other words, there is an isometry of risks between the signal and wavelet domains for estimators related by a transform.

An obvious question is why estimation in the wavelet domain should provide any advantage over estimation in the signal domain. The answer is the sparsity of the wavelet-domain representation of "typical" relatively smooth signals, assuming the mother wavelet has not been chosen poorly. Wavelet domain representations of smooth signals are generally parsimonious, concentrating energy in a few coefficients while containing many zero or near-zero coefficients. Figure 5 depicts four

[14] This indexing scheme requires one additional element to complete $\mathbf{x}$, representing a signal mean not captured by projections onto dilations and translations of the mother wavelet; this element is often denoted as $x_{-1,0}$.

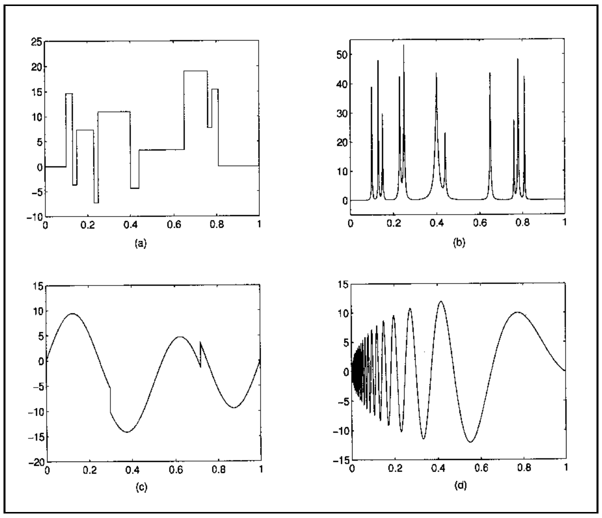

Figure 5: Four "Typical" Signals (duplicated from [13])

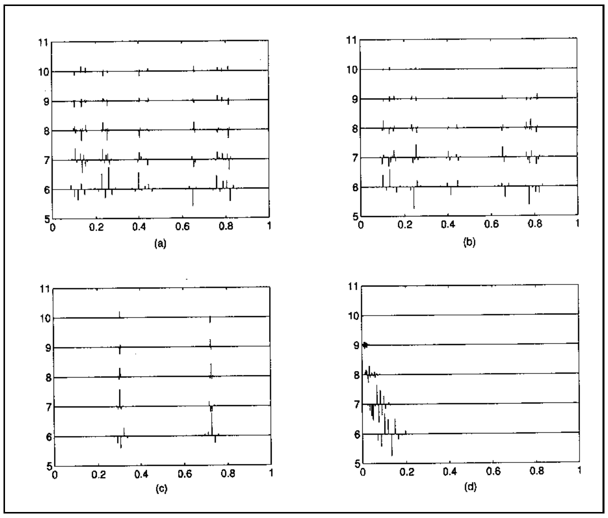

Figure 6: Wavelet Coefficients for Signals of Figure 5 (duplicated from [13])

relatively smooth signals, duplicated from [13], that might arise in various physical applications. Figure 6 depicts the wavelet transform coefficients of these signals for with $N = 2048$ for scales 6 through 10.[15] We see that most of the coefficients are indeed zero, or at least too close to zero to be resolved at the resolution of the plot.

If $\mathbf{f}$ is relatively smooth, we would expect $\mathbf{w}$ to be sparse. The observation sequence $\mathbf{y}$ is not smooth, due to the presence of additive white Gaussian noise $\mathbf{v}$; its wavelet transform $\mathbf{x}$ will lack the sparsity of $\mathbf{w}$, due to the presence of additive white Gaussian noise $\mathbf{z}$. This suggests an approach to denoising: restore sparsity to $\mathbf{x}$ by recovering the wavelet coefficients that are significantly different from zero. *WaveJS* and *SureShrink* both take this approach, but use different tools to do the job.

4.2 Derivation of *WaveJS*

Estimation of $\mathbf{w}$ from $\mathbf{x}$, given the formulation of the previous section, is simply a multivariate normal mean estimation problem with $\mathbf{x} \sim N(\mathbf{w}, \sigma^2 I)$. We lack a prior on $\mathbf{w}$ and possess only the nebulous concept of sparsity to describe what $\mathbf{w}$ should be like. This multivariate normal observation model, lack of a prior, and vague indication of a property that the estimate should possess all cry out for application of the JSE. Application of the JSE with a shrinkage target of $\mathbf{0}$ should not only help restore sparsity, but achieve a lower risk than the MLE. (Admittedly, in this situation the MLE is a dimwitted estimate that does absolutely nothing in the area of denoising.)

Donoho and Johnstone apply the JSE to the denoising problem by using a series of JSEs at different scales. (Their motivation for this approach is described below.) Let $\mathbf{x}_j$ be the collection of observed wavelet coefficients at scale j, *i.e.*, $\mathbf{x}_j = [x_{j,0}, \ldots, x_{j,2^j-1}]$. Let $\mathbf{w}_j$ and $\mathbf{z}_j$ denote similar scale-wise collections of the $w_{j,k}$ and $z_{j,k}$, respectively. (Note $\mathbf{z}_j \sim N(\mathbf{0}, \sigma^2 I)$.) Donoho and Johnstone obtain an estimate of the denoised wavelet coefficients at each scale as follows:

$$\hat{\mathbf{w}}_j^{\mathrm{JS}} = \begin{cases} \mathbf{x}_j, & 0 \leq j < L, \\ \left(1 - \frac{\sigma^2(2^j - 2)}{\mathbf{x}_j'\mathbf{x}_j}\right)^+ \mathbf{x}_j, & L \leq j \leq M-1, \end{cases} \tag{36}$$

[15] The wavelet basis used for the transform in this example, as well as the others in this section, is the "Most Nearly Symmetric Daubechies Wavelet" with $N = 8$ [34].

where $L \geq 2$ is chosen to reflect the fact that shrinkage of coarse-level signal features might be undesirable. Concatenating the $\hat{\mathbf{w}}_j^{\mathrm{JS}}$ into a coefficient vector $\hat{\mathbf{w}}^{\mathrm{JS}}$ allows an estimate of $\mathbf{f}$ to be formed directly:

$$\hat{\mathbf{f}}^{\mathrm{JS}} = W'\hat{\mathbf{w}}^{\mathrm{JS}},$$

which is precisely the estimator *WaveJS* in [13].

The decoupling of the James-Stein estimation into separate estimation problems at each scale is motivated by the following concern: if much of the signal energy is concentrated at a given scale j, then $\mathbf{x}_j'\mathbf{x}_j$ will be large and the shrinkage coefficient in (36) will be near unity, resulting in little shrinkage of the "true" signal wavelet coefficients. On the other hand, if there is little energy in scale j, then $\mathbf{x}_j'\mathbf{x}_j$ will be small and thus the shrinkage coefficient will be near zero or identically zero, reintroducing sparsity to that scale.

Summarizing, Donoho and Johnstone present the following algorithm for denoising a signal $\mathbf{y}$ according to *WaveJS*:

1. Take the discrete wavelet transform of the noisy data to yield $\mathbf{x} = W\mathbf{y}$.

2. Apply the JSE$^+$ to the collection of wavelet coefficients at each scale $j \geq L$ according to (36) to yield $\hat{\mathbf{w}}^{\mathrm{JS}}$.

3. Take the inverse wavelet transform of the estimated wavelet coefficients to yield an estimate of the denoised function $\hat{\mathbf{f}}^{\mathrm{JS}} = W'\hat{\mathbf{w}}^{\mathrm{JS}}$.

Note that *WaveJS* can be implemented with any choice of wavelet basis; the specific choice for a given application would typically depend on the details of the application.

Donoho and Johnstone demonstrate that *WaveJS* is "nearly optimal" in a certain sense for a given class of estimators. Specifically, they point out that *WaveJS* can be viewed as essentially an adaptive linear shrinkage estimator at each scale j in the wavelet domain (although the linear shrinkage coefficient used at each scale $j \geq L$ actually depends on the data in a nonlinear fashion). That is, at each scale the estimate takes the form $\hat{\mathbf{w}}_j^{\gamma} = \gamma\mathbf{x}_j$. It is easy to show that the risk of any estimator of this form is

$$J(\mathbf{w}; \hat{\mathbf{w}}_j^{\gamma}) = (1-\gamma)^2\mathbf{w}_j'\mathbf{w}_j + 2^j\sigma^2\gamma^2. \tag{37}$$

Differentiating with respect to γ and checking the second derivative, it is seen that the ideal shrinkage coefficient, *i.e.*, the choice of γ that would achieve the minimum risk, is

$$\tilde{\gamma} = \frac{\mathbf{w}_j'\mathbf{w}_j}{\mathbf{w}_j'\mathbf{w}_j + 2^j\sigma^2} = 1 - \frac{2^j\sigma^2}{\mathbf{w}_j'\mathbf{w}_j + 2^j\sigma^2}.$$

Obviously the estimator defined by this choice of shrinkage coefficient cannot be implemented, since it depends on the quantity to be estimated! Thus an optimal, *but unattainable*, linear shrinkage estimator is an oracle of the form

$$\tilde{\mathbf{w}}_j^{\mathrm{I}} = \left(1 - \frac{2^j\sigma^2}{\mathbf{w}_j'\mathbf{w}_j + 2^j\sigma^2}\right)\mathbf{x}_j. \tag{38}$$

The JSE in (36) can be viewed as an attempt to emulate this unattainable ideal estimator by estimating its shrinkage coefficient.[16] Comparing (36) and (38), we see that the JSE emulates the

[16] In fact, this is exactly the empirical Bayesian heuristic argument for the justification of the JSE provided in Section 2.2.2!

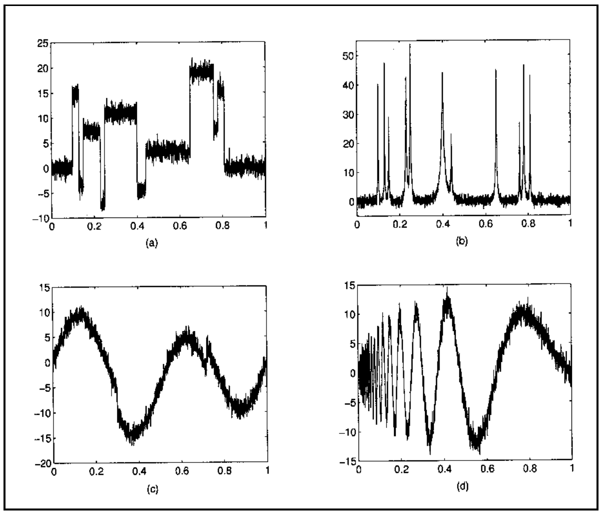

Figure 7: Signals of Figure 5 Corrupted by Noise (duplicated from [13])

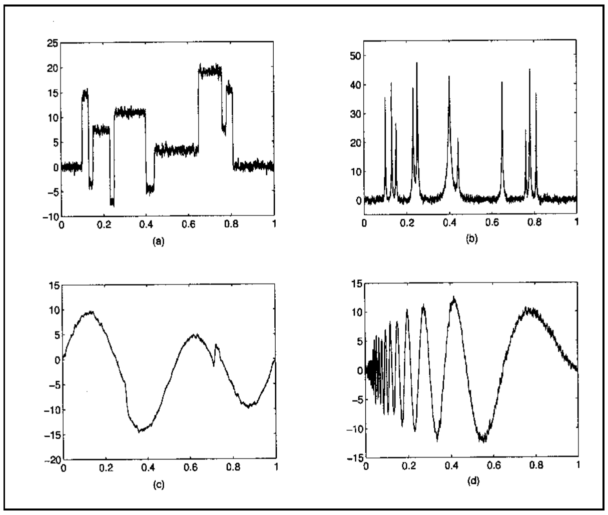

Figure 8: Performance of *WaveJS* on Noisy Signals of Figure 7 (duplicated from [13])

ideal linear shrinkage estimator by modeling $\mathbf{w}'\mathbf{w}+2^j\sigma^2$ as $\mathbf{x}'\mathbf{x}$, closely tracking $(2^j-2)\sigma^2$ by $2^j\sigma^2$, and mirroring the positive shrinkage coefficient of (38) by restriction to positivity.

The ideal linear shrinkage estimator $\tilde{\mathbf{w}}_j^{\mathrm{I}}$ in (38) can easily be shown by application of (37) to attain a MSE of

$$J(\mathbf{w};\tilde{\mathbf{w}}_j^{\mathrm{I}}) = \frac{2^j\sigma^2\mathbf{w}_j'\mathbf{w}_j}{2^j\sigma^2+\mathbf{w}_j'\mathbf{w}_j}.$$

The $\hat{\mathbf{w}}_j^{\mathrm{JS}}$ of (36), on the other hand, attains a slightly larger MSE, though the margin between the two is small. Donoho and Johnstone show that for each $j \geq L$,

$$J(\mathbf{w};\hat{\mathbf{w}}_j^{\mathrm{JS}}) \leq 2\sigma^2 + J(\mathbf{w};\tilde{\mathbf{w}}_j^{\mathrm{I}}), \tag{39}$$

again a direct analog to the empirical Bayesian result in Section 2.2.2. The scale-wise result of (39) can be used to show that

$$J_N(\mathbf{f};\hat{\mathbf{f}}^{\mathrm{JS}}) \leq J_N(\mathbf{f};\tilde{\mathbf{f}}^{\mathrm{I}}) + \frac{2^L\sigma^2+2\sigma^2\log_2(N)}{N}, \tag{40}$$

where $\tilde{\mathbf{f}}^{\mathrm{I}} = W'\tilde{\mathbf{w}}^{\mathrm{I}}$. Simply put, the result of (40) means that as we sample a function more and more finely, the per-element risk of $\hat{\mathbf{f}}^{\mathrm{JS}}$ approaches the risk of the unattainable ideal linear shrinkage estimator approximately in the manner of $\log_2 N/N$.

Figure 7 depicts the four signals of Figure 5 in the presence of additive unit-variance white Gaussian noise, scaled to give a signal-to-noise ratio $\|\mathbf{f}\|/\sigma$ of 7. Figure 8 depicts the denoised signals produced by application of *WaveJS* to the noisy signals of Figure 7 for $N = 2048$ and $L = 5$. (Both figures are reproduced from [13].) Qualitatively, *WaveJS* has obviously caused an improvement from Figure 7 to Figure 8; unfortunately, the signals of Figure 8 are not what we typically think of as smooth or denoised, especially compared to the original noiseless signals of Figure 5. We defer quantitative analysis of the performance of *WaveJS* on these signals until the next section, in which we also present an alternative wavelet-shrinkage-based denoising technique used by Donoho and Johnstone.

4.3 Derivation of *SureShrink*

In the previous section we described Donoho's and Johnstone's results showing that *WaveJS* is nearly optimal among adaptive linear shrinkage estimators. This result hints at another area for investigation: how well could a nonlinear shrinkage estimator do? *SureShrink* is a nonlinear shrinkage estimator that operates in a manner quite similar to *WaveJS*, but achieves superior results.

SureShrink produces an estimate $\hat{\mathbf{f}}^{\mathrm{SH}}$ of the noiseless signal $\mathbf{f}$ in much the same way as *WaveJS* does: it transforms the problem to the wavelet domain and attempts to recover the wavelet coefficients that are significantly different from zero. Rather than do this in a linear fashion at each scale, however, *SureShrink* employs a nonlinear threshold function that essentially makes a "keep or kill" decision on each wavelet coefficient at each scale, based on the amplitude of that coefficient. If the coefficient is near zero, *SureShrink* assumes that it represents noise and estimates the denoised coefficient as zero; if the coefficient is significantly larger than zero, then its estimated denoised amplitude is changed only slightly. *SureShrink* uses the *soft threshold* function

$$\eta_t(y) = \mathrm{sgn}(y)(|y| - t)^+$$

for this estimation. This soft threshold function returns 0 for any argument closer than t to 0; it shrinks arguments greater than t or less than $-t$ toward zero by the constant amount t. *SureShrink* estimates wavelet coefficients as

$$\hat{w}_{j,k} = \begin{cases} x_{j,k}, & 0 \le j < L, \\ \eta_{t_j}(x_{j,k}), & L \le j \le M-1. \end{cases}$$

The concatenation of these estimates yields a vector estimate $\hat{\mathbf{w}}^{\mathrm{SH}}$ that is transformed to yield $\hat{\mathbf{f}}^{\mathrm{SH}} = W'\hat{\mathbf{w}}^{\mathrm{SH}}$. This is a complete description of *SureShrink* except for one vital detail: the choice of a threshold t_j at each scale $j \ge L$.

Donoho's and Johnstone's criterion for threshold choice at each scale is based on Stein's unbiased risk estimate (SURE) described in Section 2.4. After collecting the $\hat{w}_{j,k}$ into $\hat{\mathbf{w}}_j^{\mathrm{SH}}$, massaging into the form $\hat{\mathbf{w}}_j^{\mathrm{SH}} = \mathbf{x} + \mathbf{g}(\mathbf{x})$ required to calculate SURE, and verifying that the resulting $\mathbf{g}(\mathbf{x})$ meets the criteria required for applicability of SURE, (14) can be applied to yield an unbiased estimate of the risk of $\hat{\mathbf{w}}^{\mathrm{SH}}$ for any choice of threshold t_j:

$$\mathrm{SURE}_{t_j}(\mathbf{x}_j) = 2^j\sigma^2 - 2\sigma^2\#\{k : |x_{j,k}| \le t_j\} + \sum_{k=0}^{2^j-1}(\min(|x_{j,k}|, t_j))^2, \tag{41}$$

where the # operator simply returns the cardinality of the set to which it is applied. Clearly (41) involves a fundamental tradeoff between the second and third terms on its right-hand side: choosing a larger threshold t_j increases the sub-threshold cardinality benefit by increasing the number of $x_{j,k}$ that are below the threshold, but also increases the penalty for those coefficients that are above the threshold. Given an observation $\mathbf{x}$, $\mathrm{SURE}_{t_j}(\mathbf{x}_j)$ can be minimized over t_j by performing the calculation of (41) for t_j equal to each $x_{j,k}, k = 0, \ldots, 2^j - 1$. Donoho and Johnstone set

$$t_j = \arg \min_{0 \le t \le \sigma\sqrt{2j}} \mathrm{SURE}_t(\mathbf{x}_j). \tag{42}$$

The imposition of $\sigma\sqrt{2j}$ as the maximum allowable threshold is based on a detailed argument presented in [35].[17]

[17] The main thrust of this argument is that with high probability, the maximum of 2^j independent, identically distributed $N(0, \sigma^2)$ noise samples is less than $\sigma\sqrt{2j}$, so that a vector of 2^j zero coefficients corrupted by noise will be estimated as identically zero with probability approaching 1 as $j \to \infty$.

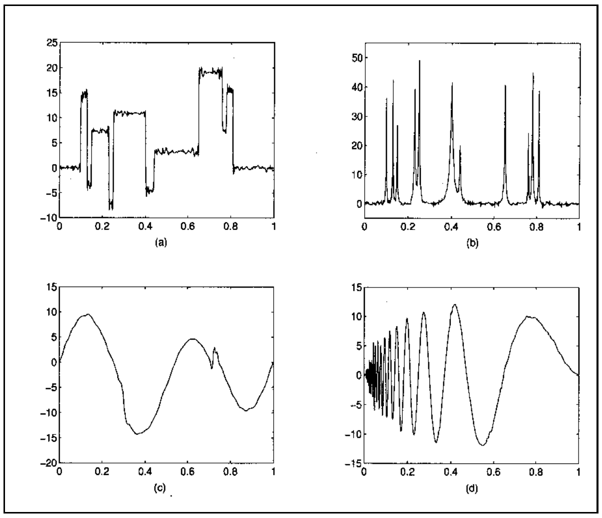

Figure 9: Performance of *SureShrink* on Noisy Signals of Figure 7 (duplicated from [13])

To summarize, *SureShrink* operates according to the following algorithm:

1. Take the discrete wavelet transform of the noisy data to yield $\mathbf{x} = W\mathbf{y}$.
2. For each scale $j \geq L$, find the threshold t_j according to (42).
3. Apply the soft threshold function to each wavelet coefficient at scale j to yield $\hat{w}_{j,k}^{\mathrm{SH}} = \eta_{t_j}(x_{j,k})$.
4. Take the inverse wavelet transform of the estimated wavelet coefficients to yield an estimate of the denoised function $\hat{\mathbf{f}}^{\mathrm{SH}} = W'\hat{\mathbf{w}}^{\mathrm{SH}}$.

Like *WaveJS*, *SureShrink* can be applied with any choice of wavelet basis. Donoho and Johnstone prove several near-optimality properties for *SureShrink* among nonlinear shrinkage estimators; these results essentially state that $\hat{\mathbf{f}}^{\mathrm{SH}}$ is nearly minimax across a wide class of functions and achieves asymptotic behavior of $J_N(\mathbf{f}; \hat{\mathbf{f}}^{\mathrm{SH}})$ that is close to optimal.[18]

To compare the performance of *SureShrink* with that of *WaveJS* Donoho and Johnstone apply *SureShrink* to the signals of Figure 5 that were examined in Section 4.2, again using $N = 2048$ and $L = 5$. The results of application of *SureShrink* to the noisy signals of Figure 7 are depicted in Figure 9; comparison to the results of *WaveJS* in Figure 8 confirms that, at least visually, the nonlinear shrinkage estimator *SureShrink* outperforms the linear shrinkage estimator *WaveJS*.

The relative performance of *WaveJS* and *SureShrink* for different values of N is quantified in Table 4, which tabulates the average per-element root mean-squared error of both techniques for multiple Monte Carlo runs on each of the four signals in Figure 5 for dyadic values of N from 128 to 16384.[19] For comparison, the per-element root mean-squared error of the noisy signals, or equivalently that of the MLE of $\mathbf{f}$ based on $\mathbf{y}$, is exactly 1. The results presented in Table 4 bolster the superior visual performance already observed for *SureShrink*. We see that for all tested numbers of samples, *SureShrink* achieves a lower risk than *WaveJS*; as the number of samples increases, both

[18]The results in [13] rely heavily on *Besov space* concepts. Besov spaces [36] are essentially collections of function classes parameterized by smoothness. An appropriate choice of parameters can give, for instance, a specific Hölder class, the L^2 Sobolev space, or the space of all functions of bounded total variation. The near-optimality and near-minimaxity results in [13] hold over Besov spaces for some range of parameter values.

[19]Donoho and Johnstone performed 20 runs for each of the $N \leq 4096$ cases, 10 runs for $N = 8192$, and one run at $N = 16384$ to obtain the values in Table 4.

N	signal (a)		signal (b)		signal (c)		signal (d)	
	WaveJS	*SureShr*	*WaveJS*	*SureShr*	*WaveJS*	*SureShr*	*WaveJS*	*SureShr*
128	.94	.89	.99	.94	.74	.73	.93	.82
256	.92	.80	.99	.85	.56	.56	.92	.74
512	.85	.78	.94	.74	.45	.44	.77	.63
1024	.77	.64	.84	.70	.34	.33	.59	.50
2048	.67	.56	.70	.52	.28	.27	.47	.38
4096	.58	.44	.54	.41	.23	.20	.38	.28
8192	.50	.37	.42	.32	.20	.17	.27	.19
16384	.43	.30	.33	.22	.16	.12	.21	.15

Table 4: Per-Element Root MSEs for *WaveJS* and *SureShrink* (from [13])

WaveJS and *SureShrink* produce an estimate with per-element risk decreasing toward zero, but the convergence of *SureShrink* toward zero appears to be faster than that of *WaveJS*.

The analysis of Donoho and Johnstone in [13] is a reminder that although the JSE dominates the MLE, it is not a panacea. In any application, some other approach might well prove better adapted to the problem at hand. In the case of denoising by wavelet shrinkage, the soft threshold provides "keep-or-kill" shrinkage that appears to be better suited to the problem than the adaptive linear shrinkage of the JSE.

In defense of the JSE and *WaveJS*, the scale-wise decomposition of $\hat{\mathbf{w}}^{\mathrm{JS}}$ robs the JSE of much of its potential benefit by greatly reducing the dimension of the estimation problem from N to some dyadic reduction of N at each scale. Certainly the motive for this decomposition cited by Donoho and Johnstone (the avoidance of excessive shrinkage at scales with significant signal energy) has intuitive appeal, but an equally compelling argument can be made against this scale-wise decomposition: why should we care about shrinking the signal if we thereby attain a smaller MSE? This is not addressed in [13]. It would be interesting to compare the results of *WaveJS* and*SureShrink* to those of an estimator that operates similarly to *WaveJS*, but shrinks all wavelet coefficients (possibly above some coarsest scale L) toward zero in ensemble rather than independently by scale. Such an approach would almost certainly result in more shrinkage of the signal along with the noise, giving a more biased denoising, but if our goal truly is a low-MSE reconstruction, this should not be a reason for concern.

5 Multiple Shrinkage

In the examples and applications we have examined so far, the use of the JSE has depended on the specification of a single point target toward which the observation vector is shrunk. We have seen that regardless of the location of the target in relation to the true mean, the JSE will dominate the MLE; however, we have also observed that meaningful reductions in MSE occur only if this shrinkage target is relatively close to the true mean vector being estimated. If the shrinkage target is a poor guess, the JSE will provide little improvement from the MLE.

There are many scenarios in which an accurate shrinkage target might be available. A reasonable target might be provided by the result of a previous experiment or by some special physical or statistical structure of the problem. In many cases, however, it is difficult to specify a single accurate point shrinkage target for use in the JSE. In this section we examine an expansion of the JSE framework that allows adaptive simultaneous shrinkage toward multiple point targets, a subspace, or multiple subspaces instead of a single point target. This modification unfetters the JSE from a single, all-or-nothing shrinkage target specification; it enables substantial reductions in risk compared to the MLE not just in the vicinity of a single point, but in the vicinity of multiple points or subspaces. The discussion in this section focuses mainly on the results of George in [14] and [17]. These papers detail much of the same work and topics and serve as complementary expositions of many of the same general ideas.

5.1 Shrinkage Toward a Subspace

In many settings, the requirement that the JSE shrink toward a point target is not accommodating to the inherent structure of the underlying problem. Consider, for instance, an observation vector comprising measurements that are believed to be independent and identically distributed. In this case, although there is a strong prior indication of what form a "reasonable" estimate of the mean $\boldsymbol{\theta} \in \mathbb{R}^p$ should take—namely, $\hat{\boldsymbol{\theta}} = \hat{\theta}\mathbf{1}_p$—the standard form of the JSE does not allow incorporation of this knowledge. More generally, if there is reason to believe that the mean belongs to a lower-dimensional subspace of $\mathbb{R}^p$, we might want to restrict any estimate to be close to this subspace. Again, however, the JSE as formulated thus far does not provide the ability to do this.

Consider again the general problem: we wish to estimate $\boldsymbol{\theta}$ from a vector observation $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$, where $\boldsymbol{\theta}$ and $\mathbf{x}$ both belong to $\mathbb{R}^p$. Suppose that instead of specifying a single point as a shrinkage target, we wish to specify an entire subspace $V \subset \mathbb{R}^p$ that captures some belief about the region of $\mathbb{R}^p$ in which a reasonable estimate might lie. Adapting the JSE, we would like to shrink toward the entire subspace V instead of toward a single point. The modification that will enable this is remarkably straightforward and intuitive. Let subspace V have dimension $p - q$, where q is required to be greater than 2. Define a projection operator P_V that projects from $\mathbb{R}^p$ into V and a projection operator $P_\perp = I - P_V$ that projects from $\mathbb{R}^p$ into the orthogonal complement of V in $\mathbb{R}^p$ (which has dimension q), so that

$$\mathbf{x} = P_V\mathbf{x} + P_\perp\mathbf{x} = \mathbf{x}_V + \mathbf{x}_\perp.$$

Now, because $\mathbf{x}_V \in V$, it corresponds to the component of the observation that conforms to the prior belief indicating V as the subspace in which a reasonable estimate should lie. On the other hand, $\mathbf{x}_\perp$ corresponds to the deviation of the observation from this prior belief. It might then be reasonable to estimate $\boldsymbol{\theta}$ by maintaining the component $\mathbf{x}_V$, but reducing the magnitude of the orthogonal component $\mathbf{x}_\perp$—in other words, shrinking $\mathbf{x}_\perp$ toward $\mathbf{0}$! Applying the JSE to this task, we obtain

$$\hat{\boldsymbol{\theta}}_V^{\mathrm{JS}} = \mathbf{x}_V + \left(1 - \frac{\sigma^2(q-2)}{\mathbf{x}_\perp'\mathbf{x}_\perp}\right)\mathbf{x}_\perp = P_V\mathbf{x} + \left(1 - \frac{\sigma^2(q-2)}{(\mathbf{x} - P_V\mathbf{x})'(\mathbf{x} - P_V\mathbf{x})}\right)(\mathbf{x} - P_V\mathbf{x}). \tag{43}$$

This is an adaptation of the JSE to the case where shrinkage is desired not toward a single point but toward an entire subspace. The standard point-target JSE can be viewed as a special case of (43), in which $V = v \in \mathbb{R}^p$, $q = p$, and $P_V\mathbf{x} = v$. The subspace-target JSE of (43) can be trivially modified to yield a dominating positive-part JSE as was the case with the point-target JSE. The general idea of modifying the JSE to shrink toward a subspace instead of a point appears to have been first proposed by Lindley [5] in 1962.

Shrinkage toward a subspace as prescribed by (43), instead of toward a point, enables greater flexibility in specification of prior information and the ability essentially to hedge bets about where exactly in $\mathbb{R}^p$ the parameter $\boldsymbol{\theta}$ lies. Of course, there is a price to this newfound flexibility: a reduction in the potential improvement over the MLE. We observed in Section 2 that the JSE's potential savings in MSE increase with the dimensionality of the problem; by shrinking toward a subspace instead of a single point, the dimension of the estimation problem is effectively reduced, resulting in a reduction of the potential improvement in MSE. Depending on the specific application, this may or may not be an acceptable price to pay.

5.2 Shrinkage Toward Multiple Targets

Suppose that vague, conflicting, or multiply indicating prior information suggests that any one of the subspaces $V_1, \ldots, V_K$ of $\mathbb{R}^p$ might be an appropriate shrinkage target for estimation of $\boldsymbol{\theta}$ from $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$. Denote by $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ the positive-part JSE formed by shrinking toward subspace V_k:

$$\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}} = P_k\mathbf{x} + \left(1 - \frac{\sigma^2(q_k - 2)}{(\mathbf{x} - P_k\mathbf{x})'(\mathbf{x} - P_k\mathbf{x})}\right)^+ (\mathbf{x} - P_k\mathbf{x}), \tag{44}$$

where P_k is the projection operator from $\mathbb{R}^p$ into V_k and where $p - q_k$ is the dimension of subspace V_k, where $q_k > 2$. The choice of which of these JSEs to use will have a great impact on the performance of the estimator.

It is natural to investigate whether the K estimators defined by (44) might be combined in some way in an attempt to reap the benefits of each, *i.e.*, to allow for substantial reductions in risk as long as $\boldsymbol{\theta}$ is close to *any* of the V_k. George studies this issue in [14], drawing a clever parallel to Bayesian estimation to derive a candidate multiple shrinkage estimator whose properties he then examines. Specifically, consider the problem of Bayesian estimation of $\boldsymbol{\theta}$ from $\mathbf{x} \sim N(\boldsymbol{\theta}, \sigma^2 I)$ under a prior $p_k(\boldsymbol{\theta})$. The Bayesian estimate $\hat{\boldsymbol{\theta}}^{\mathrm{BLS}} = E(\boldsymbol{\theta}|\mathbf{x})$ in this case can be shown to satisfy [8]

$$\hat{\boldsymbol{\theta}}^{\mathrm{BLS}} = \mathbf{x} + \sigma^2 \nabla \log m_k(\mathbf{x}), \tag{45}$$

where $m_k(\mathbf{x})$ is the marginal density for $\mathbf{x}$, *i.e.*,

$$m_k(\mathbf{x}) = \int \frac{1}{(2\pi\sigma^2)^{p/2}} \exp\left[-\frac{(\mathbf{x} - \boldsymbol{\theta})'(\mathbf{x} - \boldsymbol{\theta})}{2\sigma^2}\right] p_k(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \tag{46}$$

One way to interpret an arbitrary estimate $\hat{\boldsymbol{\theta}}$ that can be put in the form (45), then, is as a "pseudo-Bayesian" estimate of $\boldsymbol{\theta}$ generated under an assumed marginal density $m_k(\mathbf{x})$. George shows that each of the JSEs of (44) can be put in the form of (45) with

$$m_k(\mathbf{x}) = \begin{cases} \left(\frac{\sigma^2(q_k-2)}{(\mathbf{x}-P_k\mathbf{x})'(\mathbf{x}-P_k\mathbf{x})}\right)^{\frac{q_k-2}{2}} \exp\left[-\frac{q_k-2}{2}\right], & \|\mathbf{x} - P_k\mathbf{x}\|^2 \geq \sigma^2(q_k - 2), \\ \exp\left[-\frac{1}{2\sigma^2}(\mathbf{x} - P_k\mathbf{x})'(\mathbf{x} - P_k\mathbf{x})\right], & \|\mathbf{x} - P_k\mathbf{x}\|^2 < \sigma^2(q_k - 2), \end{cases} \tag{47}$$

for $k = 1, \ldots, K$. A few comments regarding these $m_k(\mathbf{x})$ are in order. First of all, these are *not* true marginal densities: examination of (47) reveals that these "marginals" do not, in general, integrate to 1. Secondly, George does not explicitly assume that the $m_k(\mathbf{x})$ actually arise from appropriately chosen priors by way of (46). The assumed marginals of (47) are essentially convenient constructions that are used to draw an analogy between James-Stein estimation and Bayesian estimation. This having been said, note that each assumed marginal has a property befitting its role as an encapsulation of some vague prior information suggesting shrinkage toward a subspace V_k: each $m_k(\mathbf{x})$ monotonically decreases as $\mathbf{x}$ becomes more distant from the corresponding V_k.

Returning to the problem of trying to combine the shrinkage estimators $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$, suppose that prior information is available regarding how likely each V_k is to be the "right" target for shrinkage. Specifically, suppose that it were possible to assign weights $w_1, \ldots, w_K$ satisfying $\sum_{k=1}^K w_k = 1$ to each of the shrinkage targets as a quantification of this information.[20] In the Bayesian analogy, this corresponds to specifying a mixture prior $p_*(\boldsymbol{\theta}) = \sum_{k=1}^K w_k p_k(\boldsymbol{\theta})$ that produces a marginal density

$$m_*(\mathbf{x}) = \sum_{k=1}^{K} w_k m_k(\mathbf{x}). \tag{48}$$

[20] Alternatively, we might simply assign $w_k = \frac{1}{K}$ for $k = 1, \ldots, K$ in the absence of any such information.

The Bayesian estimate of $\boldsymbol{\theta}$ under the prior $p_*(\boldsymbol{\theta})$ is $\hat{\boldsymbol{\theta}}_*^{\mathrm{BLS}} = \mathbf{x} + \sigma^2 \nabla \log m_*(\mathbf{x})$. This immediately suggests a scheme for generating a multiple shrinkage estimator from the ensemble of $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$. Using the assumed marginals $m_k(\mathbf{x})$ of (47) and weights w_k, define

$$\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}} = \mathbf{x} + \sigma^2 \nabla \log m_*(\mathbf{x}) \tag{49}$$

for $m_*(\mathbf{x})$ as given by (48). George shows that this $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ can be expressed as

$$\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}} = \sum_{k=1}^{K} \rho_k(\mathbf{x}) \hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}(\mathbf{x}),$$

where

$$\rho_k(\mathbf{x}) = \frac{w_k m_k(\mathbf{x})}{m_*(\mathbf{x})}. \tag{50}$$

Clearly (48) and (50) imply $\sum_{k=1}^{K} \rho_k(\mathbf{x}) = 1$. Thus $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is a convex combination of the individual shrinkage estimators $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}, k = 1, \ldots, K$, with the convex weights $\rho_k(\mathbf{x})$ adaptive functions of the data $\mathbf{x}$. This is George's multiple shrinkage estimator.

Investigation of the behavior of $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ reveals many appealing properties. First of all, when $\mathbf{x}$ is distant from all of the V_k, indicating that the prior knowledge that motivated the specification of these V_k was incorrect, the shrinkage provided by each of the $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ is very small and as a result $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is approximately the same as the MLE. Secondly, the $\rho_k(\mathbf{x})$ adapt to $\mathbf{x}$ in such a way as to provide the most shrinkage toward the closest V_k and little shrinkage toward distant V_k. The convex weights $\rho_k(\mathbf{x})$ vary proportionally with the corresponding $m_k(\mathbf{x})$; the $m_k(\mathbf{x})$ are largest when $\mathbf{x}$ is close to V_k. Thirdly, the prior weights w_k can be chosen to provide more shrinkage toward lower-dimensional subspaces to reflect the fact that proximity of $\mathbf{x}$ to a low-dimensional subspace V_k is a better validation of prior information than proximity of $\mathbf{x}$ to a high-dimensional subspace.[21] George proposes use of the following weights w_k in the absence of other prior information, in order to safeguard against this eventuality:

$$w_k = d^{(q_k-2)/2} \exp\left[\frac{q_k - 2}{2}\right], \tag{51}$$

where $d \geq 1$ is an arbitrary constant regulating how much more shrinkage should take place toward smaller-dimensional subspaces for equal distances. Using this scheme, equal-dimensional subspaces are all weighted similarly and the ρ_k reflect the shrinkage coefficients of the component $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ in most cases. Finally, regardless of the specific choice of the w_k, the multiple shrinkage estimator $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ behaves approximately like $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ in the vicinity of a single V_k.

What is the price of this adaptivity? A naive approach to achieving a good estimate would be to specify a plethora of subspaces V_k as shrinkage targets. This approach is shown to be flawed by a more careful analysis: the more subspace targets are specified, the more unwanted shrinkage toward incorrect subspaces results. In the immediate neighborhood of one of the V_k, the estimator $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ might still behave approximately like $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$, offering substantial reductions in risk; however, the neighborhood in which this is true becomes smaller and smaller as more V_k are introduced. By hedging our bets we decreasing robustness and diminishing our maximum returns.

[21] That is, a point chosen at random is less likely to be near a lower-dimensional subspace than a higher-dimensional one.

To make the concepts of the previous paragraph more precise, let us examine the expected reduction in risk from the MLE afforded by the multiple-shrinkage JSE across parameter space. Stein's unbiased risk estimate (SURE) described in Section 2.4 is particularly suited to this task. George's application of SURE to the risk of each component $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ yields the unbiased risk estimate

$$\mathrm{SURE}_k(\mathbf{x}) = p\sigma^2 - D_k(\mathbf{x}),$$

where

$$D_k(\mathbf{x}) = \begin{cases} \frac{\sigma^4(q_k-2)^2}{(\mathbf{x}-P_k\mathbf{x})'(\mathbf{x}-P_k\mathbf{x})}, & \|\mathbf{x}-P_k\mathbf{x}\|^2 \geq \sigma^2(q_k-2), \\ 2\sigma^2 q_k - (\mathbf{x}-P_k\mathbf{x})'(\mathbf{x}-P_k\mathbf{x})), & \|\mathbf{x}-P_k\mathbf{x}\|^2 < \sigma^2(q_k-2). \end{cases}$$

Application of SURE to $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ yields

$$\mathrm{SURE}_*(\mathbf{x}) = p\sigma^2 - D_*(\mathbf{x}), \tag{52}$$

where

$$D_*(\mathbf{x}) = \sum_{k=1}^{K} \rho_k(\mathbf{x}) D_k(\mathbf{x}) - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{K} \rho_k(\mathbf{x})\rho_l(\mathbf{x}) \left\| \hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}(\mathbf{x}) - \hat{\boldsymbol{\theta}}_l^{\mathrm{JS}}(\mathbf{x}) \right\|^2. \tag{53}$$

This last equation implies that the expected reduction in risk achieved by $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is roughly a convex combination of the reductions achieved with each individual $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$, with the important caveat that the reduction is offset by an increase depending on how widely spaced the ensemble of $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ are. If $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is influenced strongly by only one of its components $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$, then the remaining component shrinkage estimators will roughly equal the MLE, be relatively closely spaced and $D_*(\mathbf{x})$ will approximately equal $D_k(\mathbf{x})$. On the other hand, if $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is influenced by two or more components, each shrinking toward a different subspace and thus producing widely spaced component estimates, (53) indicates that there will be a price to pay terms of the expected reduction in risk. A cavalier proliferation of shrinkage targets embodied in a single $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ results in an increase in the probability that $\mathbf{x}$ will be pulled in multiple directions by multiple targets and thus confounds the desirable property of the JSE—namely, a significant reduction in risk near the shrinkage target. Additionally, regardless of the number of shrinkage targets, there is a price to pay for adaptivity even when $\boldsymbol{\theta} \in V_k$ for some k. Because the adaptive specification ensures that $\rho_k < 1$ except in degenerate cases (*e.g.*, when $w_k = 1$ or $\mathbf{x} \in V_k$), there will always be pressure from the other components that will draw the estimate out of V_k, and thus $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ will not perform quite as well as $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ even in the ideal situation where shrinkage target V_k is exactly right.

The previous paragraphs provide a complete description of the multiple shrinkage estimator presented by George in [14]; a straightforward extension in [17] expands the range of use of the estimator. Suppose that the details of an application suggest that mean vector $\boldsymbol{\theta}$ contains two distinct subvectors $\boldsymbol{\theta}_1 \in \mathbb{R}^{p_1}$ and $\boldsymbol{\theta}_2 \in \mathbb{R}^{p_2}$ (where $p_1 + p_2 = p$) that are unrelated. Suppose further that prior information suggests that $V_1 \subset \mathbb{R}^{p_1}$ and $V_2 \subset \mathbb{R}^{p_2}$ might be good shrinkage targets for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. It might then seem reasonable to implement a JSE that shrinks toward $V_1 \times V_2$. Another reasonable estimator, however, could be formed by concatenation of separate JSEs applied to each subvector, as $\hat{\boldsymbol{\theta}}^{\mathrm{JS}} = [\hat{\boldsymbol{\theta}}_1^{\mathrm{JS}}, \hat{\boldsymbol{\theta}}_2^{\mathrm{JS}}]$. In [17], George describes how such a concatenation JSE might be used as a component of a multiple shrinkage estimator, *i.e.*, as one of the $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ from which $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ is formed. The only modification required is the use of the product of the assumed marginals of the subvector estimators for the assumed marginal of the component

individual estimator	subspace shrinkage target(s)
$\hat{\boldsymbol{\theta}}_1^{\mathrm{JS}}$	$\mathbf{1}_{26}$
$\hat{\boldsymbol{\theta}}_2^{\mathrm{JS}}$	$\mathbf{1}_{14} \times \mathbf{1}_{12}$
$\hat{\boldsymbol{\theta}}_3^{\mathrm{JS}} = \boxed{\hat{\boldsymbol{\theta}}_{3_1}^{\mathrm{JS}}, \hat{\boldsymbol{\theta}}_{3_2}^{\mathrm{JS}}}$	$\mathbf{1}_{14}$ and $\mathbf{1}_{12}$
$\hat{\boldsymbol{\theta}}_4^{\mathrm{JS}}$	$\mathbf{1}_7 \times \mathbf{1}_7 \times \mathbf{1}_6 \times \mathbf{1}_6$
$\hat{\boldsymbol{\theta}}_5^{\mathrm{JS}} = \boxed{\hat{\boldsymbol{\theta}}_{5_1}^{\mathrm{JS}}, \hat{\boldsymbol{\theta}}_{5_2}^{\mathrm{JS}}, \hat{\boldsymbol{\theta}}_{5_3}^{\mathrm{JS}}, \hat{\boldsymbol{\theta}}_{5_4}^{\mathrm{JS}}}$	$\mathbf{1}_7, \mathbf{1}_7, \mathbf{1}_6$, and $\mathbf{1}_6$

Table 5: Multiple Shrinkage Component Estimators

estimator. In other words, if $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}} = [\hat{\boldsymbol{\theta}}_{k_1}^{\mathrm{JS}}, \ldots, \hat{\boldsymbol{\theta}}_{k_J}^{\mathrm{JS}}]$, with $\hat{\boldsymbol{\theta}}_{k_j}^{\mathrm{JS}} = \mathbf{x} + \nabla\sigma^2 \log m_{k_j}(\mathbf{x})$ as in (45), then we simply put $m_k(\mathbf{x}) = \prod_{j=1}^{J} m_{k_j}(\mathbf{x})$, which then allows calculation of ρ_k for combination with other standard or subvector-concatenation JSEs into a single multiple shrinkage estimator. The composition of an assumed marginal $m_k(\mathbf{x})$ from the product of the assumed subvector marginals $m_{k_j}(\mathbf{x})$ suggests that each subvector is independent, which presumably is the motivation for using separate subvector JSEs to form $\hat{\boldsymbol{\theta}}_k^{\mathrm{JS}}$ in the first place.

5.3 Multiple Shrinkage Estimation Example

George presents an example of a multiple shrinkage estimator in [17] with a flavor similar to that of the example of Efron and Morris [26] that was described in Section 2.5. George's example concerns the prediction of team batting averages for all 26 major league baseball teams during the 1984 season. Specifically, George uses the observed batting averages from each team's first 300 at-bats as measurements from which each team's average for the remainder of the season are to be estimated. (Generally speaking, a team will accrue more than 5000 at-bats over the course of the entire season.) George makes the same assumptions as Efron and Morris, and processes the batting average data in the same way (*i.e.*, using the variance-stabilizing transformation of [26] described in Section 2.5) in order to make it adhere more closely to a multivariate normal distribution with known covariance, thus fostering fruitful application of the JSE.

The division of baseball teams into leagues and divisions and the perception of disparities in the quality of play in these divisions in the mid-1980s provide a structure suited to illustration of a multiple shrinkage estimator based on shrinkages to different subspaces. Throughout the 1980s, the American League (AL) comprised 14 teams and the National League (NL) comprised 12 teams; each league further separated teams into Eastern and Western divisions, each of which comprised half of the teams in the respective league. In the mid-1980s, many observers—impartial or otherwise—believed that the quality of play was stronger in the American League than the National League, and particularly that the American League East was superior to the American League West.[22]

These divisions and hypotheses of superiority suggest many plausible subspace shrinkage targets for a JSE. Historical data suggest that differences in batting averages between leagues or divisions tend to be statistically insignificant in the long run [37]; this might suggest shrinkage to the subspace $\mathbf{1}_{26}$. Under the hypothesis that the quality of play is roughly homogeneous within each division of each league, but possibly varies between divisions, a reasonable subspace target might be $\mathbf{1}_7 \times \mathbf{1}_7 \times \mathbf{1}_6 \times \mathbf{1}_6$; alternatively, this hypothesis might suggest that a JSE should be formed by concatenating individual JSEs for each division. One significant and fundamental asymmetry between the American and National Leagues is the designated hitter: in the American League, the pitcher does not bat but has his place in the lineup taken by another player. Because pitchers are

[22]These biases were borne out somewhat by the 1984 World Series: the Detroit Tigers, a team from the American League East, won the championship that year.

team	division	BA: 300 at-bats $x_i = \hat{\theta}_i^{\mathrm{ML}}$	BA: rest of season θ_i	$y_i = \psi_i^{\mathrm{ML}}$	ψ_i
Boston	AL,E	.239	.284	-9.53	-7.72
New York	AL,E	.244	.277	-9.29	-8.01
Toronto	AL,E	.283	.273	-7.78	-8.17
Detroit	AL,E	.278	.271	-7.98	-8.25
Cleveland	AL,E	.256	.265	-8.83	-8.46
Milwaukee	AL,E	.272	.262	-8.19	-8.60
Baltimore	AL,E	.247	.253	-9.20	-8.96
Kansas City	AL,W	.279	.268	-7.93	-8.37
Minnesota	AL,W	.280	.264	-7.88	-8.53
Texas	AL,W	.238	.262	-9.57	-8.61
Oakland	AL,W	.244	.260	-9.33	-8.69
Seattle	AL,W	.271	.256	-8.22	-8.82
California	AL,W	.228	.250	-9.95	-9.07
Chicago	AL,W	.222	.249	-10.22	-9.11
Philadelphia	NL,E	.287	.265	-7.63	-8.46
Chicago	NL,E	.284	.259	-7.73	-8.72
New York	NL,E	.248	.258	-9.15	-8.75
Pittsburgh	NL,E	.238	.255	-9.54	-8.86
St. Louis	NL,E	.285	.250	-7.68	-9.08
Montreal	NL,E	.294	.248	-7.37	-9.14
San Francisco	NL,W	.260	.265	-8.67	-8.48
Houston	NL,W	.229	.265	-9.91	-8.47
San Diego	NL,W	.284	.258	-7.74	-8.77
Atlanta	NL,W	.194	.250	-11.41	-9.07
Los Angeles	NL,W	.244	.244	-9.31	-9.32
Cincinnati	NL,W	.265	.242	-8.49	-9.38

Table 6: 1984 Season Team Batting Averages and Transformed Values

	$\hat{\boldsymbol{\theta}}^{\mathrm{ML}}$	$\hat{\boldsymbol{\theta}}_1^{\mathrm{JS}}$	$\hat{\boldsymbol{\theta}}_2^{\mathrm{JS}}$	$\hat{\boldsymbol{\theta}}_3^{\mathrm{JS}}$	$\hat{\boldsymbol{\theta}}_4^{\mathrm{JS}}$	$\hat{\boldsymbol{\theta}}_5^{\mathrm{JS}}$	$\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}, d=1$	$\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}, d=2$	$\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}, d=5$
$\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\|^2$	26.57	4.24	4.94	7.76	7.27	10.66	5.37	4.73	4.43

Table 7: Squared Errors of MLE, component JSEs, and Multiple-Shrinkage JSE

notoriously bad batters (with relatively few exceptions), the implementation of the designated hitter rule in one league only suggests a subspace shrinkage target of $\mathbf{1}_{14} \times \mathbf{1}_{12}$, or the implementation of a concatenated JSE formed from two JSEs, one for each league.

George forms a multiple shrinkage estimator that incorporates all of these hypotheses. Specifically, George's multiple shrinkage estimator embodies five individual estimators, three of which are standard JSEs that shrink toward a single subspace and two of which are concatenations of subvector JSEs. These individual estimators, which correspond to the hypotheses forwarded in the previous paragraph, are summarized in Table 5.

The observed batting averages for each team through its first 300 at-bats is presented in Table 6. The performance of each estimator as observed by George is presented in Table 7, which presents the actual losses observed for the MLE, each of the five JSEs of Table 5, and the multiple shrinkage estimator $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ for three values of d in (51). (Recall that a larger value of d results in a heavier weight placed on those component JSEs that shrink toward lower-dimensional subspaces.) Note that these losses are in terms of the transformed, roughly identity-covariance variables, not the explicit batting averages.[23] We see from Table 7 that for each of the values of d tested, $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ outperforms at least three of the five individual JSEs; for two out of three values of d tested, $\hat{\boldsymbol{\theta}}_*^{\mathrm{JS}}$ outperforms all but the best individual JSE. (As expected, all of these estimators perform markedly better than the MLE.) This example suggests the general utility of a multiple shrinkage estimator: it makes the selection of a single shrinkage target from a handful of reasonable alternatives unnecessary and thus averts much of the peril of shrinkage target misspecification inherent in such a choice.

6 Conclusion

We have presented a concise introduction to the James-Stein estimator (JSE). The JSE is applicable to any problem that can be posed as estimation of a multivariate normal mean; under certain remarkably general conditions it achieves a lower mean-squared error than the maximum likelihood estimator (MLE). The JSE can be viewed as an attempt to emulate the Bayes least-squares estimator in the absence of a prior but given a guess of the prior mean. We have described the properties and behavior of the JSE and examined its application to three specific problems [12, 13, 14]. Two of these concern the adaptation of the JSE to topics of significant general interest to the signal processing community—state estimation for a dynamic system [12] and function denoising [13]. The third is not a specific application so much as an interesting extension of the James-Stein framework [14].

Despite its dominance of the MLE, the JSE has not been widely accepted outside of the statistics community. A search for references to James-Stein estimation in a database of papers published in IEEE journals since 1980 turns up fewer than half a dozen records. The JSE has had a storied history, generating an uproar upon its introduction in 1961, being deemed a "paradox" by many statisticians [4], and eventually being recognized as a useful tool by researchers in statistics and econometrics. It is our hope that the JSE will come to be known and used by engineers in all fields and that the JSE's tumultuous past will give way to a distinguished future.

References

[1] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Annals of Mathematical Statistics*, 10:299–326, 1939.

[2] W. James and C. Stein. Estimation with quadratic loss. In *Proceedings fo the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379, Berkeley, 1961. University of California Press.

[3] C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings fo the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, Berkeley, 1956. University of California Press.

[4] B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5):119–127, 1977.

[5] D. V. Lindley. Discussion on Professor Stein's paper. *Journal of the Royal Statistical Society, Series B*, 24:285–287, 1962.

[6] C. M. Stein. Confidence sets for the mean of a multivariate normal distribution. *Journal of the Royal Statistical Society, Series B*, 24:265–296, 1962.

[23] George does not provide data in [17] that allow $\boldsymbol{\theta}$-space losses for each team to be tabulated as in Table 1.

[7] B. M. Hill. On coherence, inadmissibility and inference about many parameters in the theory of least squares. In S. E. Fienberg and A. Zellner, editors, *Studies in Bayesian Econometrics and Statistics*, pages 555–584. North-Holland, Amsterdam, 1974.

[8] E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer, New York, 2nd edition, 1998.

[9] M. H. J. Gruber. *Improving Efficiency by Shrinkage: the James-Stein and Ridge Regression Estimators.* Marcel Dekker, New York, 1998.

[10] G. G. Judge and M. E. Bock. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics.* North-Holland, Amsterdam, 1978.

[11] E. Greenberg and Jr. C. E. Webster. *Advanced Econometrics: A Bridge to the Literature.* John Wiley & Sons, New York, 1998.

[12] J. H. Manton, V. Krishnamurthy, and H. V. Poor. James-Stein state filtering algorithms. *IEEE Transactions on Signal Processing*, 46(9):2431–2447, September 1998.

[13] D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, December 1995.

[14] E. I. George. Minimax multiple shrinkage estimation. *The Annals of Statistics*, 14(1):188–205, January 1986.

[15] H. L. Van Trees. *Detection, Estimation, and Modulation Theory Part I: Detection, Estimation, and Linear Modulation Theory.* John Wiley & Sons, 1968.

[16] B. Efron and C. N. Morris. Limiting the risk of Bayes and empirical Bayes estimators—part II: The empirical Bayes case. *Journal of the American Statistical Association*, 67:130–139, 1972.

[17] E. I. George. Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, 81(394):437–445, June 1986.

[18] M. E. Bock. Minimax estimators of the mean of a multivariate normal distribution. *The Annals of Statistics*, 3(1):209–218, January 1975.

[19] W. E. Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, 42:385–388, 1971.

[20] P. Y.-S. Shao and W. E. Strawderman. Improving on the James-Stein positive-part estimator. *The Annals of Statistics*, 22:1517–1538, 1994.

[21] Y. Y. Guo and N. Pal. A sequence of improvements over the James-Stein estimator. *Journal of Multivariate Analysis*, 42(2):302–317, 1992.

[22] B. Efron and C. Morris. Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68:117–130, 1973.

[23] A. J. Baranchik. Multiple regression and estimation of the mean of a multivariate normal distribution. Technical Report 51, Department of Statistics, Stanford University, Stanford, CA, 1964.

[24] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

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

[26] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[27] Y. Theodor and U. Shaked. Robust discrete-time minimum-variance filtering. *IEEE Transactions on Signal Processing*, 44(2):181–189, February 1996.

[28] L. Xie, Y. C. Soe, and C. E. de Souza. Robust Kalman filtering for uncertain discrete-time systems. *IEEE Transactions on Automatic Control*, 39:1310–1314, 1994.

[29] H. W. Sorenson and D. L. Alspach. Recursive Bayesian estimation using Gaussian sums. *Automatica*, 7:465–479, 1971.

[30] A. M. Makowski, W. S. Levine, and M. Asher. The nonlinear MMSE filter for partially observed systems driven by non-Gaussian white noise, with application to failure estimation. In *Proceedings of the 23rd IEEE Conference on Decision and Control*, December 1984.

[31] S. M. Tonissen and A. Logothetis. Estimation of multiple target trajectories with time varying amplitudes. In *Proceedings of the 8th IEEE Signal Processing Workshop on Statistical Signal and Array Processing*, pages 32–35, June 1996.

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

[33] S. Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, San Diego, 1998.

[34] I. Daubechies. *Ten Lectures on Wavelets.* Number 61 in CBMS-NSF Series in Applied Mathematics. SIAM, Philadelphia, 1992.

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

[36] H. Triebel. *Theory of Function Spaces.* Birkhäuser Verlag, Bassel, 1983.

[37] *The Baseball Encyclopedia: The Complete and Definitive Record of Major League Baseball.* Macmillan, New York, 9th edition, 1993.