ON MULTIVARIATE ESTIMATION BY THRESHOLDING

Alyson K. Fletcher, Vivek K Goyal, and Kannan Ramchandran

University of California, Berkeley
Department of Electrical Engineering
alyson@eecs.berkeley.edu, v.goyal@ieee.org, kannanr@eecs.berkeley.edu

ABSTRACT

Despite their simplicity, scalar threshold operators effectively remove additive white Gaussian noise from wavelet detail coefficients of many practical signals. This paper explores the use of multivariate estimators that are almost as simple as scalar threshold operators. Şendur and Selesnick have recently shown the effectiveness of joint threshold estimation of parent and child wavelet coefficients. This paper discusses analogous results in two situations. With a frame representation, a simple joint threshold estimator is derived and it is shown that its generalization is equivalent to a type of ℓ_1-regularized denoising. Then, for the case where multiple independent noisy observations are available, the counter-intuitive results by Chang, Yu, and Vetterli on combining averaging and thresholding are explained as a fortuitous consequence of randomization.

1. INTRODUCTION

Consider the problem of estimating a random vector x from a noisy observation

$$y = x + n \tag{1}$$

where the joint distribution of the signal $x \in \mathbb{R}^N$ and noise $n \in \mathbb{R}^N$ is known. In principle, one can use the maximum a posteriori probability (MAP) or minimum mean-squared error (MMSE) criterion to determine a function g to generate estimates through $\hat{x} = g(y)$. The problem, of course, is that g may be very difficult to determine or apply. Henceforth we make the standard assumption that n is independent of x and has the i.i.d. Gaussian distribution $\mathcal{N}(0, \sigma^2 I_N)$; but this does not change the fact that an optimal estimator can be difficult to determine.

To reduce the complexity of the estimation procedure, one may require that g lie in a class of simple functions. For example, it is well known that one can require g to be a linear function and still obtain optimal (MAP or MMSE) estimates *when the signal is Gaussian.* Similarly, when the components of x are independent Laplacian random variables, the MAP estimate is obtained with component-wise soft thresholding of y. In this work, we explore situations where there are interesting estimators that are more complicated than acting component-wise (scalar), but have a simple and similar structure.

For review and motivation, scalar thresholding and a recent bivariate method [6] are discussed in Section 2. Then, in Section 3, a model motivated by the overcomplete expansion of signal is given. In a bivariate situation, this leads to a simple vector threshold estimator. More generally, this gives a new interpretation to a recent maximum entropy method [5]. Finally, we consider the case where $x_1 = x_2 = \cdots = x_N$, i.e., there are multiple noisy observations of a single random quantity. We demonstrate that the optimal estimate must be expressible as a function of $N^{-1}\sum_{i=1}^{N} y_i$ and reconcile this fact with the conclusions of [2].

2. MAP ESTIMATORS AND THRESHOLDS

Consider first the scalar ($N = 1$) version of (1) where x has the Laplacian distribution with zero mean, i.e., let x have probability density function $f_x(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$. A straightforward calculation shows that the MAP estimate of x from y is given by the standard *soft-thresholding operator*

$$\Lambda_\rho(y) = \begin{cases} y - \operatorname{sgn}(y)\cdot\rho & \text{if } |y| > \rho; \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

with *threshold* $\rho = \lambda\sigma^2$. The MMSE estimate has a more complicated closed form, but is also approximated by a soft threshold operator. The two estimators are shown in Fig. 1 for $\lambda = 1$ and $\sigma = 1$.

Empirically, it is reasonable to model the (detail) wavelet coefficients of images as independent Laplacian random variables. Treating x in (1) as a vector of independent Laplacian random variables allows the estimation procedure to be broken down to N component-wise operations. Thus, the ability to effectively estimate Laplacian signals with the estimator (2) provides a basic justification for thresholding wavelet detail coefficients as a denoising method for images. (More mathematical justifications based on signal smoothness classes are given in [4].)

One source of significant improvement in image denoising over the basic wavelet thresholding of [4] is to use spa-

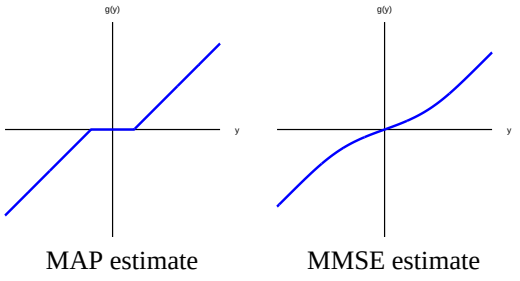

Fig. 1. MAP and MMSE estimates of a Laplacian signal.

tially adapted thresholds as in [1]. Although this makes the denoising operation not component-wise separable in the wavelet domain, this is not what we refer to as *vector thresholding*; information across subbands and from the spatial neighborhood is used to adjust the threshold ρ but not to change the argument to the threshold function Λ_ρ.

One method for vector thresholding is due to Şendur and Selesnick [6]. They consider a detail coefficient with its parent and use a density model where these are not independent. Then, they derive the MAP estimate for the detail coefficient given the noisy detail coefficient and noisy parent coefficient. This estimate has the qualitative aspects that make it a vector threshold operator applied to two-tuple.

3. MAP ESTIMATION OF VECTORS

When one uses nonorthogonal transformations—including overcomplete transformations—on a noisy signal, the noise components are dependent. One can exploit this dependence in estimation even when no specific form of signal dependence is assumed.

For example, consider our estimation problem with $N = 2$ and let the non-orthogonal transform matrix T be given by

$$T = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$$

for some constant a. Let $X = Tx$ and $Y = Ty$ be the transforms of the true and noisy vectors. A simple signal model that can be reasonably validated for detail coefficients of images is for X_1 and X_2 to be i.i.d. Laplacian variables. Fig. 2 plots the MAP estimate of the component X_1 as a function of Y_1 and Y_2 for the case when $a = 0.5$. It can be seen that X_1 is a threshold-like function of Y_1, but with a dependence on Y_2. The value $a = 0.5$ is motivated, for example, by the approximate correlation between a Daubechies highpass filter and its shift by one. (The orthogonality property holds for *even* shifts.)

The remainder of this section connects this type of thresholding that operates jointly over the components with the "maximum entropy" (MAXENT) method developed by Ishwar and Moulin [5] and gives further examples.

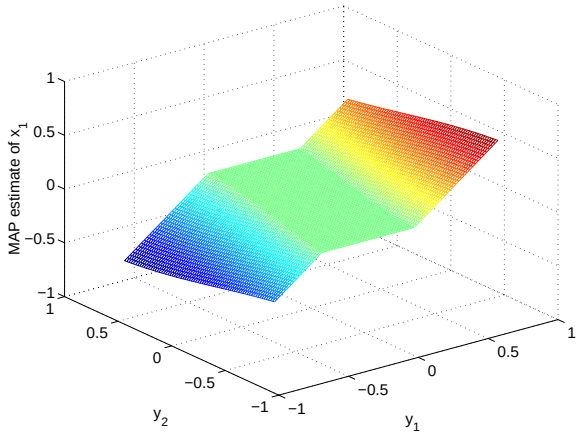

Fig. 2. Two-dimensional MAP estimate with a non-orthogonal transform T. Qualitatively, it has the form of a soft threshold applied to a linear combination of Y_1 and Y_2, with small weight on Y_2.

3.1. Maximum Entropy Method

The MAXENT approach provides a systematic way to combine information from several wavelet bases. We will see that it exhibits a number of "threshold-like" phenomena.

A simplified version is as follows: Take (1) to be in time domain. The wavelet transform of x is modeled by an $M \times N$ matrix T. In the single-basis problem, $M \leq N$ and T would represent the transform to the detail coefficients of the wavelet transform in a single basis. For the multibasis problem, we can take $T = [T_1^T \ \cdots \ T_K^T]^T$, where the T_k's are transforms to the detail coefficients in K different wavelet bases.

Given a transform matrix T, the MAXENT method assumes a *prior* distribution, $f(x)$, on the unknown signal x, of the form,

$$f(x) = C \exp(-\lambda \|Tx\|_p^p) \tag{3}$$

where $p \geq 1$, C and λ are constants and $\|\cdot\|_p$ represents the p-norm. The MAXENT method then estimates x from y by the standard maximum *a posteriori* (MAP) estimate $\arg\max_x f(x|y)$. Using the Gaussian distribution on n, the MAP estimate can be rewritten as

$$\hat{x} = \arg\min_x \left[\|y - x\|_2^2 + 2\sigma^2 \lambda \|Tx\|_p^p \right]. \tag{4}$$

To motivate the prior distribution $f(x)$ above, note that $f(x)$ is a product of Generalized Gaussian distributions on the components of Tx. In general, the exponent p of the distribution is selected such that $p < 2$, making the distributions "heavy-tailed". The heavy-tailed distribution models that the distribution of the detail coefficients of the wavelet transform of images tends to be sparse. Using the product

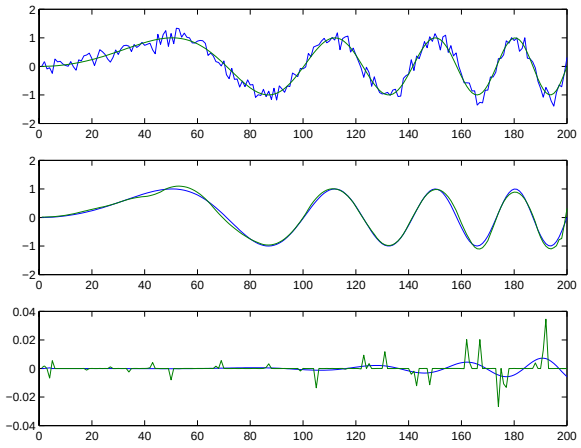

Fig. 3. MAXENT denoising of a chirp signal. Top: True and noisy signal (SNR=11.5 dB). Middle: True signal and estimate (SNR=19.0 dB). SNR with single-basis denoising is 14.3 dB and with cycle spinning is 14.8 dB. Bottom: Detail coefficients of the true signal and the estimate.

distribution models the wavelet coefficients as if they were independent, although, when the transform matrix T is tall, it is impossible for the components of Tx to be independent. Nevertheless, it is shown in [5] that the product distribution on the components of Tx is the maximum entropy distribution under a certain constraint on the pth moment $\mathbf{E}\|Tx\|_p^p$. Information theoretically, the maximum entropy distribution is a "maximally noncommittal" distribution in the sense that it makes minimal assumptions on the relationship between components of Tx.

3.2. Single Basis MAXENT Estimation

In the case of a single wavelet basis, the MAXENT method reduces to the classical wavelet threshold denoising. To see this, let W be an $N \times N$ matrix representing an orthogonal wavelet transform, and partition W into T_s and T_d to produce the scaling and detail components, respectively. In (4), if we take $p = 1$ and $T = T_d$ and use the orthogonality of W, the MAXENT estimate reduces to

$$\hat{x} = T_s^T T_s y + T_d^T \Lambda_\rho (T_d y), \tag{5}$$

where Λ_ρ is as given in (2) and $\rho = \lambda\sigma^2$. The estimator in (5) is precisely the standard soft-threshold estimate of x given y. The estimate is the sum of the scaling components of the noisy signal, along with thresholded detail coefficients.

3.3. Shift-Invariant Denoising Example

Now consider the use of MAXENT estimation for shift-invariant denoising. The wavelet transform is not in general shift-invariant. Consequently, different estimates can be obtained by shifting the noisy signal, applying standard wavelet thresholding and shifting the estimate back. If the wavelet transform has J stages, one can obtain up to 2^J different estimates in this manner. In [3], Donoho and Johnstone propose a method called "cycle spinning" which simply averages these 2^J estimates.

MAXENT estimation provides an alternative way to combine the information from the different shifts. It is not difficult to see that the components from the different shifts of the noisy signal can be obtained from performing a single wavelet transform without any decimation. Given an N-length input signal, the undecimated wavelet transform will result in $(J+1)N$ coefficients, JN of which will be detail coefficients. To use the MAXENT method above, we can let x and y be the true and noisy time-domain signals and let T be the $JN \times N$ matrix representing the undecimated transform for the detail coefficients.

Consider denoising the chirp signal $x[t]$ shown in the top panel in Fig. 3. For simplicity, we use a single-stage wavelet transform with the Daubechies D_4 filter pair. The MAXENT estimate was found by solving (4) with $p = 1$, using quadratic programming to perform the minimization.

A chirp signal is not a natural candidate for wavelet thresholding since it does not have any sharp discontinuities. Indeed, the bottom panel of Fig. 3 shows the undecimated detail coefficients of the signal, and we can see that the signal is not sparse in the wavelet domain. Nevertheless, we see two interesting phenomena. Firstly, the MAXENT estimate is able to denoise the signal well in comparison to single-basis denoising and cycle spinning. Secondly, as shown in the bottom panel of Fig. 3, although the true signal is not sparse in the wavelet domain, the MAXENT estimator finds an estimate that is sparse. In this sense, the MAXENT estimate can be seen as a vector thresholding operation which zeros out most coefficients while preserving the key components to model the signal.

4. ESTIMATION FROM MULTIPLE OBSERVATIONS

Now suppose that in (1) we have x_1 Laplacian and $x_1 = x_2 = \cdots = x_N$, i.e., N noisy observations of the same random variable. A paper by Chang *et al.* [2] optimizes and compares two methods: (a) averaging the observations (the y_is) and then applying a soft threshold; and (b) applying a soft threshold separately to each observation and then averaging. The main result of [2] is that the choice between these methods that gives lower mean-squared error (MSE) depends on the number of observations and the input signal-to-noise ratio (SNR). Specifically, choice (b) is superior when $N = 2$, 3, or 4, and the SNR is larger than 0 dB.

We will presently describe why we consider the supe-

riority of method (b) to be counterintuitive. Then we will explain the performance of method (b). This explanation suggests how to design vector threshold functions that will give small improvements over method (b). For notational convenience, we consider only the case of $N = 2$.

The estimates in [2] are the average of thresholds

$$\hat{x}_{\mathrm{AT}} = \tfrac{1}{2}\left(\Lambda_{\rho_{\mathrm{AT}}}(y_1) + \Lambda_{\rho_{\mathrm{AT}}}(y_2)\right)$$

and the threshold of averages

$$\hat{x}_{\mathrm{TA}} = \Lambda_{\rho_{\mathrm{TA}}}\left(\tfrac{1}{2}\left(y_1 + y_2\right)\right).$$

The latter seems unfounded because all of the information about x in the vector (y_1, y_2) is contained in the average $\frac{1}{2}(y_1+y_2)$. To understand why, consider the transformation to sum s and difference d:

$$\begin{aligned} s &= y_1 + y_2 = 2x_1 + n_1 + n_2 \\ d &= y_1 - y_2 = n_1 - n_2. \end{aligned}$$

Since n_1 and n_2 are i.i.d. Gaussian, $n_1 + n_2$ and $n_1 - n_2$ are independent; thus, d is independent of x_1. Thus, the optimal estimate under any criterion (e.g., MAP or MMSE) is a function of s alone, independent of d. However, this does not mean that optimizing over a constrained set of estimation functions of s (e.g., optimizing $\Lambda_\rho(\frac{1}{2}s)$ over ρ) will always give the best estimate.

With careful consideration of the nine regimes created by $y_i \in (-\infty, -\rho)$, $[-\rho, \rho]$, or (ρ, ∞), $i = 1, 2$, one can write $\hat{x}_{\mathrm{AT}}$ as a function of s and d. For $s > 2\rho$,

$$\hat{x}_{\mathrm{AT}} = \begin{cases} \frac{1}{2}s - \rho, & |d| < s - 2\rho; \\ \frac{1}{4}s + \frac{1}{4}|d| - \frac{1}{2}\rho, & |d| \in [s - 2\rho, s + 2\rho]; \\ \frac{1}{2}s, & |d| > s + 2\rho. \end{cases} \tag{6}$$

For $s \in [0, 2\rho]$,

$$\hat{x}_{\mathrm{AT}} = \begin{cases} 0, & |d| < 2\rho - s; \\ \frac{1}{4}s + \frac{1}{4}|d| - \frac{1}{2}\rho, & |d| \in [2\rho - s, 2\rho + s]; \\ \frac{1}{2}s, & |d| > 2\rho + s. \end{cases} \tag{7}$$

This can easily be extended for $s < 0$. Note that d in (6)–(7) need not be derived from the observations; the results are identical in distribution if d is generated randomly, independent of the observations, with the $\mathcal{N}(0, 2\sigma^2)$ distribution.

In (6)–(7), $\hat{x}_{\mathrm{AT}}$ is not a simple, threshold-like function of s. The strategy of thresholding first and then averaging uses the otherwise irrelevant value of d to randomize the choice of $\hat{x}_{\mathrm{AT}}$ so as to soften the transition of the threshold function. In particular, when ρ is chosen optimally, $\mu(s) = E[\hat{x}_{\mathrm{AT}}|s]$ can be a better match to the MMSE estimate given $\frac{1}{2}s$ (see Fig. 1) than $\Lambda_{\rho_{\mathrm{TA}}}(\frac{1}{2}s)$ for any ρ_{TA}. However, the bias reduction caused by the randomization comes with a variance $E[(\hat{x}_{\mathrm{AT}} - \mu(s))^2|s]$. The latter offsets the potential advantage of a smoother estimator.

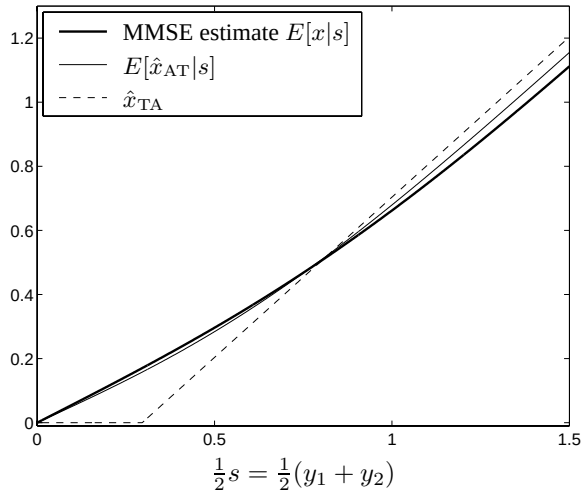

Fig. 4. Functions for estimating a Laplacian signal from two noisy observations. (Each is an odd function of s.)

Consider a unit-variance source, i.e., $\lambda = \sqrt{2}$, and noise variance $\sigma^2 = \frac{9}{16}$. According to [2, Fig. 1], these values approximately maximize the advantage of $\hat{x}_{\mathrm{AT}}$ over $\hat{x}_{\mathrm{TA}}$. By numerical search, one can determine that $\rho_{\mathrm{TA}} \approx 0.2962$ and $\rho_{\mathrm{AT}} \approx 0.3476$ are MSE-minimizing values. With these parameter choices, Fig. 4 compares the MMSE estimate given s to the estimate $\hat{x}_{\mathrm{TA}}$ and the expected value of the estimate $\hat{x}_{\mathrm{AT}}$ given s. The latter is a closer match to the MMSE estimate, but this is offset by the variance of $\hat{x}_{\mathrm{AT}}|s$ so that $\hat{x}_{\mathrm{AT}}$ is only slightly superior to $\hat{x}_{\mathrm{TA}}$. A vector threshold function can improve upon $\hat{x}_{\mathrm{AT}}$ by reducing the variance.

5. REFERENCES

[1] S. G. Chang, B. Yu, and M. Vetterli. Spatially adaptive wavelet thresholding with context modeling for image denoising. *IEEE Trans. Image Proc.*, 9(9):1522–1531, Sept. 2000.

[2] S. G. Chang, B. Yu, and M. Vetterli. Wavelet thresholding for multiple noisy image copies. *IEEE Trans. Image Proc.*, 9(9):1631–1635, Sept. 2000.

[3] R. R. Coifman and D. L. Donoho. Translation-invariant denoising. In *Wavelets and Statistics*, vol. 103 of *Springer Lecture Notes in Statistics*, pp. 125–150, New York, 1995. Springer-Verlag.

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

[5] P. Ishwar and P. Moulin. Multiple-domain image modeling and restoration. In *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, pp. 362–366, Kobe, Japan, Oct. 1999.

[6] L. Şendur and I. W. Selesnick. Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans. Signal Proc.*, 50(11):2744–2756, Nov. 2002.