Optimal Signal Estimation Using Cross-Validation

Robert D. Nowak, *Member, IEEE*[*]

Abstract—**This letter develops an optimal, nonlinear estimator of a deterministic signal in noise. The methods of penalized least squares and cross validation balance the bias-variance tradeoff and lead to a closed form expression for the estimator. The estimator is simultaneously optimal in a "small-sample," predictive sum of squares sense and asymptotically optimal in the mean square sense.**

I. Introduction

This letter considers the problem of estimating a deterministic signal $\mathbf{s} \in \mathbb{R}^d$ in a zero mean noise $\boldsymbol{\eta}$. We assume $n > 1$, i.i.d. observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i = \mathbf{s} + \boldsymbol{\eta}_i$. The noise distribution is unknown, and we only assume $\mathrm{E}[\boldsymbol{\eta}^T\boldsymbol{\eta}] < \infty$.

The sample mean $\widehat{\mathbf{s}} = n^{-1}\sum_{j=1}^n \mathbf{x}_j$ is an unbiased estimator of $\mathbf{s}$. We may also represent $\mathbf{s}$ with respect to the orthonormal basis $\{\mathbf{b}_i\}_{i=1}^d$ as $\mathbf{s} = \sum_{i=1}^d \theta_i \mathbf{b}_i$, where θ_i are the coordinates of the signal in this basis. In matrix notation $\mathbf{s} = \mathbf{B}\boldsymbol{\theta}$, where $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_d]$ and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_d]^T$. The problem of estimating $\mathbf{s}$ is then equivalent to estimating the coefficients $\boldsymbol{\theta}$.

It is well-known that estimation performance is greatly improved when the basis efficiently represents the unknown signal. Selection of an appropriate basis is one way to incorporate prior signal information into the estimation procedure. For example, smooth wavelet bases are optimal bases for estimation of signals that may contain some points of discontinuity but otherwise are smooth [2]. In this letter, we assume that the basis is fixed and concentrate on estimating $\boldsymbol{\theta}$ rather than $\mathbf{s}$ directly. The sample estimator of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = n^{-1}\sum_{j=1}^n \mathbf{B}^T\mathbf{x}_j \tag{1}$$

Note that $\widehat{\mathbf{s}} = \mathbf{B}\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}} = \mathbf{B}^T\widehat{\mathbf{s}}$.

The estimator $\widehat{\boldsymbol{\theta}}$ may also be written as the solution to the least squares problem

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \|\widehat{\mathbf{s}} - \mathbf{B}\,\boldsymbol{\theta}\|^2, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm.

Now define the subspace $\Theta_m = \{\boldsymbol{\theta} : \theta_i = 0,\ i > m\}$ and the constrained estimator

$$\bar{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta_m} \|\widehat{\mathbf{s}} - \mathbf{B}\,\boldsymbol{\theta}\|^2. \tag{3}$$

The advantage of the constrained estimator is that it is insensitive to the noise component in the span of $\{\mathbf{b}_i\}_{i>m}$ and consequently may have a smaller variance than the sample estimator. However, the reduction in estimator variance may be obtained at the expense of introducing an estimator bias.

This letter considers a penalized least squares approach to balancing the trade-off between bias and variance. Using a predictive sum of squares (PRESS) criterion and the method of cross validation [1, 5], we determine the proper amount of penalization. We show that the PRESS-optimal solution can be computed in closed form.

The results in this paper are related to many well known methods in statistical signal processing. For example, see [1,3,5,6]. These methods are based on the linear statistical model. The linear statistical model[1] assumes a single, indirect observation of the unknown signal parameters $\boldsymbol{\theta}$:

$$\mathbf{y} = \mathbf{H}\,\boldsymbol{\theta} + \mathbf{n}, \tag{4}$$

where $\mathbf{H}$ is a known matrix and $\mathbf{n}$ is a vector of scalar, zero mean, i.i.d. noises. In this model, the signal $\mathbf{H}\,\boldsymbol{\theta}$ has a parametric form specified by the columns of $\mathbf{H}$.

Our problem formulation is quite different from the linear statistical model. We directly observe *multiple vector observations* of a deterministic signal $\mathbf{s}$, each with i.i.d. *vector noises*. The most distinguishing feature of our formulation is that *the distribution of the noise vectors is completely unknown* and the scalar noises within each vector are not necessarily i.i.d.

II. Penalized Least Squares Estimation of $\boldsymbol{\theta}$

A penalized least squares estimator for $\boldsymbol{\theta}$ is

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Lambda}) = \arg\min_{\boldsymbol{\theta}} \|\widehat{\mathbf{s}} - \mathbf{B}\,\boldsymbol{\theta}\|^2 + \boldsymbol{\theta}^T\boldsymbol{\Lambda}\boldsymbol{\theta}, \tag{5}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with entries $\lambda_i \geq 0$. The term $\boldsymbol{\theta}^T\boldsymbol{\Lambda}\boldsymbol{\theta}$ is called the *penalizing functional.* Increasing λ_i increases the cost associated with signal estimates having a large component in the $\mathbf{b}_i$ direction. Hence, the penalizing functional can be designed to weight undesirable or unlikely solutions more heavily than desirable or likely solutions.

It is easily verified that the unique minimizer is $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Lambda}) = (\mathbf{I} + \boldsymbol{\Lambda})^{-1}\,\widehat{\boldsymbol{\theta}}$. Furthermore, because $\boldsymbol{\Lambda}$ is diagonal, $(\mathbf{I} + \boldsymbol{\Lambda})^{-1} = \mathbf{I} - \boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is diagonal with entries $\gamma_i = \frac{\lambda_i}{1+\lambda_i}$. Since each λ_i ranges over $[0, \infty)$, we have $\gamma_i \in [0, 1]$. The γ_i are referred to as *regularization parameters.*

Rewriting $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Lambda})$ as a function of $\boldsymbol{\Gamma}$ produces

$$\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Gamma}) = \widehat{\boldsymbol{\theta}} - \boldsymbol{\Gamma}\,\widehat{\boldsymbol{\theta}}, \tag{6}$$

where $\widehat{\boldsymbol{\theta}}$ is the sample mean estimator (1). Equation (6) shows that $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Lambda})$ is a type of shrinkage estimator [3, 6];

[*]Supported by the National Science Foundation, grant no. MIP–94–57438, and the Office of Naval Research, grant no. N00014–95–1–0849. The author is with the Department of Electrical and Computer Engineering, Rice University, P.O. Box 1892, Houston, TX 77251–1892, USA. Email: nowak@ece.rice.edu

[1] This model is also referred to as the linear regression model in the statistics literature.

$\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Gamma})$ "shrinks" $\widehat{\boldsymbol{\theta}}$ closer to the origin. Setting $\gamma_i = 0$, $i = 1,\ldots,m$ and $\gamma_i = 1$, $i = m+1,\ldots,d$ yields the constrained estimator (3).

Note that we can study the relationship between $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Gamma})$ and $\boldsymbol{\Gamma}$ along each coordinate $\widetilde{\theta}_i(\gamma_i) = \widehat{\theta}_i - \gamma_i\,\widehat{\theta}_i$ independently. The theoretically optimal γ_i, in the MSE sense, is easily expressed. The value of γ_i minimizing $\mathrm{E}[|\theta_i - \widetilde{\theta}_i(\gamma_i)|^2]$ is

$$\gamma_i^{(\mathrm{MSE})} = \frac{\sigma_i^2}{\sigma_i^2 + n\,\theta_i^2}, \tag{7}$$

where $\sigma_i^2 = \mathrm{E}[(\mathbf{b}_i^T(\mathbf{x}-\mathbf{s}))^2]$ and $\theta_i^2 = (\mathbf{b}_i^T\mathbf{s})^2$, the noise and signal power, respectively, in the subspace spanned by $\mathbf{b}_i$. The MSE-optimal estimator is $\widetilde{\theta}_i^{(\mathrm{MSE})} = \widehat{\theta}_i - \gamma_i^{(\mathrm{MSE})}\,\widehat{\theta}_i$.

A simple approach at this point would be to compute sample estimates of σ_i^2 and θ_i^2 from the data $\mathbf{x}_1,\ldots,\mathbf{x}_n$ and then plug these estimates into (7) to obtain an estimate of $\gamma_i^{(\mathrm{MSE})}$. However, this estimate directly involves the estimate of the unknown parameter θ_i. Alternatively, we consider choosing γ_i using the method of cross validation.

III. Estimating the Regularization Parameters by Cross-Validation

Cross validation is a standard procedure for assessing the performance of an estimator [1]. To formulate a cross validation procedure for this problem, first compute a penalized least squares estimate of $\boldsymbol{\theta}$ with the jth data vector $\mathbf{x}_j$ omitted: $\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Gamma})_j = (\mathbf{I}-\boldsymbol{\Gamma})\,\frac{1}{n-1}\sum_{k\neq j}\mathbf{B}^T\mathbf{x}_k$. This "leaving-one-out" estimate is then used to predict $\mathbf{B}^T\mathbf{x}_j$. The predictive sum of squares (PRESS)

$$V(\boldsymbol{\Gamma}) = \frac{1}{n}\sum_{j=1}^{n}\|\widetilde{\boldsymbol{\theta}}(\boldsymbol{\Gamma})_j - \mathbf{B}^T\mathbf{x}_j\|^2, \tag{8}$$

may be viewed as a small-sample optimality criterion measuring the quality of the estimator and of the parameters $\boldsymbol{\Gamma}$. The objective is to choose $\boldsymbol{\Gamma}$ to minimize $V(\boldsymbol{\Gamma})$.

A closed form expression for the minimizer exists. After some algebra, we have

$$V(\boldsymbol{\Gamma}) = \left(\frac{n}{n-1}\right)\sum_{i=1}^{d}\left[\widehat{\sigma}_i^2 - \gamma_i\,2n^{-1}\widehat{\sigma}_i^2 + \gamma_i^2\,n^{-2}(\widehat{\sigma}_i^2 + n(n-1)\widehat{\theta}_i^2\right],$$

where $\widehat{\theta}_i = n^{-1}\sum_{j=1}^{n}\mathbf{b}_i^T\mathbf{x}_j$, and unbiased estimate of θ_i, and $\widehat{\sigma}_i^2 = (n-1)^{-1}\sum_{j=1}^{n}[\mathbf{b}_i^T\mathbf{x}_j - \widehat{\theta}_i]^2$, an unbiased estimate of σ_i^2.

We can study the quadratic relationship between $V(\boldsymbol{\Gamma})$ and $\boldsymbol{\Gamma}$ along each coordinate independently. Setting $\frac{dV}{d\gamma_i} = 0$ and solving for γ_i produces

$$\gamma_i^{(\min)} = \frac{\widehat{\sigma}_i^2}{n^{-1}\widehat{\sigma}_i^2 + (n-1)\widehat{\theta}_i^2}. \tag{9}$$

Note that $\gamma_i^{(\min)} \geq 0$. It is easily verified that $\gamma_i^{(\min)} > 1$ implies that V is strictly decreasing on $0 \leq \gamma_i \leq 1$. Therefore, if $\gamma_i^{(\min)} \geq 1$, then the PRESS is minimized at $\gamma_i = 1$. Hence, the PRESS-optimal choice of γ_i is

$$\gamma_i^{(\mathrm{PRESS})} = \tau\left(\frac{\widehat{\sigma}_i^2}{n^{-1}\widehat{\sigma}_i^2 + (n-1)\widehat{\theta}_i^2}\right), \tag{10}$$

where τ is the threshold nonlinearity

$$\tau(\gamma) = \begin{cases} \gamma, & \gamma < 1 \\ 1, & \gamma \geq 1. \end{cases} \tag{11}$$

The PRESS-optimal estimator is $\widetilde{\theta}_i^{(\mathrm{PRESS})} = \widehat{\theta}_i - \gamma_i^{(\mathrm{PRESS})}\widehat{\theta}_i$.

Note that $\gamma_i^{(\mathrm{PRESS})}$ is quite different from simply estimating $\gamma_i^{(\mathrm{MSE})}$ from the sample statistics. In fact, plugging the unbiased estimates $\widehat{\sigma}_i^2$ and $\widehat{\theta}_i^2$ into (7) produces a estimate $\widehat{\gamma}_i^{(\mathrm{MSE})} \leq \gamma_i^{(\mathrm{PRESS})}$. Hence, the estimate $\widehat{\gamma}_i^{(\mathrm{MSE})}$ may be over-conservative in the PRESS sense.

It is easily verified that $0 < \gamma_i^{(\mathrm{PRESS})} < 1$ whenever $0 < n^{-1}\widehat{\sigma}_i^2 < \widehat{\theta}_i^2$. The two extremes at $\gamma_i^{(\mathrm{PRESS})} = 1$ and $\gamma_i^{(\mathrm{PRESS})} = 0$ are especially interesting.

$\gamma_i^{(\mathrm{PRESS})} = 1$ ($\widetilde{\theta}_i^{(\mathrm{PRESS})} = 0$) if and only if $\widehat{\theta}_i^2 \leq n^{-1}\widehat{\sigma}_i^2$. Given n i.i.d. observations, the quantity $n^{-1}\widehat{\sigma}_i^2$ is an estimate of the averaged noise power in the subspace $\mathcal{S}_i$ spanned by $\mathbf{b}_i$ (the true noise power is $n^{-1}\sigma_i^2$). $\widehat{\theta}_i^2$ is an estimate signal power in $\mathcal{S}_i$. Hence, the PRESS-optimal estimate of θ_i is zero if and only if the SNR in $\mathcal{S}_i$ is less than or equal to 0dB.

$\gamma_i^{(\mathrm{PRESS})} = 0$ ($\widetilde{\theta}_i^{(\mathrm{PRESS})} = \widehat{\theta}_i$) if and only if $\widehat{\sigma}_i^2 = 0$. This shows that the sample mean estimator $\widehat{\theta}_i$ is suboptimal in the PRESS sense *except* when there is no noise in $\mathcal{S}_i$. This situation does not occur very often: In [4] it is shown that if the distribution of $\mathbf{x}$ is absolutely continuous, then $\mathrm{Prob}[\gamma_i^{(\mathrm{PRESS})} = 0] = 0$. It is also shown that under the same condition $\mathrm{Prob}[\gamma_i^{(\mathrm{MSE})} = 0] = 0$, where $\gamma_i^{(\mathrm{MSE})}$ is the MSE-optimal regularization parameter (7). One interpretation of this result is the following. If $\mathbf{x}$ is absolutely continuous, then it is always possible to reduce the estimator variance by an amount $\epsilon > 0$ at the expense of adding a squared bias $< \epsilon$. This type of result is widely known in the statistical literature [3].

In addition to the PRESS-optimality of $\widetilde{\boldsymbol{\theta}}^{(\mathrm{PRESS})}$, we can show that $\widetilde{\boldsymbol{\theta}}^{(\mathrm{PRESS})}$ is asymptotically optimal in the mean square sense, that is $\gamma_i^{(\mathrm{PRESS})}/\gamma_i^{(\mathrm{MSE})} \to 1$ *w.p.*1. The result is obtained by a simple application of the strong law of large numbers; the details are given in [4]. Arguing along similar lines, it is easily established that $\widetilde{\boldsymbol{\theta}}^{(\mathrm{PRESS})} \to \mathbf{B}^T\mathbf{s}$ *w.p.*1.

IV. Numerical Example

To demonstrate the performance of the PRESS-optimal estimator, consider the problem of estimating the intensity of a spatially varying Poisson process. This problem arises in planar nuclear medicine imaging, for example.

The maximum likelihood estimate (MLE) of the intensity is proportional to the the total counts (over the entire observation period T) in each detector bin. We advocate splitting the observation period T into n intervals, each of time duration T/n, providing n independent observations needed to form the PRESS-optimal estimator.

The true intensity $\mathbf{s}$ depicted in Fig. 1 (a). The performance of the MLE and the PRESS-optimal estimator are compared in 50 independent trials. The intensity of the simulated process is adjusted so that the maximum number of counts/pixel is approximately 100 (total counts over the entire observation period T). The observation period T is split into $n = 25$ intervals. The Haar wavelet basis is chosen to represent the unknown intensity, because of its excellent localization and approximation properties. Table I summarizes the performance of the MLE and PRESS-optimal estimator. All quantities are normalized by the true intensity. The MLE and PRESS-optimal estimate obtained in a typical trial are shown in Fig. 1 (b) and (c), respectively.

Table 1: Comparison of MLE and PRESS-optimal estimators of Poisson intensity.

Estimator	Bias Squared	Variance	M.S.E.
MLE $\widehat{\mathbf{s}}$	0.0012	0.0551	0.0563
$\widetilde{\mathbf{s}}^{(\text{PRESS})}$	0.0018	0.0246	0.0263

V. Related Work and Conclusions

In this letter we have developed an optimal signal estimator based on the method of cross validation. Applications of the estimator include moment estimation in array and signal processing, in which case the moment plays the rôle of signal and the noise is the variability of the data about the moment.

The PRESS-optimal estimator is closely related to James-Stein and related "shrinkage" estimators [2, 3, 5, 6]. Stone was apparently the first to draw the connection between shrinkage estimators and cross validation in the context of the linear statistical model [5].

One of the interesting features of the PRESS-optimal estimator is that the amount of shrinkage is determined independently for each coordinate and this amount is related to the estimated SNR in the coordinate. In this respect, the PRESS-optimal estimator is similar to the data adaptive rank shaping methods developed in [6]. The methods in [6] are based on the linear statistical model and assume an additive white noise with known variance. Their methods take full advantage of this prior information and produce a suite nonlinear, data-adaptive filters that dramatically outperform the classical Wiener filter in many cases.

In contrast, the PRESS-optimal estimator does not require any prior knowledge of the noise, except that the noise is zero mean has finite second order moments. We do, however, require $n > 1$ i.i.d. noisy observations of the signal. The methods of [6] only require a single observation. Also, the PRESS-optimal estimator is simultaneously optimal in the small-sample PRESS sense and asymptotically optimal in the MSE sense. This is due to the fact that the PRESS-optimal estimator is derived using the method of CV. CV is not considered in [6].

Finally, note that the PRESS and MSE optimality of the PRESS-optimal estimator is *with respect to the basis* $\mathbf{B}$. Although the rate of convergence does not depend on $\mathbf{B}$, the small-sample error is affected by the choice $\mathbf{B}$. Intuition suggests that the estimator will provide best results (small PRESS and MSE) if the basis $\mathbf{B}$ is well-matched to the unknown signal. Although it is difficult to characterize the PRESS-optimal basis, this intuition is partially justified by considering the MSE-optimal basis. The following result is proved in [4]: Assume that the noise is white; that is, $\sigma_i^2 = \sigma^2 \ \forall i$, independent of the basis. Then the MSE, $\sum_{i=1}^{d} \mathrm{E}[(\theta_i - \widetilde{\theta}_i^{(\text{MSE})})^2]$, is minimized by every orthonormal basis with $\mathbf{b}_1 = \mathbf{s}/||\mathbf{s}||$. Ongoing work is aimed at the basis selection problem.

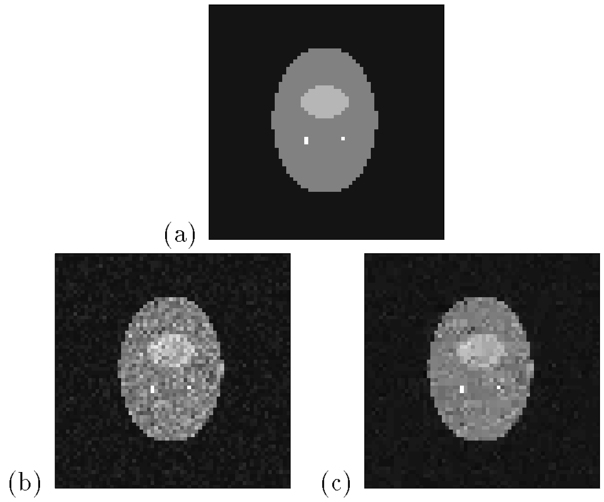

Figure 1: Estimation of spatial Poisson intensity. (a) True Poisson intensity, (b) MLE, (c) PRESS-optimal estimate.

References

[1] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, Feb. 1974.

[2] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Tran. on Info. Theory*, pp. 613-627, vol. 41, no. 3, May, 1995.

[3] N. R. Draper and R. C. Van Nostrand, "Ridge regression and James-Stein estimation: review and comments," *Technometrics*, vol. 21, no. 4, Nov. 1979.

[4] R. D. Nowak, "Optimal signal estimation using cross validation," *Rice University Technical Report*, ECE-TR-9601, Rice University, Houston, TX, Jan. 1996.

[5] M. Stone, Discussion, *J. Roy. Statist. Soc., B-35*, pp. 408-409, 1973.

[6] A. J. Thorpe and L. L. Scharf, "Data adaptive rank-shaping methods for solving least squares problems," *IEEE Tran. Signal Proc.*, vol. 43, no. 7, pp. 1591-1601, July 1995.