CCIT Report #644 November 2007

Generalized SURE for Exponential Families: Applications to Regularization

Yonina C. Eldar

Abstract

Stein's unbiased risk estimate (SURE) was proposed by Stein for the independent, identically distributed (iid) Gaussian model in order to derive estimates that dominate least-squares (LS). In recent years, the SURE criterion has been employed in a variety of denoising problems for choosing regularization parameters that minimize an estimate of the mean-squared error (MSE). However, its use has been limited to the iid case which precludes many important applications. In this paper we begin by deriving a SURE counterpart for general, not necessarily iid distributions from the exponential family. This enables extending the SURE design technique to a much broader class of problems. Based on this generalization we suggest a new method for choosing regularization parameters in penalized LS estimators. We then demonstrate its superior performance over the conventional generalized cross validation approach in the context of image deblurring. The SURE technique can also be used to design estimates without predefining their structure. However, allowing for too many free parameters impairs the performance of the resulting estimates. To address this inherent tradeoff we propose a regularized SURE objective. Based on this design criterion, we derive a wavelet denoising strategy that is similar in sprit to the standard soft-threshold approach but can lead to improved MSE performance.

Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel. Phone: +972-4-8293256, fax: +972-4-8295757, E-mail: yonina@ee.technion.ac.il. This work was supported in part by the Israel Science Foundation.

I. Introduction

Estimation in multivariate problems is a fundamental topic in statistical signal processing. One of the most common recovery strategies for deterministic unknown parameters is the well-known maximum likelihood (ML) method. The ML estimator enjoys several appealing properties, including asymptotic efficiency under suitable regularity conditions. Nonetheless, its mean-squared error (MSE) can be improved upon in the non-asymptotic regime in many different settings.

In their seminal work, Stein and James showed that for the independent, identically-distributed (iid) linear Gaussian model, it is possible to construct a nonlinear estimator with lower MSE than that of ML for all values of the unknowns [1], [2]. Various modifications of the James-Stein method have since been developed that are applicable to the non-iid Gaussian case as well [3], [4], [5], [6], [7].

The James-Stein approach is based on the Stein unbiased risk estimate (SURE) [8], [9], which is an unbiased estimate of the MSE. Since the MSE in general depends on the true unknown parameter values it cannot be used as a design objective. However, using the SURE principle leads to a relatively simple technique for determining methods that have lower MSE than ML. The idea is to choose a class of estimates, and then select the member from the class that minimizes the SURE estimate of the MSE. This strategy has been applied to a variety of different denoising techniques [10], [11], [12], [13]. Typically, in these problems, implicit prior information on the signal to be recovered is incorporated into the chosen structure of the estimate. For example, in wavelet denoising the signal is assumed to be sparse in the wavelet domain which motives the use of threhoslding. Only the value of the threshold is determined by the SURE principle.

The SURE method is appealing as it allows to directly approximate the MSE of an estimate from the data, without requiring knowledge of the true parameter values. However, it has two main drawbacks which severely limit its use in practical applications. The first restriction is that it was originally limited to the iid Gaussian case. Several extensions have been developed for different independent models. In particular, a SURE principle for iid, infinitely divisible random variables with finite variance is derived in [14]. Extensions to independent variables from a continuous exponential family are treated in [15], [16], while the discrete exponential case is discussed in [17]. All of these generalizations are confined to the independent case which precludes a variety of important applications such as image deblurring.

The second drawback of using SURE as a design criterion is that in order to get meaningful estimators

the basic structure of the estimate must be determined in advance. If no parametrization is assumed, then there are too many free variables to be optimized, and the SURE method will typically not lead to good MSE behavior.

In this paper we extend the basic SURE principle in two directions, in order to circumvent the two fundamental drawbacks outlined above. First, in Section III, we generalize SURE to multivariate, possibly non-iid exponential families. In particular, we develop an unbiased estimate of the MSE for a general Gaussian vector model. Exponential probability density functions (pdfs) play an important role in statistics due to the Pitman-Koopman-Darmois theorem [18], [19], [20], which states that among distributions whose domain does not vary with the parameter being estimated, only in exponential families is there a sufficient statistic with bounded dimension as the sample size increases [21]. Furthermore, efficient estimators exist only when the underlying model is exponential. Many known distributions are of the exponential form, such as Gaussian, gamma, chi-square, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, and geometric distributions. Our result has important practical value as it extends the applicability of the SURE technique to more general estimation models, and in particular to scenarios in which the observations are dependent. This is the case, for example, when using overcomplete wavelet transforms, and in image deblurring.

An immediate application of this extension is to the general linear Gaussian model. In this setup, we seek to estimate a parameter vector $\boldsymbol{\theta}$ from noisy, blurred measurements $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{w}$ is a Gaussian noise vector. One of the most popular recovery strategies in this context is the regularized least-squares method. In this approach, the estimate is designed to minimize a regularized least-squares objective where a typical choice of penalization is the ℓ_2 norm. This technique is commonly referred to as Tikhonov regularization [22]. An important aspect of the Tikhonov technique, which significantly impacts its performance, is selecting the regularization parameter. A variety of different methods have been proposed for regularization selection [23], [24], [25], [26], [27], [28]. One of the most popular techniques is generalized cross-validation (GCV) [29]. Here, we suggest an alternative strategy based on our extended SURE criterion. Specifically, we use SURE to evaluate the MSE of the Tikhonov approach for any choice of regularization, and then select the value that minimizes the SURE estimate. This allows SURE-based optimization of a broad class of deblurring methods. Using several test images, we demonstrate that this strategy can lead to significant performance improvement over the standard GCV criterion in the context of image deblurring.

Finally, to circumvent the need for pre-defining a particular structure when applying SURE, in Section VI we propose an alternative approach based on regularizing the SURE objective. Specifically, we suggest adding a regularization term to the SURE expression and choosing an estimate that minimizes the regularized function. In this way, we can control the properties of the estimate without having to aprioiri assume a specific structure. We then illustrate this strategy in the context of wavelet denoising. Instead of assuming a threshold estimate and choosing the threshold to minimize the SURE criterion, as in [10], we design an estimate that minimizes an ℓ_1 regularized SURE objective. The resulting denoising scheme has the form of a threshold with a particular form of shrinkage, that is different than that obtained when using soft or hard thresholding. To evaluate our method, we compare it with SureShrink of Donoho and Johnstone, by repeating the simulations reported in their paper. As we show, the recovery results tend to be better using our technique. Moreover, our approach is general as it is not tailored to a specific problem. We thus believe that using a regularized SURE principle together with the generalized SURE developed here can extend the applicability of SURE-based estimators to a broad class of problems.

The remaining of the paper is organized as follows. In Section II we introduce the basic concept of MSE estimation. An extension of SURE to exponential families is developed in Section III. We then specialize the results to the linear Gaussian model in Section IV. Applications to regularization selection are discussed in Section V. The regularized SURE criterion, together with an application to wavelet denoising, are developed in Section VI.

II. MSE Estimation

We denote vectors by boldface lowercase letters, *e.g.,* $\mathbf{x}$, and matrices by boldface uppercase letters *e.g.,* $\mathbf{A}$. The ith component of a vector $\mathbf{y}$ is written as y_i, and $\hat{(\cdot)}$ is an estimated vector. The identity matrix is written as $\mathbf{I}$, $\mathbf{A}^T$ is the transpose of $\mathbf{A}$, and $\mathbf{A}^\dagger$ denotes the pseudo-inverse.

We consider the class of problems in which our goal is to estimate a deterministic parameter vector $\boldsymbol{\theta}$ from observations $\mathbf{x}$ which are related through a pdf $f(\mathbf{x};\boldsymbol{\theta})$. We further assume that the pdf belongs to the exponential family of distributions and can be expressed in the form

$$f(\mathbf{x};\boldsymbol{\theta}) = r(\mathbf{x})\exp\{\boldsymbol{\theta}^T\phi(\mathbf{x}) - g(\boldsymbol{\theta})\}, \tag{1}$$

where $r(\mathbf{x})$ and $\phi(\mathbf{x})$ are functions of the data only, and $g(\boldsymbol{\theta})$ depends on the unknown parameter $\boldsymbol{\theta}$.

As an example of an application where the model (1) can occur, consider the location problem of estimating a parameter vector $\boldsymbol{\theta} \in \mathcal{R}^m$ from observations $\mathbf{x} \in \mathcal{R}^n$ related through the linear model:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}, \tag{2}$$

where $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance $\mathbf{C} \succ 0$. The pdf of $\mathbf{x}$ is then given by (1) with

$$\begin{aligned} r(\mathbf{x}) &= \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C})}} \exp\{-(1/2)\mathbf{x}^T\mathbf{C}^{-1}\mathbf{x}\}; \\ \phi(\mathbf{x}) &= \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}; \\ g(\boldsymbol{\theta}) &= (1/2)\boldsymbol{\theta}^T\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\boldsymbol{\theta}. \end{aligned} \tag{3}$$

Other examples of distributions in the exponential family include Poisson with unknown mean, exponential with unknown mean, gamma, and Bernoulli or binomial with unknown success probabilities.

Given the model (1), a sufficient statistic for estimating $\boldsymbol{\theta}$ is given by

$$\mathbf{u} = \phi(\mathbf{x}). \tag{4}$$

Therefore, any reasonable estimate of $\mathbf{x}$ will be a function of $\mathbf{u}$. More specifically, from the Rao-Blackwell theorem [30] it follows that if $\hat{\boldsymbol{\theta}}$ is an estimate of $\boldsymbol{\theta}$ which is not only a function only $\mathbf{u}$, then the estimate $E\{\hat{\boldsymbol{\theta}}|\mathbf{u}\}$ has lesser or equal MSE than that of $\hat{\boldsymbol{\theta}}$, for all $\boldsymbol{\theta}$. Therefore, in the sequel, we only consider methods that depend on the data via $\mathbf{u}$.

Let $\hat{\boldsymbol{\theta}}_0$ be a particular estimate of $\boldsymbol{\theta}$, and suppose we would like to improve its MSE, where the MSE of an estimate $\hat{\boldsymbol{\theta}}$ is defined by $E\{\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\}$. Thus, our goal is to design a method with lower MSE for all values of $\boldsymbol{\theta}$. To this end we may consider estimators of the form

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 + h(\mathbf{u}), \tag{5}$$

for some function $h(\mathbf{u})$, and then choose $h(\mathbf{u})$ to minimize the MSE. Denoting by $\epsilon(\boldsymbol{\theta})$ the MSE of $\hat{\boldsymbol{\theta}}_0$, we can

express the MSE of $\hat{\boldsymbol{\theta}}$ as

$$\begin{aligned} E\left\{\|\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\right\} = & \\ = & E\left\{\|\hat{\boldsymbol{\theta}}_0-\boldsymbol{\theta}+h(\mathbf{u})\|^2\right\} \\ = & \epsilon(\boldsymbol{\theta})+E\left\{\|h(\mathbf{u})\|^2\right\}-2E\left\{h^T(\mathbf{u})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)\right\}. \end{aligned} \tag{6}$$

In order to minimize the MSE over $h(\mathbf{u})$ we need to explicitly evaluate the expression

$$f(h,\boldsymbol{\theta}) = E\left\{\|h(\mathbf{u})\|^2\right\}-2E\left\{h^T(\mathbf{u})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_0)\right\}. \tag{7}$$

Evidently, the MSE will depend in general on $\boldsymbol{\theta}$, which is unknown, and therefore cannot be minimized. Instead, we may seek an unbiased estimate of $f(h,\boldsymbol{\theta})$ and then choose h to minimize this estimate. Specifically, suppose we construct a function $g(h(\mathbf{u}))$ that depends only on $\mathbf{u}$ (and not on $\boldsymbol{\theta}$), such that

$$E\left\{g(h(\mathbf{u}))\right\} = E\left\{h^T(\mathbf{u})\boldsymbol{\theta}\right\}. \tag{8}$$

Then

$$\hat{f}(h) = \|h(\mathbf{u})\|^2-2(g(h(\mathbf{u}))-h^T(\mathbf{u})\hat{\boldsymbol{\theta}}_0), \tag{9}$$

is an unbiased estimate of $f(h,\boldsymbol{\theta})$, since clearly $E\{\hat{f}(h)\} = f(h,\boldsymbol{\theta})$. A reasonable strategy therefore is to select $h(\mathbf{u})$ to minimize our assessment $\hat{f}(h)$ of the MSE. This approach was first proposed by Stein in [8], [9] for the iid Gaussian model (2) with $\mathbf{C}=\mathbf{I}$ and $\mathbf{H}=\mathbf{I}$.

To apply the design technique outlined above we need to construct an unbiased estimate $g(h(\mathbf{u}))$ of $E\{h^T(\mathbf{u})\boldsymbol{\theta}\}$. In the next section we derive such a function for the general exponential model (1). In practice, it is typically assumed that $h(\mathbf{u})$ has a particular structure, so that it is parameterized by some vector α. The value of α is then chosen to minimize the SURE estimate of the MSE. In Sections IV and V we apply this technique to several examples and propose an alternative to the popular GCV method for Tikhonov regularization. In Section VI, we suggest a regularized SURE strategy for determining $h(\mathbf{u})$ without the need for parametrization, and demonstrate its performance in the context of wavelet denoising.

III. Extended SURE Principle

The following theorem provides an unbiased estimate of $E\left\{h^T(\mathbf{u})\boldsymbol{\theta}\right\}$ which depends only on $\mathbf{u}$ and not on the unknown parameters $\boldsymbol{\theta}$.

Theorem 1: Let $\mathbf{x}$ denote a random vector with exponential pdf given by (1), and let $\mathbf{u} = \phi(\mathbf{x})$ be a sufficient statistic for estimating $\boldsymbol{\theta}$ from $\mathbf{x}$. Let $h(\mathbf{u})$ be an arbitrary function of $\boldsymbol{\theta}$ that is weakly differentiable in $\mathbf{u}$ and such that $E\left\{|h_i(\mathbf{u})|\right\}$ is bounded where $h_i(\mathbf{u})$ is the ith component of $h(\mathbf{u})$. Then

$$E\left\{h^T(\mathbf{u})\boldsymbol{\theta}\right\} = -E\left\{\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right)\right\} - E\left\{h^T(\mathbf{u})\frac{d\ln q(\mathbf{u})}{d\mathbf{u}}\right\}, \tag{10}$$

where

$$q(\mathbf{u}) = \int r(\mathbf{x})\delta(\mathbf{u} - \phi(\mathbf{x}))d\mathbf{x}, \tag{11}$$

and $\delta(\mathbf{x})$ is the Kronecker delta function.

From the theorem, it follows that

$$-\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right) - h^T(\mathbf{u})\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} \tag{12}$$

is an unbiased estimate of $E\{h^T(\mathbf{u})\boldsymbol{\theta}\}$.

Note, that as we show in the proof of the theorem, the pdf $f_u(\mathbf{u})$ of $\mathbf{u}$ is given by

$$f_u(\mathbf{u}) = \exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}q(\mathbf{u}). \tag{13}$$

Therefore, an alternative to computing $q(\mathbf{u})$ using (11) is to evaluate the pdf of $\mathbf{u}$ and then use (13).

Proof: To prove the theorem we first determine the pdf of $\mathbf{u}$. Since $\mathbf{u} = \phi(\mathbf{x})$ we have that [30, p. 127]

$$f_u(\mathbf{u}) = \int f(\mathbf{x};\boldsymbol{\theta})\delta(\mathbf{u} - \phi(\mathbf{x}))d\mathbf{x}. \tag{14}$$

Substituting (1) into (14),

$$\begin{aligned} f_u(\mathbf{u}) &= \exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\} \int r(\mathbf{x})\delta(\mathbf{u} - \phi(\mathbf{x}))d\mathbf{x} \\ &= \exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}q(\mathbf{u}). \end{aligned} \tag{15}$$

Now,

$$\begin{aligned} &E\left\{h^T(\mathbf{u})\boldsymbol{\theta}\right\} = \\ &= \int h^T(\mathbf{u})\boldsymbol{\theta}\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}q(\mathbf{u})d\mathbf{u} \\ &= \sum_{i=1}^{m}\int h_i(\mathbf{u})\theta_i\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}q(\mathbf{u})d\mathbf{u}. \end{aligned} \tag{16}$$

Noting that

$$\theta_i\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\} = \frac{d\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}}{du_i}, \tag{17}$$

we have

$$\begin{aligned} &\int_{-\infty}^{\infty} h_i(\mathbf{u})\theta_i\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}q(\mathbf{u})du_i = \\ &= \int_{-\infty}^{\infty} h_i(\mathbf{u})q(\mathbf{u})\frac{d\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}}{du_i}du_i \\ &= -\int_{-\infty}^{\infty}\frac{dh_i(\mathbf{u})q(\mathbf{u})}{du_i}\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}du_i, \end{aligned} \tag{18}$$

where we used the fact that $|h_i(\mathbf{u})q(\mathbf{u})\exp\{\boldsymbol{\theta}^T\mathbf{u} - g(\boldsymbol{\theta})\}| \to 0$ for $|u_i| \to \infty$ since $E\{h_i(\mathbf{u})\}$ is bounded. Now,

$$\frac{dh_i(\mathbf{u})q(\mathbf{u})}{du_i} = \frac{dh_i(\mathbf{u})}{du_i}q(\mathbf{u}) + \frac{dq(\mathbf{u})}{du_i}h_i(\mathbf{u}). \tag{19}$$

Substituting (18) and (19) into (16),

$$\begin{aligned} & E\left\{h^T(\mathbf{u})\boldsymbol{\theta}\right\} = \\ &= -\sum_{i=1}^{m}\int \frac{dh_i(\mathbf{u})q(\mathbf{u})}{du_i}\exp\{\boldsymbol{\theta}^T\mathbf{u}-g(\boldsymbol{\theta})\}d\mathbf{u} \\ &= \sum_{i=1}^{m}\left(-E\left\{\frac{dh_i(\mathbf{u})}{du_i}\right\}-E\left\{\frac{dq(\mathbf{u})}{du_i}\frac{h_i(\mathbf{u})}{q(\mathbf{u})}\right\}\right) \\ &= -E\left\{\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right)\right\}-E\left\{h^T(\mathbf{u})\frac{d\ln q(\mathbf{u})}{d\mathbf{u}}\right\}, \end{aligned} \tag{20}$$

which completes the proof of the theorem. ■

Based on Theorem 1 we can develop a generalized SURE principle for estimating an unknown parameter vector $\boldsymbol{\theta}$ in an exponential model. Specifically, let $\hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ be an ML estimate of $\boldsymbol{\theta}$ based on the data $\mathbf{x}$, let $\epsilon_{\mathrm{ML}}(\boldsymbol{\theta})$ be its MSE and let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})$ be an arbitrary estimate of $\boldsymbol{\theta}$ where $h(\mathbf{u})$ satisfies the regularity conditions of Theorem 1. Then, combining (6) and Theorem 1, an unbiased estimate of the MSE of $\hat{\boldsymbol{\theta}}$ is given by

$$\begin{aligned} S(h) &= \epsilon_{\mathrm{ML}}(\boldsymbol{\theta}) + \|h(\mathbf{u})\|^2 \\ &\quad +2\,\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right) + 2h^T(\mathbf{u})\left(\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} + \hat{\boldsymbol{\theta}}_{\mathrm{ML}}\right). \end{aligned} \tag{21}$$

We may then design $\hat{\boldsymbol{\theta}}$ by choosing $h(\mathbf{u})$ to minimize $S(h)$.

Another application of the SURE approach is to the problem of determining unknown regularization parameters which comprise a given estimation strategy. In the context of wavelet denoising, this method is used in the popular SureShrink method [10]. Extending this technique, our general SURE objective (21) can be used to select regularization parameters in more general models. We next discuss these ideas in the context of linear Gaussian problems.

IV. Linear Gaussian Model

In this section we specialize our results to the linear Gaussian model (2) with $\mathbf{H}$ an $n \times m$ matrix with $n \geq m$.

To use Theorem 1 we need to compute the pdf $q(\mathbf{u})$ of $\mathbf{u}$. This can be done by defining a transformation between $\mathbf{u}$ and $\mathbf{x}$. Since $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$, it must lie in the range $\mathcal{R}(\mathbf{H}^T)$ of $\mathbf{H}^T$. If $\mathbf{H}$ has full column-rank,

then the range is the entire space $\mathcal{R}^m$. Given $\mathbf{u} \in \mathcal{R}(\mathbf{H}^T)$, the possible choices of $\mathbf{x}$ are

$$\mathbf{x} = \mathbf{H}(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\mathbf{u} + \mathbf{C}\mathbf{v}, \tag{22}$$

where $\mathbf{v}$ is an arbitrary vector in the null space $\mathcal{N}(\mathbf{H}^T)$ of $\mathbf{H}^T$. It follows that given $\mathbf{u} \in \mathcal{R}(\mathbf{H}^T)$ there is a one-to-one correspondence between the vector $\mathbf{y} = [\mathbf{u}^T \quad \mathbf{v}^T]^T$ and $\mathbf{x}$. Furthermore, this relationship is linear. Thus we can write $\mathbf{y} = \mathbf{A}\mathbf{x}$ for some invertible matrix $\mathbf{A}$. Now,

$$\begin{aligned} q(\mathbf{u}_0) &= \int r(\mathbf{x})\delta(\mathbf{u}_0 - \phi(\mathbf{x}))d\mathbf{x} \\ &= |\mathbf{A}^{-1}| \int\int r(\mathbf{A}^{-1}\mathbf{y})\delta(\mathbf{u}_0 - \mathbf{u})d\mathbf{u}d\mathbf{v} \end{aligned} \tag{23}$$

where we used the change of variables $\mathbf{y} = \mathbf{A}\mathbf{x}$ and the fact that $\mathbf{y} = [\mathbf{u}^T \quad \mathbf{v}^T]^T$. Since $\mathbf{A}^{-1}\mathbf{y} = \mathbf{x}$, we can use $\mathbf{x}$ of (22) to write

$$\begin{aligned} r(\mathbf{A}^{-1}\mathbf{y}) = K \exp\{-(1/2)\mathbf{u}^T(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\mathbf{u}\}\cdot \\ \exp\{-(1/2)\mathbf{v}^T\mathbf{C}^{-1}\mathbf{v}\}, \end{aligned} \tag{24}$$

where K is a constant and we used the fact that $\mathbf{H}^T\mathbf{v} = 0$. Substituting (24) into (23),

$$\begin{aligned} q(\mathbf{u}_0) = K \exp\{-(1/2)\mathbf{u}_0^T(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\mathbf{u}_0\} \\ \int \exp\{-(1/2)\mathbf{v}^T\mathbf{C}^{-1}\mathbf{v}\}d\mathbf{v}. \end{aligned} \tag{25}$$

Therefore,

$$\frac{d \ln q(\mathbf{u})}{d\mathbf{u}} = -(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^\dagger\mathbf{u}. \tag{26}$$

It then follows from Theorem 1 that

$$E\left\{h^T(\mathbf{u})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\right\} = -E\left\{\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right)\right\}, \tag{27}$$

where we denoted

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}. \tag{28}$$

Note, that $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is an ML estimate of $\boldsymbol{\theta}$ for the model (2). If $\mathbf{H}$ has full column rank, then this is the unique ML solution.

We summarize our results for the linear Gaussian model in the following proposition.

Proposition 1: Let $\mathbf{x}$ denote measurements of an unknown parameter vector $\boldsymbol{\theta}$ in the linear Gaussian model (2), where $\mathbf{w}$ is a zero-mean Gaussian random vector with covariance $\mathbf{C} \succ 0$. Let $h(\mathbf{u})$ with $\mathbf{u} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{x}$ be an arbitrary function of $\boldsymbol{\theta}$ that is weakly differentiable in $\mathbf{u}$ and such that $E\{|h_i(\mathbf{u})|\}$ is bounded. Then

$$E\left\{h^T(\mathbf{u})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})\right\} = -E\left\{\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right)\right\}, \tag{29}$$

where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is an ML estimate of $\boldsymbol{\theta}$ and is given by (28). An unbiased estimate of the MSE of $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})$ is

$$S(h) = \epsilon_{\mathrm{ML}}(\boldsymbol{\theta}) + \|h(\mathbf{u})\|^2 + 2\,\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right). \tag{30}$$

Note that in deriving (30) we used (21) and (26).

A special case of Proposition 1 is the iid Gaussian model in which $\mathbf{H} = \mathbf{I}$ and $\mathbf{C} = \sigma^2\mathbf{I}$. This example was originally treated by Stein in [8], [9]. For this setup, he showed that

$$E\left\{h^T(\mathbf{x})(\boldsymbol{\theta} - \mathbf{x})\right\} = -\sigma^2\sum_{i=1}^{n} E\left\{\frac{dh(\mathbf{x})}{d\mathbf{x}_i}\right\}, \tag{31}$$

based on which he suggested the SURE estimate of the MSE:

$$n\sigma^2 + \|h(\mathbf{x})\|^2 + 2\sigma^2\sum_{i=1}^{n}\frac{dh(\mathbf{x})}{d\mathbf{x}_i}. \tag{32}$$

It is easy to see that (31) and (32) are a special case of Proposition 1. Indeed, in the iid model we have that $\mathbf{u} = (1/\sigma^2)\mathbf{x}$, and $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{x}$. Consequently $dh(\mathbf{u})/d\mathbf{u} = \sigma^2 dh(\mathbf{x})/d\mathbf{x}$ and $\epsilon_{\mathrm{ML}} = n\sigma^2$.

A. Examples

To illustrate the use of the SURE principle, suppose that we consider estimators of the form $\hat{\boldsymbol{\theta}} = \alpha\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ where $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is given by (28), and we would like to select a good choice of α. To this end, we minimize the

SURE unbiased estimate of the MSE given by Proposition 1 with $h(\mathbf{u}) = (\alpha - 1)\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$.

For this choice of $h(\mathbf{u})$, minimizing $S(h)$ is equivalent to minimizing

$$(1-\alpha)^2\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2 + 2(\alpha-1)\,\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\right). \tag{33}$$

The optimal choice of α is

$$\alpha = 1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}, \tag{34}$$

resulting in the estimate

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}\right)\hat{\boldsymbol{\theta}}_{\mathrm{ML}}. \tag{35}$$

The estimate of (35) coincides with the balanced blind minimax method proposed in [7, Eq. (45)], which was derived based on a minimax framework. Here we see that the same technique results from applying our generalized SURE criterion. A striking feature of this estimate, proved in [7], is that when $\mathbf{H}^*\mathbf{C}^{-1}\mathbf{H}$ is invertible and its effective dimension is larger than 4, it dominates ML for all values of $\boldsymbol{\theta}$ (see Theorem 3 in [7]). This means that its MSE is always lower than that of the ML method, regardless of the true value of $\boldsymbol{\theta}$.

When $\mathbf{H} = \mathbf{I}$ and $\mathbf{C} = \sigma^2\mathbf{I}$, (35) reduces to

$$\hat{\boldsymbol{\theta}} = \left(1 - \frac{n\sigma^2}{\|\mathbf{x}\|^2}\right)\mathbf{x}, \tag{36}$$

which coincides with Stein's estimate [1]. This technique is known to dominate ML for $n \geq 3$.

If in addition we require that $\alpha \geq 0$, then the estimate of (35) becomes

$$\hat{\boldsymbol{\theta}} = \left[1 - \frac{\mathrm{Tr}\left((\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{\dagger}\right)}{\|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}\|^2}\right]_+ \hat{\boldsymbol{\theta}}_{\mathrm{ML}}, \tag{37}$$

where we used the notation

$$[x]_+ = \begin{cases} x, & x \geq 0; \\ 0, & x \leq 0. \end{cases} \tag{38}$$

The method of (37) is a positive-part version of (35). In the iid case, it reduces to the positive-part Stein's estimate [31], which is known to dominate the standard Stein approach (36).

Next, consider the case in which $\mathbf{H} = \mathbf{I}$ and $\mathbf{C} = \mathbf{D}$ with $\mathbf{D} = \mathrm{diag}\,(\sigma_1^2, \ldots, \sigma_n^2)$ and suppose we seek a

diagonal estimate of the form $\hat{\theta}_i = \alpha_i x_i$. Since the ML solution in this case is $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbf{x}$, $h_i(\mathbf{u}) = (\alpha_i - 1)x_i$, and the unbiased estimate of (30) becomes

$$\sum_{i=1}^{n} \sigma_i^2 + \sum_{i=1}^{n} (1-\alpha_i)^2 x_i^2 + 2\sum_{i=1}^{n} \sigma_i^2(\alpha_i - 1). \tag{39}$$

Minimizing with respect to α_i yields

$$\alpha_i = 1 - \frac{\sigma_i^2}{x_i^2}. \tag{40}$$

Restricting the coefficients α_i to be non-negative leads to the estimate

$$\hat{\theta}_i = \left[1 - \frac{\sigma_i^2}{x_i^2}\right]_+ x_i. \tag{41}$$

In contrast to $\hat{\boldsymbol{\theta}}$ of (35), which dominates the ML method, it can be proved that the estimate of (41) is not dominating. Thus, we see that by allowing for too many free parameters, we impair the performance of the SURE-based estimate. On the other hand, assuming strong structure, as in (35), severely restricts the class of estimators and consequently limits the possible performance advantage which can be obtained. In Section VI we suggest a regularized SURE strategy in order to overcome this inherent tradeoff between over-parametrization and performance.

V. Application to Regularization Selection

A popular strategy for solving inverse problems of the form (2) is to use regularization techniques in conjunction with a least-squares objective. Specifically, the estimate $\hat{\boldsymbol{\theta}}$ is chosen to minimize a regularized least-squares criterion:

$$(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}})\mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\hat{\boldsymbol{\theta}}) + \lambda\|\mathbf{L}\hat{\boldsymbol{\theta}}\| \tag{42}$$

where the norm is arbitrary. Here $\mathbf{L}$ is some regularization operator such as the discretization of a first or second order differential operator that accounts for smoothness properties of $\boldsymbol{\theta}$, and λ is the regularization parameter [25], [24]. An important problem in practice is the selection of λ, which strongly effects the recovery performance. One of the most popular approaches to choosing λ is the generalized cross-validation (GCV) method [29].

Based on our generalized SURE criterion, we propose a new method for regularization selection. Specifically,

we choose λ to minimize the SURE objective (30). As we demonstrate for the case in which the norm in (42) is the ℓ_2-norm, this method can dramatically outperform GCV in practical applications.

For concreteness, suppose that the squared-ℓ_2 norm is used in (42). The solution then has the form

$$\hat{\boldsymbol{\theta}} = (\mathbf{Q} + \lambda \mathbf{L}^T \mathbf{L})^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}, \tag{43}$$

where for brevity we denoted

$$\mathbf{Q} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}. \tag{44}$$

The estimate (43) is commonly referred to as Tikhonov regularization [22]. In the GCV method, λ is chosen to minimize

$$G(\lambda) = \frac{1}{\mathrm{Tr}^2(\mathbf{I} - (\mathbf{Q} + \lambda \mathbf{L}^T \mathbf{L})^{-1} \mathbf{Q})} \sum_{i=1}^{n} (\mathbf{x}_i - [\mathbf{H}\hat{\boldsymbol{\theta}}]_i)^2. \tag{45}$$

To apply the SURE criterion, we rewrite the estimate (43) as $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})$ where $\mathbf{u} = \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x}$ and

$$h(\mathbf{u}) = -\lambda (\mathbf{Q} + \lambda \mathbf{L}^T \mathbf{L})^{-1} \mathbf{L}^T \mathbf{L} \mathbf{Q}^{-1} \mathbf{u}. \tag{46}$$

We then suggest choosing the value of λ that minimizes

$$S(\lambda) = \|h(\mathbf{u})\|^2 - 2\lambda \,\mathrm{Tr}\left((\mathbf{Q} + \lambda \mathbf{L}^T \mathbf{L})^{-1} \mathbf{L}^T \mathbf{L} \mathbf{Q}^{-1} \right), \tag{47}$$

which can be easily determined numerically.

Since $S(0) = 0$ it follows immediately that if the optimal λ is not zero, then $S(\lambda) < 0$ for all $\mathbf{x}$ which implies that $E(S(\lambda)) < 0$. Since

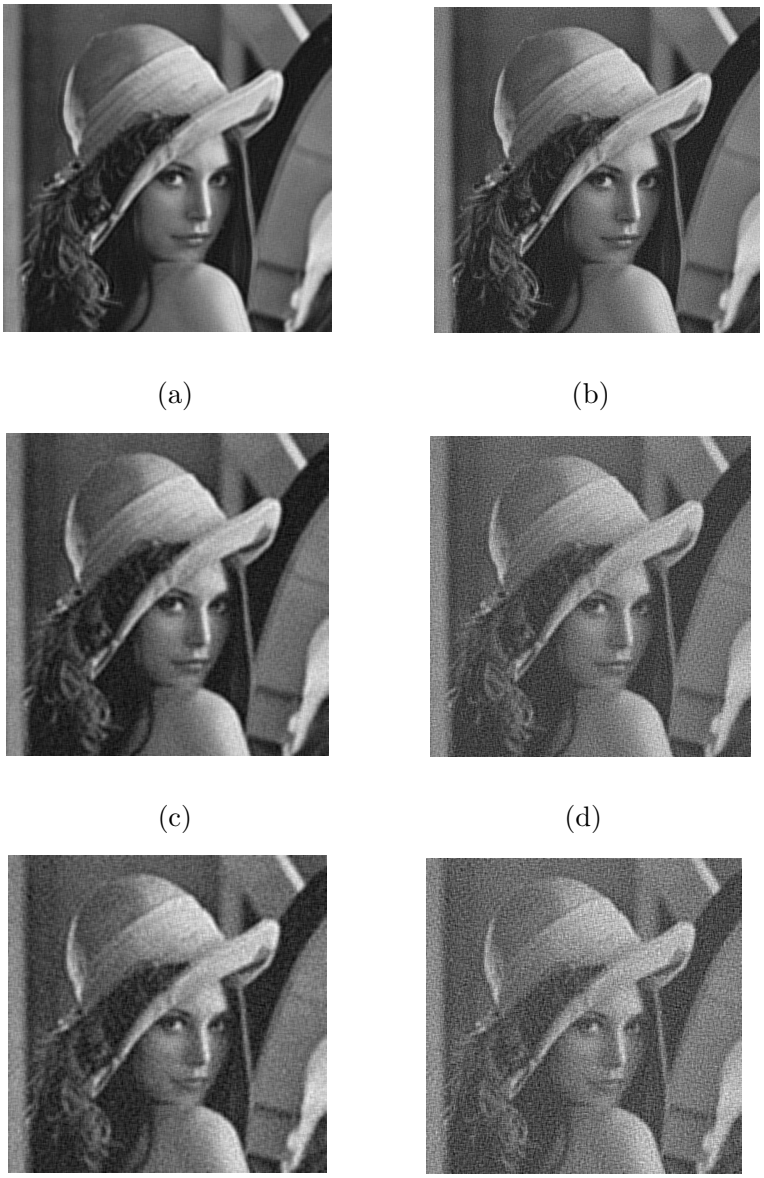
$$E\left\{ \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right\} - \epsilon_{\mathrm{ML}}(\boldsymbol{\theta}) = E(S(\lambda)), \tag{48}$$

we conclude that the resulting Tikhonov estimate will lead to smaller MSE than the ML technique for all values of $\boldsymbol{\theta}$. Note, that for the GCV choice of λ it is no longer true in general that $S(\lambda) < 0$.

To demonstrate the performance of our new regularization method, we tested it in the context of image deblurring using the HNO deblurring package for Matlab[1] based on [32]. We chose several test images, and

[1]The package is available at `http://www2.imm.dtu.dk/~pch/HNO/`.

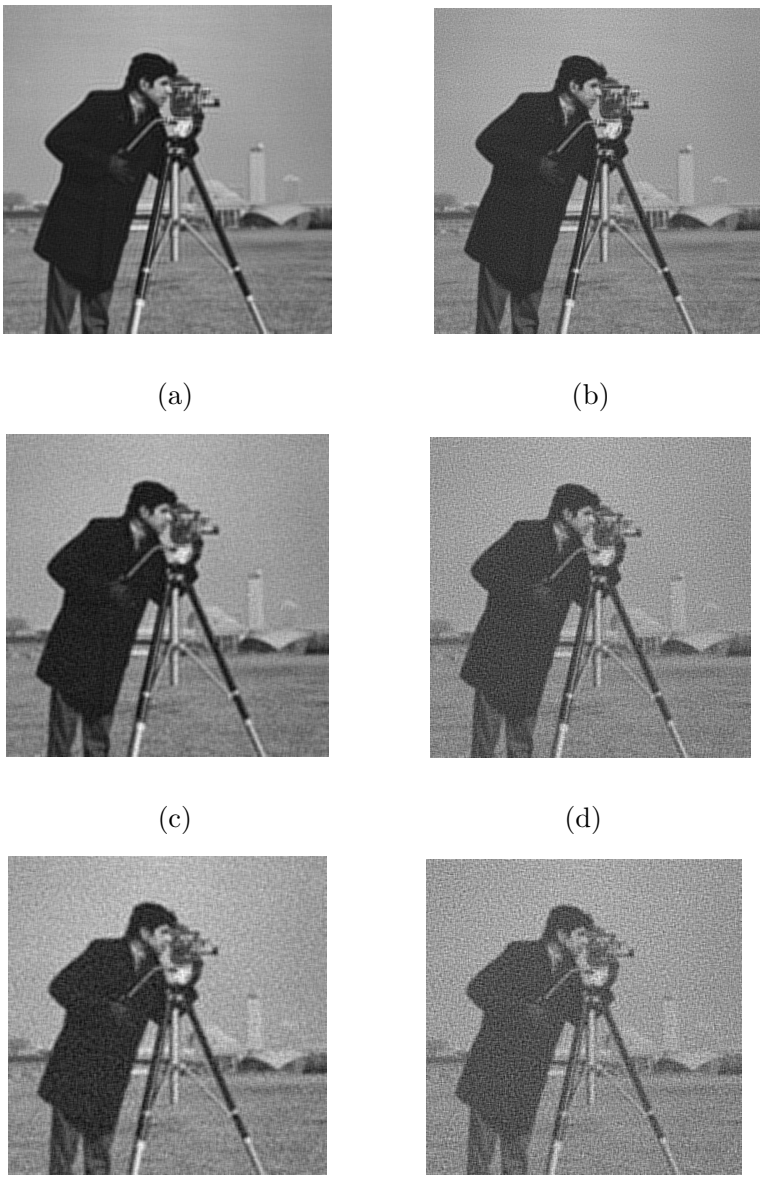
blurred them using a Gaussian point-spread function of dimension 9 with standard deviation 6. We then added zero-mean, Gaussian white noise with variance σ^2. In Figs. 1 and 2 we compare the deblurred images resulting from using the Tikhonov estimate (43) with $\mathbf{L} = \mathbf{I}$ where the regularization parameter is chosen according to our new SURE criterion (left) and the GCV method (right), for different noise levels.



(a) (b) (c) (d) (e) (f)

Fig. 1. Deblurring of Lena using Tikhonov regularization with SURE (left) and GCV (right) choices of regularization and different noise levels: (a), (b) $\sigma = 0.01$ (c),(d) $\sigma = 0.05$ (e),(f) $\sigma = 0.1$.

As can be seen from the figures, our SURE based approach leads to a substantial performance improvement over the standard GCV criterion. This can also be seen in Tables I and II in which we report the resulting



(a) (b)

(c) (d)

(e) (f)

Fig. 2. Deblurring of Cameraman using Tikhonov regularization with SURE (left) and GCV (right) choices of regularization and different noise levels: (a), (b) $\sigma = 0.01$ (c),(d) $\sigma = 0.05$ (e),(f) $\sigma = 0.1$.

MSE values.

VI. Regularized SURE Method

A crucial element in guaranteeing success of the SURE method is to choose a good parameterization of $h(\mathbf{u})$. However, in many contexts, such a structure may be hard to find. On the other hand, letting the SURE criterion select many free parameters can deteriorate its performance. One way to treat this inherent tradeoff

TABLE I
MSE FOR TIKHONOV DEBLURRING OF LENA

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$
GCV	0.0022	0.0077	0.0133
SURE	0.0011	0.0025	0.0042

TABLE II
MSE FOR TIKHONOV DEBLURRING OF CAMERAMAN

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$
GCV	0.0033	0.0121	0.0221
SURE	0.0016	0.0039	0.0064

is by regularization. Thus, instead of minimizing the SURE objective we suggest minimizing a regularized version:

$$\begin{aligned} S(h,\lambda) &= S(h) + \lambda g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})) \\ &= \epsilon_{\mathrm{ML}}(\boldsymbol{\theta}) + \|h(\mathbf{u})\|^2 + 2\,\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right) \\ &\quad +2h^T(\mathbf{u})\left(\frac{d\ln q(\mathbf{u})}{d\mathbf{u}} + \hat{\boldsymbol{\theta}}_{\mathrm{ML}}\right) + \lambda g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})), \end{aligned} \tag{49}$$

where λ is a regularization parameter and $g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u}))$ is a regularization function. For example, we may choose $g(\mathbf{v}) = \|\mathbf{v}\|$ where the norm is arbitrary. The parameter λ is determined by applying the conventional SURE (21) to the estimate $h(\mathbf{u}, \lambda)$ resulting from solving (49) with λ fixed. When $\mathbf{u}$ is Gaussian, (49) reduces to

$$\begin{aligned} S(h,\lambda) &= \epsilon_{\mathrm{ML}}(\boldsymbol{\theta}) + \|h(\mathbf{u})\|^2 \\ &\quad +2\,\mathrm{Tr}\left(\frac{dh(\mathbf{u})}{d\mathbf{u}}\right) + \lambda g(\hat{\boldsymbol{\theta}}_{\mathrm{ML}} + h(\mathbf{u})). \end{aligned} \tag{50}$$

As an example, consider the iid Gaussian model in which $\mathbf{x} = \boldsymbol{\theta} + \mathbf{w}$ where $\mathbf{w}$ is a Gaussian noise vector with iid zero-mean components of variance σ^2. Assuming that $\boldsymbol{\theta}$ represents the wavelet coefficients of some underlying signal $\mathbf{x}$, a popular estimation strategy is wavelet denoising in which each component of $\mathbf{x}$ is replaced by a soft or hard-thresholded version. In particular, in their landmark paper, Donoho and Johnstone

[10] developed a soft-threshold wavelet denoising method in which

$$\hat{\theta}_i = \begin{cases} |x_i| - t, & |x_i| \geq t; \\ 0, & |x_i| \leq t, \end{cases} \tag{51}$$

where t is a threshold value. They suggest selecting t to minimize the SURE criterion, and refer to the resulting estimate as SureShrink (to be more precise, in SureShrink t is determined by SURE only if it lower than some upper limit). In developing the SureShrink method, the function $h(\mathbf{x})$ is restricted to be a component-wise soft threshold. The motivation for this choice is that the wavelet coefficients below a certain level tend to be sparse. It is well known that soft-thresholding can be obtained as the solution to a least-squares criterion with an ℓ_1 penalty:

$$\min \left\{ \|\mathbf{x} - \boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}. \tag{52}$$

Thus, in principle we can view the SureShrink approach as a 2-step procedure: We first determine the estimate that minimizes an ℓ_1 penalized least-squares criterion. We then choose the penalization factor to minimize SURE.

Instead, we suggest choosing an estimate that directly minimizes an ℓ_1 regularized SURE objective, where the only assumption we make is that the processing is performed component wise. Thus, $\hat{\theta}_i = \alpha_i x_i$ for some coefficients $\alpha_i(\mathbf{x}) \geq 0$. Since $\mathbf{u} = (1/\sigma^2)x_i$, $h_i(\mathbf{u}) = \sigma^2(\alpha_i - 1)u_i$. With this choice of $h(\mathbf{u})$, minimizing (50) is equivalent to minimizing the following objective:

$$\mathcal{L}(\alpha) = \sum_{i=1}^{n} (\alpha_i - 1)^2 x_i^2 + 2\sigma^2 \sum_{i=1}^{n} \alpha_i + \lambda \sum_{i=1}^{n} |\alpha_i||x_i|. \tag{53}$$

The optimal choice of $\alpha_i \geq 0$ is

$$\alpha_i = \left[1 - \frac{\sigma^2 + \lambda |x_i|}{x_i^2} \right]_+ . \tag{54}$$

The resulting estimate can be viewed as a soft-thresholding method, with a particular choice of shrinkage (different than the standard approach (51)) when the value of x_i exceeds the threshold. The precise threshold value is equal to the largest value x_i for which $\alpha_i = 0$ and is given by

$$t = \frac{1}{2} \left(\lambda + \sqrt{\lambda^2 + 4\sigma^2} \right). \tag{55}$$

To choose λ, we substitute the estimate $\hat{\theta}_i = \alpha_i(\lambda)x_i$ with $\alpha_i(\lambda)$ given by (54) into the SURE criterion (30), and minimize with respect to λ. For this choice, the SURE objective becomes

$$n\sigma^2 + \sum_{i=1}^{n} \min\left(x_i^2, \left(\frac{\sigma^2}{|x_i|} + \lambda\right)^2\right) + 2\sigma^2 \sum_{i=1}^{n} s_i(\lambda), \tag{56}$$

where

$$s_i(\lambda) = \begin{cases} -1, & \sigma^2 + \lambda|x_i| \geq |x_i|^2 \\ \frac{\sigma^2}{x_i^2}, & \sigma^2 + \lambda|x_i| < |x_i|^2. \end{cases} \tag{57}$$

The value of λ can be easily determined numerically.

To demonstrate the advantage of our method over conventional soft-thresholding we implemented our approach on the examples taken from [10]. Specifically, we used the test functions Blocks, Bumps, HeaviSine and Doppler defined in [10]. The length of all signals is 2048 and the noise variance is $\sigma^2 = 4$. We used the Daubechies 8 symmetrical wavelet, and $L = 5$ levels are considered. In Table III we report the empirical MSE values of the original noisy signals, and 3 wavelet denoising schemes: SureShrink which is the method of [10] with the threshold selected using SURE, our proposed regularized SURE method (RSURE), and OracleShrink which is a soft-threshold where the threshold value is selected to minimize the squared-error between the true unknown wavelet coefficient, and its denoised version. Clearly this approach is only for comparison and serves as a benchmark on the best possible performance that can be obtained using any soft threshold. As can

TABLE III
MSE for Different Soft Denoising Schemes

	Blocks	Bumps	HeaviSine	Doppler
Original	4.054	4.072	4.153	3.945
SureShrink	0.744	0.875	0.205	0.290
RSure	0.694	0.816	0.169	0.273
OracleShrink	0.690	0.828	0.118	0.283

be seen from the table, the regularized SURE method performs better in all cases than SureShrink. It is also interesting to see that it sometimes even outperforms OracleShrink which is based on the true unknown $\boldsymbol{\theta}$. The reason the performance can be better than the oracle is that the shrinkage performed in RSURE is different than the conventional soft threshold.

In Table IV we repeat our experiments where now we use the estimates resulting from the standard SURE

criterion. Specifically, we consider the positive-part Stein estimate (35) referred to as SteinShrink and the estimate (41) which we refer to as ScalarShrink. Evidently, using the SURE estimate without regularization

TABLE IV
MSE for Different Denoising Schemes

	Blocks	Bumps	HeaviSine	Doppler
ScalarShrink	1.043	1.362	0.161	0.594
SteinShrink	1.681	1.730	1.508	1.413

deteriorates the performance significantly. Thus, SURE alone is not generally sufficient to obtain good estimates. However, adding regularization dramatically improves the behavior without the need to pre-specify the desired structure.

Finally, in Table V we repeat the experiments of Table III to determine the threshold values, but once the values are found we apply hard-thresholding on the coefficients. As can be seen from the table, even though

TABLE V
MSE for Different Hard-Thresholding Schemes

	Blocks	Bumps	HeaviSine	Doppler
SureShrink	1.902	1.961	0.988	0.630
RSure	1.560	1.912	0.766	0.700

the thresholding operation is now the same in both methods, RSURE performs significantly better. Thus, the threshold determined from this method is superior to that resulting from the SURE criterion without regularization. Here again the importance of regularization is demonstrated.

VII. Conclusion

In this paper, we developed an unbiased estimate of the MSE in multivariate exponential families by extending the SURE method. This generalized principle can now be used in exponential multivariate estimation problems to develop estimators with improved performance over existing approaches. As an application, we suggested a new strategy for choosing the regularization parameter in penalized inverse problems. We demonstrated via several examples that when using ℓ_2 regularization this method can significantly improve the MSE over the standard GCV approach. We also suggested a regularized SURE criterion for selecting estimators without the need for pre-specifying their structure. Applying this objective in the context of wavelet denoising, we proposed a new type of soft-thresholding which minimizes a penalized estimate of the MSE.

As we demonstrated, this strategy can lead to improved MSE behavior in comparison with soft and hard thresholding methods.

The main contribution of this work is in introducing the generalized SURE criterion and the regularized SURE method and demonstrating their applicability in several examples. In future work, we intend to develop these applications in more detail and further explore the practical use of the proposed design objectives.

VIII. Acknowledgement

The author would like to thank Zvika Ben-Haim and Michael Elad for fruitful discussions, and Amir Beck for help with the deblurring examples.

References

[1] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proc. Third Berkeley Symp. Math. Statist. Prob.* 1956, vol. 1, pp. 197–206, University of California Press, Berkeley.

[2] W. James and C. Stein, "Estimation of quadratic loss," in *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* 1961, vol. 1, pp. 361–379, University of California Press, Berkeley.

[3] W. E. Strawderman, "Proper Bayes minimax estimators of multivariate normal mean," *Ann. Math. Statist.*, vol. 42, pp. 385–388, 1971.

[4] K. Alam, "A family of admissible minimax estimators of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 1, pp. 517–525, 1973.

[5] J. O. Berger, "Admissible minimax estimation of a multivariate normal mean with arbitrary quadratic loss," *Ann. Statist.*, vol. 4, no. 1, pp. 223–226, Jan. 1976.

[6] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimators: Improving on least squares estimation," in *IEEE Workshop on Statistical signal Processing (SSP'05), Bordeaux, France*, July 2005.

[7] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Trans. Inform. Theory*, vol. 53, pp. 3145–3157, Sep. 2007.

[8] C. M. Stein, "Estimation of the mean of a multivariate distribution," *Proc. Prague Symp. Asymptotic Statist.*, pp. 345–381, 1973.

[9] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, no. 6, pp. 1135–1151, Nov. 1981.

[10] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *J. Am. Stat. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[11] X. P. Zhang and M. D. Desai, "Adapting denoising based on SURE risk," *IEEE Signal Process. Lett.*, vol. 5, no. 10, pp. 265–267, 1998.

[12] F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 593–606, 2007.

[13] A. Benazza-Benyahia and J.-C. Pesquet, "Building robust wavelet estimators for multicomponent images using Stein's principle," *IEEE Trans. Image Process.*, vol. 14, no. 11, pp. 1814–1830, 2005.

[14] R. Averkamp and C. Houdre, "Stein estimation for in
nitely divisible laws," *ESAIM: Probability and Statistics*, p. 269, 2006.

[15] H. M. Hudson, "A natural identity for exponential families with applications in multiparameter estimation," *Ann. Statist.*, vol. 6, no. 3, pp. 473–484, 1978.

[16] J. Berger, "Improving on inadmissible estimators in continuous exponential families with applications to simultaneous estimation of gamma scale parameters," *Ann. Stat.*, vol. 8, no. 3, pp. 545–571, 1980.

[17] J. T. Hwang, "Improving upon standard estimators in discrete exponential families with applications to Poisson and negative binomial cases," *Ann. Statist.*, vol. 10, no. 3, pp. 857–867, 1982.

[18] E. Pitman, "Sufficient statistics and intrinsic accuracy," *Proc. Camb. phil. Soc.*, vol. 32, pp. 567–579, 1936.

[19] G. Darmois, "Sur les lois de probabilites a estimation exhaustive," *C.R. Acad. sci. Paris*, vol. 200, pp. 1265–1266, 1935.

[20] B. Koopman, "On distribution admitting a sufficient statistic," *Trans. Amer. math. Soc.*, vol. 39, pp. 399–409, 1936.

[21] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, New York, NY: Springer-Verlag, Inc., second edition, 1998.

[22] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, Washington, DC: V.H. Winston, 1977.

[23] N. P. Galatsanos and A. K. Katsaggelos, "Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation," *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, 1992.

[24] P. C. Hansen, "The use of the L-curve in the regularization of discrete ill-posed problems," *SIAM J. Sci. Stat. Comput.*, vol. 14, pp. 1487–1503, 1993.

[25] M. Hanke and P. C. Hansen, "Regularization methods for large-scale problems," *Surveys Math. Indust.*, vol. 3, no. 4, pp. 253–315, 1993.

[26] A. Björck, *Numerical Methods for Least-Squares Problems*, Philadelphia, PA: SIAM, 1996.

[27] R. Molina, A. K. Katsaggelos, and J. Mateos, "Bayesian and regularization methods for hyperparameter estimation in image restoration," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, 1999.

[28] W. C. Karl, "Regularization in image restoration and reconstruction," in *Handbook of Image and Video Processing*, A. Bovik, Ed., pp. 183–202. ELSEVIER, 2nd edition, 2005.

[29] G.H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, May 1979.

[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Upper Saddle River, NJ: Prentice Hall, Inc., 1993.

[31] A. J. Baranchik, "Multiple regression and estimation of the mean of a multivariate normal distribution," Tech. Rep. 51, Stanford University, 1964.

[32] P. C. Hansen, J. G. Nagy, and D. P. OLeary, *Deblurring Images: Matrices, Spectra, and Filtering*, Philadelphia, PA: SIAM, 2006.