A SELF-CONSISTENT WAVELET METHOD FOR DENOISING IMAGES WITH MISSING PIXELS

Thomas C. M. Lee[*]

Colorado State University
Department of Statistics
Fort Collins, CO 80523-1877, USA

Xiao-Li Meng[†]

Harvard University
Department of Statistics
Science Center, One Oxford Street
Cambridge, MA 02138-2901,USA

ABSTRACT

In this work we consider the problem of wavelet image denoising when some of the pixel values are unobserved. Our approach is to treat those unobserved pixels as missing data and adopt the self-consistency principle to *define* a "best" wavelet estimate for the true image. We propose fast and effective algorithms for computing such self-consistent wavelet estimate. The practical performance of our proposal is evaluated via a simulation study. A possible application of this work is image inpainting.

1. INTRODUCTION

Wavelet techniques have long been a popular approach for denoising images; e.g., see [2, 7, 8, 14]. Main reasons for this are that wavelet estimators enjoy excellent minimax properties and that they are capable of adapting to spatial and frequency inhomogeneities [4, 5]. In addition, they are backed up by a fast algorithm [9].

In the process of capturing an image, due to detector malfunction or some other reasons, it is not uncommon that some of the pixel values are not recorded. With the presence of such unobserved pixel values, most existing wavelet denoising techniques cannot be directly applied to recover the original images. The goal of this paper is to propose a method for handling this problem. Our approach is first to treat those unobserved pixels as missing data, and then invoke the *self-consistency* principle [12] to *define* wavelet estimators for images with missing pixels. As stated in [12], self-consistency is a fundamental concept in statistics, and is a general statistical principle for retaining as much as possible the information in the data.

A more precise statement of the problem is as follows. Let $\boldsymbol{f} = (f_1, \ldots, f_N)^T$ be the true image that we want to recover and $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ be the noisy version of $\boldsymbol{f}$ satisfying

$$y_i = f_i + e_i, \quad e_i \sim \text{iid}\,\mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, N. \qquad (1)$$

It is assumed that the number of rows and columns of $\boldsymbol{f}$ are both integer powers of 2 so that a 2D discrete wavelet transform (DWT) can be applied. The aim is, using wavelet methods, to estimate $\boldsymbol{f}$ when only a portion of the elements of $\boldsymbol{y}$ is observed.

2. A SELF-CONSISTENCY CRITERION FOR WAVELET IMAGE DENOISING WITH MISSING PIXELS

This section presents the proposed self-consistency criterion for wavelet image denoising when missing pixels are present. First we fix the notation. We partition the elements of $\boldsymbol{y}$ into two subsets: those that are observed and those that are not observed (i.e., missing). We denote, respectively, the number and the subset of the observed elements as $n < N$ and $\boldsymbol{y}_{\text{obs}}$, and write the subset of the missing elements as $\boldsymbol{y}_{\text{mis}}$. We shall call $\boldsymbol{y}$ the complete data and re-name it as $\boldsymbol{y}_{\text{com}}$; i.e., $\boldsymbol{y} = \boldsymbol{y}_{\text{com}} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}}\}$. In addition, we define $\boldsymbol{I}_{\text{obs}}$ as the "observed data index set": $i \in \boldsymbol{I}_{\text{obs}} \Leftrightarrow y_i$ is observed.

Suppose that when the complete data $\boldsymbol{y}_{\text{com}}$ is available, we have a method for computing the "best" estimate $\hat{\boldsymbol{f}}_{\text{com}}$ for the true image $\boldsymbol{f}$. Now, given only the observed data $\boldsymbol{y}_{\text{obs}}$, we propose estimating $\boldsymbol{f}$ with the solution $\hat{\boldsymbol{f}}_{\text{obs}}$ that solves following self-consistent equation:

$$E\left[\hat{\boldsymbol{f}}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}_{\text{obs}}\right] = \hat{\boldsymbol{f}}_{\text{obs}}(\cdot). \qquad (2)$$

Observe that definition (2) is not restricted to wavelet image denoising, nor to any Gaussian noise assumptions.

The use of the self-consistent equation (2) for the present problem is motivated by the fact that many previous applications of the self-consistency principle often led to the most efficient estimating procedures. Examples include, in the parametric setting, maximum likelihood estimation via the EM algorithm [3] and, in the nonparametric setting, the Kaplan-Meier estimator [6]. In view of this, it is expected that the proposed wavelet estimator $\hat{\boldsymbol{f}}_{\text{obs}}$, the solution to (2), would possess excellent statistical properties.

3. THREE ALGORITHMS

The self-consistent estimator $\hat{\boldsymbol{f}}_{\text{obs}}$ would not be of much practical value if (2) could not be solved with reasonable computational effort. To solve (2), two steps are involved. The first is to carry out the conditional expectation on the left-hand side, and the second is to solve the equation, in analogous to the E-step and M-step of the EM algorithm, respectively. However, unlike many common EM applications where the E-step is in closed form, in the wavelet applications, the exact E-step is typically analytically infeasible. It is because, due to the shrinkage operation, $\hat{\boldsymbol{f}}_{\text{com}}$ is a highly complicated non-linear function of the missing data $\boldsymbol{y}_{\text{mis}}$.

[*] This work was partially supported by the National Science Foundation under Grant No. 0203901.

[†] This work was partially supported by the National Science Foundation under Grant No. 0204552.

There are two general approaches for dealing with such a problem. The first is to use Monte Carlo E-step, as in [10, 13], and the second is to trade the exactness for simplicity by making certain approximations to the conditional expectation. In below we propose three algorithms for computing $\hat{\boldsymbol{f}}_{\text{obs}}$, one is based on the first Monte Carlo approach, while the remaining two follow the second approximation approach.

3.1. A Multiple Imputation Self-Consistent (MISC) Algorithm

Our first algorithm assumes that a complete-data wavelet denoising procedure has been chosen. Starting with initial estimates $\hat{\boldsymbol{f}}^{(0)}$ (for $\boldsymbol{f}$) and $\hat{\sigma}^{(0)}$ (for the noise standard deviation σ), the algorithm iterates the following three steps for $t = 1, \ldots$:

Step 1 For $\ell = 1, \ldots, m$, simulate $\boldsymbol{y}^{\ell}_{\text{mis}}$ independently from

$$P(\boldsymbol{y}_{\text{mis}} | \boldsymbol{y}_{\text{obs}}; \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}, \sigma = \hat{\sigma}^{(t-1)}).$$

Step 2 For $\ell = 1, \ldots, m$, apply the chosen complete-data wavelet denoising procedure to the *pseudo completed data* $\boldsymbol{y}^{\ell} = \{\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}^{\ell}_{\text{mis}}\}$ and obtain $\hat{f}^{\ell}_i$, $i = 1, \ldots, N$.

Step 3 Compute the t-th iterative estimate of $\boldsymbol{f}$ as

$$\hat{f}^{(t)}_i = \frac{1}{m} \sum_{\ell=1}^{m} \hat{f}^{\ell}_i, \quad i = 1, \ldots, N. \tag{3}$$

Also, use the residuals $\{y_i - \hat{f}^{(t)}_i : i \in \boldsymbol{I}_{\text{obs}}\}$ to obtain an efficient estimate $\hat{\sigma}^{(t)}$ for σ.

Throughout the whole paper, a constant image with greyvalue equal to the average of $\boldsymbol{y}_{\text{obs}}$ is taken as the initial estimate $\hat{\boldsymbol{f}}^{(0)}$, while $\hat{\sigma}^{(0)}$ is set to the standardized sum of the squared pixelwise differences between $\hat{\boldsymbol{f}}^{(0)}$ and $\boldsymbol{y}_{\text{obs}}$.

In the statistics literature, the repeated simulation of $\boldsymbol{y}^{\ell}_{\text{mis}}$ in Step 1 above is known as *multiple imputation* [11], and hence the name MISC. In this algorithm the larger the m, the better results one would expect, but at the expense of increased computational time. Our numerical experience seems to suggest that, as long as m is larger than a minimum cutoff, the additional improvement on $\hat{\boldsymbol{f}}$ computed with a larger m is not largely significant. Our numerical experience suggests that a conservative cutoff is $m = 100$, although $m = 10$ is also often sufficient.

It is evident that the above is a generic algorithm, in the sense that it is not restricted by the specific form of the complete-data wavelet denoising procedure. On one hand, this is a great advantage as it is extremely flexible and the additional programming, relative to that for the complete-data procedure, is minimal as long as it is easy to draw from the conditional distribution in the first step, which typically is the case for independent Gaussian errors. It also provides a benchmark and basis for developing more specialized and sophisticated algorithms. On the other hand, it is a "brute force" algorithm, and is thus quite inefficient as a numerical algorithm. Thus it presents the need for faster algorithms.

3.2. A Simple Approximated Algorithm

To construct a fast algorithm for computing $\hat{\boldsymbol{f}}_{\text{obs}}$, we consider replacing the costly multiple imputation computation in the MISC algorithm by a simple analytical approximation. We label the resulting algorithm as the *simple algorithm*, and it is designed for a specific type of denoising methods, namely, when the thresholding value is a known function $g(\hat{\sigma})$ of $\hat{\sigma}$, where $\hat{\sigma}$ is an estimate of σ. A classical example for $g(\hat{\sigma})$ is the universal thresholding scheme of [4], for which $g(\hat{\sigma}) = \hat{\sigma}\sqrt{2 \log N}$.

Starting with $\hat{\boldsymbol{f}}^{(0)}$ and $\hat{\sigma}^{(0)}$, this simple algorithm iterates, at the t-th iteration, the following steps:

Step 1 For each i such that $i \notin \boldsymbol{I}_{\text{obs}}$, impute the corresponding missing y_i by $y^{(t)}_i = \hat{f}^{(t-1)}_i$. Thus this creates the complete data $\boldsymbol{y}^{(t)} = \{y_i : i \in \boldsymbol{I}_{\text{obs}}\} \cup \{y^{(t)}_i : i \notin \boldsymbol{I}_{\text{obs}}\}$.

Step 2 Apply a DWT to $\boldsymbol{y}^{(t)}$ to obtain the empirical wavelet coefficients $\boldsymbol{w}^{(t)} = \boldsymbol{W}\boldsymbol{y}^{(t)}$, where $\boldsymbol{W}$ is the 2D DWT matrix.

Step 3 Obtain a robust estimate $\tilde{\sigma}^{(t)}$ of σ from $\boldsymbol{w}^{(t)}$, for example, the median absolute deviation method used by [4]. We call $\tilde{\sigma}^{(t)}$ the *unadjusted* estimate for σ.

Step 4 Use the following *variance inflation formula* to obtain an *adjusted* estimate $\hat{\sigma}^{(t)}$ for σ:

$$\hat{\sigma}^{(t)} = \sqrt{[\tilde{\sigma}^{(t)}]^2 + C_m[\hat{\sigma}^{(t-1)}]^2}, \tag{4}$$

where $C_m = 1 - \frac{n}{N}$ is the fraction of missing observations. (Recall that n is the number of observed pixel values.)

Step 5 Compute $\hat{\boldsymbol{w}}^{(t)}$ by thresholding $\boldsymbol{w}^{(t)}$ with the thresholding value $g(\hat{\sigma}^{(t)})$.

Step 6 Apply the inverse DWT to $\hat{\boldsymbol{w}}^{(t)}$ and obtain the t-th iterative estimate $\hat{\boldsymbol{f}}^{(t)} = \boldsymbol{W}^T \hat{\boldsymbol{w}}^{(t)}$.

Convergence can be declared if $|\hat{\sigma}^{(t)} - \hat{\sigma}^{(t-1)}|/\hat{\sigma}^{(t)} < \epsilon$. Upon convergence, estimates of $\boldsymbol{f}$, as well as σ, will be obtained.

It is obvious that computationally this simple algorithm is much less intensive than the MISC algorithm. It is because within each iteration there is only one complete-data wavelet shrinkage computation, in contrast to m sets of such computation with the MISC algorithm.

A key component of this algorithm is the variance inflation formula (4), which takes into account the effect of those imputed $y^{(t)}_i$'s on the estimation of σ^2. Extensive numerical experiments suggest that, not carrying out this variance inflation adjustment would lead to an underestimation of σ^2, which would in turn lead to noticeably poorer wavelet estimators. However, this variance inflation adjustment does not accounts for all the uncertainty in the thresholding due to missing data, and hence it does not work well when the percentage of missing pixels is large.

This variance inflation formula was derived with the following approximation. To simplify presentation, for the rest of this paper we will use single-indexing w_l instead of the usual double-indexing w_{jk} notation to denote a wavelet coefficient. At the t-th step when we calculate

$$\hat{\boldsymbol{f}}^{(t)} = E\left[\hat{\boldsymbol{f}}_{\text{com}} \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right],$$

we pretend that the correct conditional expectation

$$\tilde{w}^{(t)}_l \equiv E\left[1_{|w_l| \geq g(\tilde{\sigma})} w_l \middle| \boldsymbol{y}_{\text{obs}}, \boldsymbol{f} = \hat{\boldsymbol{f}}^{(t-1)}\right], \tag{5}$$

where w_l's and $\tilde{\sigma}$ are respectively the complete-data empirical wavelet coefficients and variance estimate, can be approximated by thresholding the conditional expectation of w_l with $g(\hat{\sigma})$ using the adjusted $\hat{\sigma}$. This approximation performed very well in our simulation experiments when the percentage of missing data is small (say less than 30%). However, this approximation will fail when the missing percentage is large.

3.3. A Refined Fast Algorithm

In view of the problem with the above approximation, a more refined approximation is derived. This new approximation is obtained by pretending that we know σ, or more precisely, by ignoring the conditional variability in $\tilde{\sigma}$ when calculating the E-step. Under this situation, it can be shown that (5) becomes

$$\tilde{w}_l^{(t)} = \alpha(w_l^{(t)}, \eta_l) + \beta(w_l^{(t)}, \eta_l) \times w_l^{(t)}, \tag{6}$$

where the α and β functions are given by

$$\alpha(w,\eta) = \frac{\eta\sigma}{\sqrt{2\pi}} \left\{ e^{-\frac{1}{2}\left(\frac{c+w}{\eta\sigma}\right)^2} - e^{-\frac{1}{2}\left(\frac{c-w}{\eta\sigma}\right)^2} \right\}, \tag{7}$$

$$\beta(w,\eta) = 2 - \Phi\left(\frac{c-w}{\eta\sigma}\right) - \Phi\left(\frac{c+w}{\eta\sigma}\right), \tag{8}$$

with Φ being the cumulative distribution function of $\mathcal{N}(0,1)$. In (6), the $w_l^{(t)}$ is the empirical wavelet coefficient obtained from Step 2 of the simple algorithm, and $\eta_l^2\sigma^2$ is its *conditional* variance given the observed data (and the known σ^2). Note that η_l can be easily computed at the outset of iteration as the lth diagonal element of $I - \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$, where $\boldsymbol{R}$ is an $N \times N$ matrix whose off-diagonal elements are all zero, and whose lth diagonal elements is one if y_l is observed and zero otherwise. Thus, one can treat that the diagonal of $\boldsymbol{R}$ is a "response indicator" vector. The quantity η_l is a measure of the percentage of missing information in w_l due to the missing data. Notice that $0 \leq \eta_l \leq 1$, and η_l is one when there is no information in the observed data about w_l and zero when w_l is fully observed. For simplicity, in our practical calculation for $\tilde{w}_l^{(t)}$ we use the approximation $\eta_l \approx 1 - \frac{n}{N}$ for all l.

The resulting algorithm is identical to the simple algorithm except that we replace its Step 5 by (6) to (8), where we use $\sigma = \hat{\sigma}^{(t)}$ and $c = g(\hat{\sigma}^{(t)})$. Notice that the thresholding is now performed via an almost exact (exact when σ is known) closed-form E-step calculation. Thus, computationally, this new and refined algorithm is almost as fast as the simple algorithm, and it is also straightforward to program as only standard functions are involved in (6) to (8). In addition, as it provides a much more refined E-step, the statistical efficiency of the resulting estimator is expected to be much closer to that of the MISC estimator with infinite number of imputation. An intriguing insight suggested by (6) is that even when we choose to use hard thresholding with complete data, we should use "soft" thresholding with incomplete data, as (6) is a form of soft thresholding as long as $\eta_l > 0$; note in particularly $0 < \beta(w,\eta) < 1$ when $\eta > 0$. When $\eta \to 0$, $\alpha(w,\eta) \to 0$ and $\beta(w,\eta) \to 1_{|w| \geq c}$, and thus (6) goes back to the original hard-thresholding, as it should be.

4. SIMULATION STUDY

A simulation study was conducted to evaluate the practical performances of the MISC (Section 3.1, with $m = 10$), the simple (Section 3.2) and the refined (Section 3.3) algorithms. In this study two testing images of size 256 × 256 were used: the Lena image displayed in Figure 1(a) and the Airplane image displayed in Figure 1(b). Also, two snrs and three missing data percentages were tested: snr $= (5, 7)$ and $C_m = (10\%, 30\%, 50\%)$. Lastly, two missing data formation mechanisms were tested. The first one is missing at random, in which missing pixel locations were randomly selected from the image, while in the second mechanisms the missing pixels were clustered together.

For each of the above experimental factor combinations, 100 noisy images were generated, and the three algorithms presented in Section 3 were applied to reconstruct the corresponding true images. The following adjusted universal thresholding value was used: $\hat{\sigma}\sqrt{2\log N - \log(1 + 256\log N)}$. In [1] it is shown that this adjusted thresholding value is superior to the classical universal thresholding value $\hat{\sigma}\sqrt{2\log N}$ of [4]. As to provide a benchmark for comparison, for each noisy image, we also applied the universal denoising method [4] (with the same adjusted thresholding value) to reconstruct the corresponding true image *using the complete data* $\boldsymbol{y}_{\text{com}}$. In below we refer this method as UniComp. As UniComp has the full information from $\boldsymbol{y}_{\text{com}}$, it is expected that it would produce better reconstructed images than the other three algorithms.

For every reconstructed images, we calculated

$$\text{MSE}_{\text{obs}} = \frac{1}{n} \sum_{i \in \boldsymbol{I}_{\text{obs}}} \{f_i - \hat{f}_i\}^2$$

as a measure of reconstruction quality for the observed pixels. Similar values for the missing pixels (MSE$_{\text{mis}}$, sum over $i \notin \boldsymbol{I}_{\text{obs}}$) and the complete data (MSE$_{\text{com}}$, sum over $i = 1, \ldots, N$) were also calculated. In addition we also computed the following MSE ratio:

$$r_{\text{com}}(\text{MISC}) = \frac{\text{MSE}_{\text{com}} \text{ of MISC}}{\text{MSE}_{\text{com}} \text{ of UniComp}}.$$

Similar MSE ratios for the observed (r_{obs}(MISC)) and missing data (r_{mis}(MISC)), and for the simple and refined algorithms (r_{com}(simple), r_{obs}(simple), r_{mis}(simple), r_{com}(refined), r_{obs}(refined) and r_{mis}(refined)) were also calculated. Since UniComp reconstructed the images with the complete data, it is expected that all these MSE ratios are bigger than 1. For snr $= 7$ and $C_m = 30\%$, boxplots of these MSE ratios are given in Figure 2. Boxplots for other experimental settings are similar and hence omitted. From Figure 2 some major empirical conclusions can be obtained. First, as all r_{obs}(MISC) values are fairly close to 1, the easy-to-implement benchmark MISC algorithm performs reasonably well for those observed pixels. Secondly, the refined algorithm is superior to the other two algorithms, as it does not require multiple imputation (as opposed to MISC) and it uses a better approximation than the simple algorithm. Lastly, a surprising (and unexpected) observation is that, r_{obs}(refined) is in fact less than 1 when the locations of the missing data are clustered together.

For the purpose of visual inspection, two degraded versions of Lena are displayed in Figures 3(a) and 4(a). Those black pixels represent the locations of the missing values. The snr is 7 and the missing percentage is 10%. Figures 3(b) and 4(b) display the corresponding reconstructed images obtained from the refined algorithm.

5. CONCLUSIONS AND FUTURE WORK

In this paper we presented a self-consistency criterion for wavelet image denoising when missing pixels are present. We proposed fast algorithms for computing our image estimates. Results from a simulation study suggest that the refined algorithm described in Section 3.3 is the preferred algorithm. Important extensions of this work include the development of fast algorithms for other more sophisticated wavelet image denoising techniques (e.g., [7, 14]) and the application of the self-consistency principle to image classification with missing pixels. Another important extension is wavelet regression for 1D signal estimation with irregularly-spaced data.

(a) Lena

(b) Airplane

Fig. 1. Images used in the simulation study.

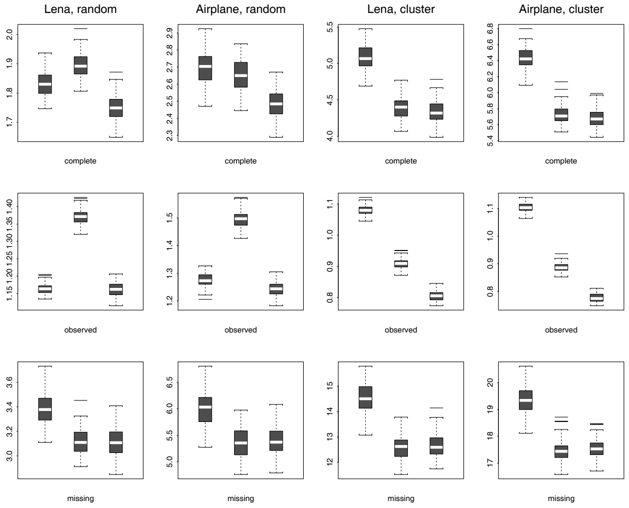
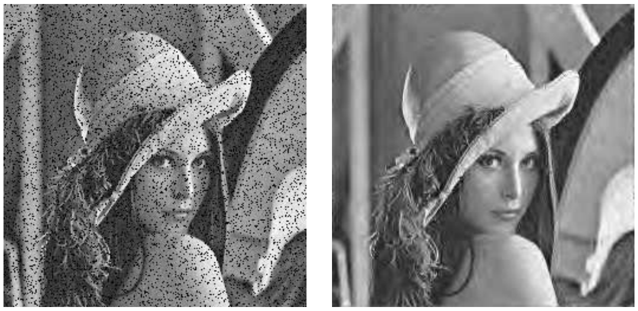


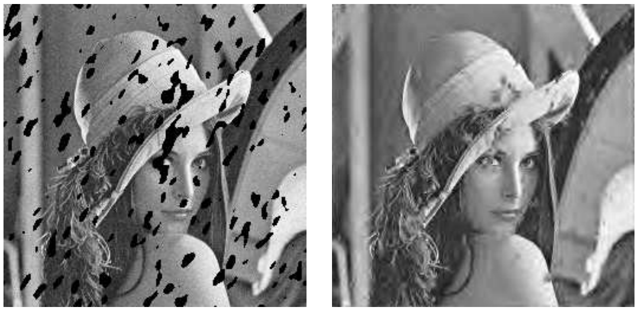
Fig. 2. Boxplots of the MSE ratios resulted from the simulation study in Section 4. In each panel the left, middle and right boxplots correspond, respectively, to the MISC, simple and refined algorithm.



(a) Degraded Lena

(b) Reconstructed Lena

Fig. 3. Degraded (a) and reconstructed (b) Lena when the pixels are missing at random.



(a) Degraded Lena

(b) Reconstructed Lena

Fig. 4. Similar to Figure 3 but for clustered missing pixels.

6. REFERENCES

[1] A. Antoniadis and J. Fan. Regularization of wavelet approximations (with discussion). *Journal of the American Statistical Association*, 96:939–967, 2001.

[2] M. Belge, M. E. Kilmer, and E. L. Miller. Wavelet domain image restoration with adaptive edge–preserving regularization. *IEEE Transactions on Image Processing*, 9:597–608, 2000.

[3] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

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

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

[6] B. Efron. The two sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume IV, pages 831–851. Berkeley: University of California Press, 1967.

[7] M. Hansen and B. Yu. Wavelet thresholding via MDL for natural images. *IEEE Transactions on Information Theory*, 46:1778–1788, 2000.

[8] M. Jansen and A. Bultheel. Multiple wavelet threshold estimation by generalized cross validation for images with correlated noise. *IEEE Transactions on Image Processing*, 8:947–953, 1999.

[9] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[10] X.-L. Meng and S. Schilling. Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, 91:1254–1267, 1996.

[11] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York, 1987.

[12] T. Tarpey and B. Flury. Self–Consistency: A fundamental concept in statistics. *Statistical Science*, 11:229–243, 1996.

[13] G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704, 1990.

[14] J. Xie, D. Zhang, and W. Xu. Spatially adaptive wavelet denoising using the minimum description length principle. *IEEE Transactions on Image Processing*, 13:179–187, 2004.