A Helicopter View of The Self-Consistency Framework for Wavelets and Other Signal Extraction Methods In the Presence of Missing and Irregularly Spaced Data

Xiao-Li Meng
Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

ABSTRACT

A common frustration in signal processing and, more generally, information recovery is the presence of irregularities in the data. At best, the standard software or methods will no longer be directly applicable when data are missing, incomplete or irregularly spaced (e.g., as with wavelets). Self-consistency is a very general and powerful statistical principle for dealing with such problems. Conceptually it is extremely appealing, for it is essentially a mathematical formalization of iterating common-sense "trial-and-error" methods until no more improvement is possible. Mathematically it is elegant, with one fixed-point equation to solve and a general projection theorem to establish optimality. Practically it is straightforward to program because it directly uses the regular/complete-data method for iteration. Its major disadvantage is that it can be computationally intensive. However, increasingly efficient (approximate) implementations are being discovered, such as for wavelet de-noising with hard and soft thresholding. This brief overview summarizes the author's keynote presentation on those points, based on joint work with Thomas Lee on wavelet applications and with Zhan Li on the theoretical properties of the self-consistent estimators.

Keywords: EM Algorithm, information recovery, Iterative algorithm, Multiple imputation, Non-parametric regression, Semi-parametric regression, Signal Processing

1. A MISSING-DATA PERSPECTIVE

In real-life statistical analysis, data are almost never complete. Putting it differently, if one is presented with a "clean and complete" data set, it would be safe to assume that someone has manipulated the original raw data to make it so. To a non-statistician, it might seem to be rather odd to worry about this "cleaning process", as many investigators would be more than happy if someone else has dealt with their "dirty data", making them easier to use, for example, ready for a standard software. But for those of us who make a living by dealing with the collection and analysis of data, we all understand the importance of knowing how the data were collected, and the potentially serious distortion of information the cleaning process can cause if the "data cleaner" is not professionally trained. For example, common methods of imputing the missing values by some "best fit", such as sample averages or least-squares estimates, typically lead to serious under-estimation of noise level if a user is unaware of this cleaning step or fails to take it into account in his/her analysis.[1]

The abundance and seriousness of missing data problems has been a driving force behind two lines of statistical research, collectively known as the "missing-data perspective" or "incomplete-data perspective".[2] One line, essentially started with the seminal work of Donald Rubin in 1976,[3] is concerned with understanding when and how a missing-data mechanism (i.e., the process that prevented us from observing the intended complete data) distorts the final analysis if it is not dealt with properly. The other line is focussed on developing missing-data methods, both statistical and computational, that would allow general practitioners to deal with missing data problems by using essentially *only the complete-data methods and software*. Two of the most significant examples are the EM algorithm (and its generalizations)[4,5] and Multiple Imputation.[6]

A somewhat unexpected but exceedingly fruitful byproduct of the second line of research is that many hard complete-data problems can be solved by viewing them as an incomplete-data problem for a larger "augmented" complete-data problem.[1,4] For example, an irregularly-spaced data set can be viewed as a regularly-spaced data

Further author information: E-mail: meng@stat.harvard.edu

Wavelets XII, edited by Dimitri Van De Ville, Vivek K. Goyal, Manos Papadakis,
Proc. of SPIE Vol. 6701, 670124, (2007) · 0277-786X/07/$18 · doi: 10.1117/12.735291

set but with missing values. This observation has enabled us[7,8] to develop wavelet methods with irregularly-spaced data sets, as reviewed below. Although purposely making a problem larger (e.g., increasing dimension) seems to be rather counter-productive, the central idea underlying it, known as the method of *auxiliary variables* in statistical physics[9] or *data augmentation*[10–12] in statistics, has been an exceedingly fruitful technique for efficient implementation of Markov chain Monte Carlo.[13]

Self-consistency is an old concept in statistics[14] for dealing with missing data. It can be viewed as an early seed for much of the later work in the second line of research and has been successfully used in a number of contexts, the most famous of which are the Kaplan-Meier estimator[15] (which actually stimulated the concept of self-consistency[14]) and the EM algorithm itself.[4] A more recent development was given by Tarpey and Flury,[16] who emphasized that self-consistency is "a fundamental concept in statistics". The purpose of this brief overview is to introduce the concept and method to some signal processing problems, including wavelet regression, and more generally semi-parametric and non-parametric regression. The overview is based on joint work on wavelet applications[7,8] with my collaborator Thomas Lee of Colorado State University and Chinese University of Hong Kong, and on-going theoretical work with my Ph.D. student Zhan Li at Harvard University. A full paper on setting up a very general "Self-consistent Projection" framework for parametric, semi-parametric, and non-parametric inference will be available shortly.

2. DEFINING SELF-CONSISTENT ESTIMATORS IN L^2 SPACE

Self-consistent estimators can be defined in any space, but for simplicity of presentation, in this overview I will focus on the L^2 space. Indeed, this is the most commonly used space in statistics, and in general, because of the ubiquitous use of variance and mean-squared errors in practice. The mathematical theory for L^2 space is also the simplest because it admits orthogonal projection. This is particularly relevant for a self-consistent estimator because it is defined, as we shall see below, as a solution of a fixed-point equation based on a projection.

2.1 An Illustration via Linear Regression

Consider the simplest linear regression setting, where the signal is represented by βx with x being a known scalar quantity and β to be estimated from the noisy data, which are linked to the signal via the usual regression model:

$$y_i = \beta x_i + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where e_i's are identically and independently distributed noises with mean zero and variance σ^2. The least-squares estimator of β is given by

$$\hat{\beta}_n \equiv \hat{\beta}_n(y_1, \ldots, y_n) = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}. \tag{2}$$

Imagine that a Mr. Littlestat was asked by his boss to fit a least-squares line to a dataset with $n = 13$, but he had no knowledge of the least-squares method, so formula (2) was not available to him. Instead, he had access to a "black-box" program that would compute (2) but only when $n = 2^4$ (hypothetical for this illustration of course, but recall the standard requirements on the data configuration for fast wavelet computation). Is it possible for him to use this black-box to compute what he needs, namely, $\hat{\beta}_{13}$?

At first sight, this seems to be a hopeless task, since the program simply would not run with 13 data points. Out of desperation, Mr. Littlestat added 3 arbitrary data points. That is, he arbitrarily picked up three 3 new x values – let's label them $\{x_{14}, x_{15}, x_{16}\}$, and then three corresponding y values – let's label them $\{y_{14}^{(0)}, y_{15}^{(0)}, y_{16}^{(0)}\}$. The program then ran and outputted an estimator of β, which we label as $\hat{\beta}_{13}^{(1)}$. Of course Mr. Littlestat knew that this $\hat{\beta}_{13}^{(1)}$ cannot be the $\hat{\beta}_{13}$ he was after. However, the output allowed him to draw the least-squares line, i.e., $y = \hat{\beta}_{13}^{(1)} x$. This gave him an idea: why not adjust the three arbitrarily added data points, which are off the line he had just drawn, to the line? That is, use $\{(x_i, y_i^{(1)}); i = 14, 15, 16\}$, where $y_i^{(1)} = \hat{\beta}_{13}^{(1)} x_i$, instead of the original three arbitrary points $\{(x_i, y_i^{(0)}); i = 14, 15, 16\}$? After all, these are fake data points to start with, and by adjusting them to the line, his gut feeling was that the result should be closer to what he was looking for.

He then ran the "black box" again, with this new updated 16-point data set. It gave a new estimate for β, $\hat{\beta}_{13}^{(2)}$, which allowed him to draw a new line $y = \hat{\beta}_{13}^{(2)} x$. He then observed that the updated imputed points $\{(x_i, y_i^{(1)}); i = 14, 15, 16\}$ are, again, off the new line, but they are much closer to the line than the previous points $\{(x_i, y_i^{(0)}); i = 14, 15, 16\}$ had been to the previous line, $y = \hat{\beta}_{13}^{(1)} x$.

Encouraged, Mt. Littlestat repeated the above process several times, until the line stopped moving as far as he could visually tell. He certainly did not have enough knowledge to know whether this "stopped" line is actually what he was looking for, but his intuition told him that it must be something "good" because it cannot be further improved upon given his resource, namely, the black-box for computing $\hat{\beta}_{16}$.

Indeed, what Mr. Littlestat obtained from his "trial-and-error" approach is the correct answer. This is because what he did can be expressed mathematically as an iteration of the form

$$\hat{\beta}_{13}^{(t+1)} = \hat{\beta}_{16}(y_1, \ldots, y_{13}, y_{14}^{(t)}, y_{15}^{(t)}, y_{16}^{(t)}), \tag{3}$$

where the $\hat{\beta}_{16}$ function is given by (2) as a function of $\{y_1, \ldots, y_{16}\}$, t indexes iteration, and $y_i^{(t)} = \hat{\beta}_{13}^{(t)} x_i$ ($i = 14, 15, 16$). Because of (2), the limit of (3), denoted by $\hat{\beta}_{13}$, must satisfy

$$\hat{\beta}_{13} = \frac{\sum_{i=1}^{13} y_i x_i + \hat{\beta}_{13} \sum_{i=14}^{16} x_i^2}{\sum_{i=1}^{16} x_i^2}.$$

But this is the same as $\hat{\beta}_{13} = \sum_{i=1}^{13} y_i x_i / (\sum_{i=1}^{13} x_i^2)$, the least-squares estimate with 13 data points.

This, of course, is too remarkable to be just a mathematical coincidence. Rather, it is a consequence of the self-consistency property of the least-squares estimator. Specifically, $\hat{\beta}_N$ has the following property: for any subset of n data points, where $n < N$, and, without loss of generality we assume it is the first n points $\{y_1, \ldots, y_n\}$, and as long as $\sum_{i=n+1}^{N} x_i^2 > 0$ (recall all design points x_i's are assumed to be deterministic in the usual regression setting), then

$$E\left[\hat{\beta}_N \middle| y_1, \ldots, y_n; \beta = \hat{\beta}_n\right] = \hat{\beta}_n. \tag{4}$$

Here $E[g(\boldsymbol{y}_{\rm com})|\boldsymbol{y}_{\rm obs}; \beta]$ denotes the expectation of $g(\boldsymbol{y}_{\rm com})$, a function of the *complete data*, $\boldsymbol{y}_{\rm com} = \{y_1, \ldots, y_N\}$, with respect to the *conditional distribution* $p(\boldsymbol{y}_{\rm com}|\boldsymbol{y}_{\rm obs}; \beta)$, where $\boldsymbol{y}_{\rm obs} = \{y_1, \ldots, y_n\}$ represents the observed data. (For simplicity, we suppress the dependence of this conditional distribution on the nuisance parameter σ^2, which is irrelevant for computing the conditional expectation of $\hat{\beta}_N$.) That is, the best estimator $\hat{\beta}_n$ is the one that is "self-consistent" in the sense that given $\beta = \hat{\beta}_n$, the projection of $\hat{\beta}_N$ will be exactly the same as $\hat{\beta}_n$, i.e., no more improvement is possible.

And the fact that the least-squares estimator (2) satisfies the self-consistency equation (4) itself is a consequence of a more general result on the self-consistency of parametric maximal likelihood estimators (MLE), stated in Section 2.4 below.

2.2 A General Definition Under L^2 Norm

Consider a more general regression setting, where the noisy observations, y_i's, are linked to the signal f (a regression function) via

$$y_i = f(x_i) + e_i, \quad i = 1, 2, ... \tag{5}$$

where x_i's are the *design points* (e.g., location of a pixel) and e_i's are the error terms. Note that here we do not need to assume that the errors are signal independent nor that they are uncorrelated with each other, and nor that they are homogenous. The self-consistency framework is applicable in general, since it relies on the complete-data method to handle all these complications.

Now suppose we have a method for estimating/reconstructing the signal f when the data are complete/regular; e.g., in a wavelet setting with $x_i = i/N$ and $N = 2^k$, where k is some positive integer. Denote this estimator by $\hat{f}_{\rm com}$. But our observations are incomplete or irregularly spaced, that is, $\boldsymbol{y}_{\rm obs} = \{y_i; i = 1, \ldots, n\}$ is a subset of the (possibly imaginary) ideal $\boldsymbol{y}_{\rm com} = \{y_i; i = 1, \ldots, N\}$, where $N \geq n$. Similarly to (4), the self-consistent

estimator $\hat{f}_{\text{obs}}$, our estimate of f given $\boldsymbol{y}_{\text{obs}}$, is defined to be the solution of the following fixed-point equation (note we always condition on x_i's):

$$E\left[\hat{f}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}}; f=\hat{f}_{\text{obs}}\right]=\hat{f}_{\text{obs}}(\cdot). \tag{6}$$

Note that the conditional expectation E with respect to $p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}}, f)$ is the projection operator under the mean integrated squared loss:

$$||\hat{f}_{\text{com}}-f||=\left\{E\left[\int_x(\hat{f}_{\text{com}}(x)-f(x))^2dx\right]\right\}^{1/2}\equiv\left\{\int_{\boldsymbol{y}_{\text{com}}}\int_x(\hat{f}_{\text{com}}(x)-f(x))^2p(\boldsymbol{y}_{\text{com}}|f)dx\ d\boldsymbol{y}_{\text{com}}\right\}^{1/2}, \tag{7}$$

where $p(\boldsymbol{y}_{\text{com}}|f)$ is the density function of the complete data under our model specification. This is because $E\left[\hat{f}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}},;f\right]$ is the projection of $\hat{f}_{\text{com}}$ onto a subspace in L^2, denoted by $\mathcal{G}_f=\{g_{\text{obs},f}(\cdot)\}$, which consists of all L^2 integrable functions that only depend on $\boldsymbol{y}_{\text{obs}}$ and the *given* f. That is, the projection is a $g\in\mathcal{G}_f$ that minimizes the conditional norm:

$$\left\{\int_{\boldsymbol{y}_{\text{com}}}\int_x(\hat{f}_{\text{com}}(x)-g(x))^2p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}};f)dx\ d\boldsymbol{y}_{\text{com}}\right\}^{1/2}. \tag{8}$$

The importance of recognizing this projection property is two-fold. First, it provides a geometric insight on why self-consistency is a good principle. It essentially provides the best solution to the "chicken or egg" problem: if we knew the signal f, then the best prediction for the unobserved $\hat{f}_{\text{com}}$ is its projection onto the observed data space under $p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}};f)$. On the other hand, if we knew this projection, then of course it is our best estimate for f given the method for $\hat{f}_{\text{com}}$ and our observed data. But which one comes first? Chicken or egg? The answer is that they come at the same time, in the sense that we can "try and err" until the two are the same. That is, the projection is the same as our estimate based on the observed data, exactly as Mr. Littlestat did with his 13 data points. In this sense, the self-consistency principle is simply a mathematical formulation of the common-sense "trial-and-error" approach.

Second, the recognition of the projection property immediately allows us to generalize beyond the L^2 norm. For example, it is not uncommon to consider the L^1 norm or the L^0 norm in wavelet settings or other sparse representation settings. The self-consistency fixed-point equation will then change accordingly, because the projection is the (conditional) median under L^1, and is the (conditional) mode under L^0, in contrast to the (conditional) mean under L^2. That is, we will replace the conditional expectation operator on the left hand side of (6) by the conditional median operator and conditional mode operator, respectively for L^1 and L^0. The corresponding theoretical investigation of the resulting solution is a bit more involved because the general L^p space and its projection operator is not as easy to handle as when $p=2$. But conceptually there is no difficulty in using any norm – as long as the corresponding projection under $p(\boldsymbol{y}_{\text{com}}|\boldsymbol{y}_{\text{obs}};f)$ can be calculated.

2.3 An Optimality Result Under L^2

Let $M(f;\boldsymbol{y}_{\text{com}})=E\left[\hat{f}_{\text{com}}(\cdot)\middle|\boldsymbol{y}_{\text{obs}},f\right]$ be the projection of $\hat{f}_{\text{com}}$ under the L^2 norm as defined in (7), and let $M(\hat{f})\equiv M(f=\hat{f},\boldsymbol{y}_{\text{obs}})$ be the corresponding induced mapping from $\mathcal{F}_{\text{obs}}=$\{all L^2-integrable estimators of f based on $\boldsymbol{y}_{\text{obs}}$\} into itself. Then we have the following result.

THEOREM 2.1. *Suppose the mapping M satisfies the following contraction mapping property:*

$$||M(\hat{f}_1)-M(\hat{f}_2)||\le\delta||\hat{f}_1-\hat{f}_2||,\ \text{for some constant}\ \delta\in(0,1). \tag{9}$$

Then

(I) There exists a unique (with respect to the L^2 norm) solution to $||M(\hat{f}_{\text{obs}})-\hat{f}_{\text{obs}}||=0$; and

(II) The error norm of the solution $\hat{f}_{\rm obs}$ must satisfy

$$||\hat{f}_{\rm obs} - f|| \leq \frac{||\hat{f}_{\rm com} - f||}{1-\delta}. \tag{10}$$

The proof of this theorem is quite simple. Result (I) is a direct consequence of the well-known contraction mapping theorem,[17] since $\mathcal{F}_{\rm obs}$ is a complete metric space. Result (II) is easily established by noting that

$$||\hat{f}_{\rm obs} - f|| \leq ||M(\hat{f}_{\rm obs}) - M(f)|| + ||M(f) - f||, \tag{11}$$

and that $||M(\hat{f}_{\rm obs}) - M(f)|| \leq \delta||\hat{f}_{\rm obs} - f||$ because of (9) and $||M(f) - f|| \leq ||\hat{f}_{\rm com} - f||$ because $M(f)$ is an orthogonal projection of $\hat{f}_{\rm com}$ under L^2.

The simplicity and elegance of the proof of this theorem is an indication of the generality of the self-consistency approach. Clearly the result is not restricted to wavelet regression or any particular way of estimating/constrcuting $\hat{f}_{\rm com}$. However, for wavelet hard or soft thresholding, it can be shown that, asymptotically, $\delta = \sqrt{1 - \frac{n}{N}}$, when the observed n data points are a random sample of the N (imaginary) complete data points. This makes intuitive sense, because $r = 1 - \frac{n}{N}$ is the missing percentage; the larger r is, the larger the error $||\hat{f}_{\rm obs} - f||$ should be. The significance of (10) is that it ensures that the self-consistent estimator keeps the same *rate of convergence* as its parental complete-data estimator $\hat{f}_{\rm com}$. That is, if $||\hat{f}_{\rm com} - f|| = O(N^{-\alpha})$, then $||\hat{f}_{\rm obs} - f|| = O(n^{-\alpha})$, as long as r is bounded away from one (i.e., 100% missing) as N grows. In this sense, the self-consistent approach is an optimal approach for dealing with missing data, because it maintains the same rate of convergence as the parental complete-data estimator.

2.4 Connection with Maximum Likelihood Estimation

The optimality of the self-consistent estimator perhaps can be best seen in parametric cases, because it can be shown that under mild regularity conditions, parametric maximum likelihood estimators (MLE) must be self-consistent, at least asymptotically. The linear regression example in Section 2.1 has already hinted at this possibility, as there, the least-squares estimator for the regression coefficient, which is the MLE under homogenous normal error, is exactly self-consistent.

To see this more generally, suppose our complete set $\boldsymbol{y}_{\rm com}$ is of size N, and let $\ell(\theta|\boldsymbol{y}_{\rm com})$ be the complete-data log-likelihood function with parameter θ, and $S(\theta|\boldsymbol{y}_{\rm com}) = \ell'(\theta|\boldsymbol{y}_{\rm com})$ and $I(\theta) = E[-\ell''(\theta|\boldsymbol{y}_{\rm com})|\theta]$ be the corresponding score function and (expected) Fisher information. For simplicity of illustration, here we assume θ is a scalar quantity, but the results below extend straightforwardly to multivariate θ. Denoting the complete-data MLE by $\hat{\theta}_{\rm com}$, then under standard regularity conditions, $S(\hat{\theta}_{\rm com}|\boldsymbol{y}_{\rm com}) = 0$. Expanding this identity around $S(\theta|\boldsymbol{y}_{\rm com})$ yields the standard Taylor expansion or asymptotic representation[18] of $\hat{\theta}_{\rm com}$,

$$\hat{\theta}_{\rm com} - \theta = \frac{S(\theta|\boldsymbol{y}_{\rm com})}{I(\theta)} + R_N, \tag{12}$$

where R_N is a remainder term typically of order $o_p(N^{-1/2})$, a notation that means that $\sqrt{N}R_N$ converges to zero, in probability, as $N \to \infty$. Taking the conditional expectation of both sides of (12) with respect to $p(\boldsymbol{y}_{\rm com}|\boldsymbol{y}_{\rm obs};\theta)$, we obtain

$$E[\hat{\theta}_{\rm com}|\boldsymbol{y}_{\rm obs};\theta] - \theta = \frac{E[S(\theta|\boldsymbol{y}_{\rm com})|\boldsymbol{y}_{\rm obs};\theta]}{I(\theta)} + \tilde{R}_n, \tag{13}$$

where $\tilde{R}_n = E[R_N|\boldsymbol{y}_{\rm obs};\theta]$ is $o_p(n^{-1/2})$ (n is the size of $\boldsymbol{y}_{\rm obs}$) under mild regularity conditions on R_N (e.g., $\sqrt{N}R_N$ converges to zero in L^p for some $p > 1$). Because of the Fisher's identity[19] (which is the fundamental identity underlying the EM algorithm[4,20])

$$E[S(\theta|\boldsymbol{y}_{\rm com})|\boldsymbol{y}_{\rm obs};\theta] = S(\theta|\boldsymbol{y}_{\rm obs}), \tag{14}$$

and since $S(\theta|\boldsymbol{y}_{\rm obs})$ is the observed-data score function, hence $S(\hat{\theta}_{\rm obs}|\boldsymbol{y}_{\rm obs}) = 0$, we see that the observed-data MLE $\hat{\theta}_{\rm obs}$ must satisfy

$$E[\hat{\theta}_{\rm com}|\boldsymbol{y}_{\rm obs}, \theta = \hat{\theta}_{\rm obs}] = \hat{\theta}_{\rm obs} + o_p(n^{-1/2}). \tag{15}$$

Consequently, the MLE is self-consistent, asymptotically, to the order of $n^{-1/2}$. In this sense, the self-consistency principle is a more basic principle than the maximum likelihood principle for statistical inference.

3. WAVELET DE-NOISING WITH MISSING AND IRREGULARLY-SPACED DATA

Of course the self-consistency principle will remain largely as a principle only, if the self-consistency equation is hard to solve in general. Fortunately, this is not the case, although, like the MLE, it is not always easy to solve. Here we review our recent work on applying this general method to wavelet de-noising with missing and irregularly-spaced data. Details can be found in the full papers,[7,8] including literature review and comparisons with other methods.

3.1 The MISC Algorithm

Our first algorithm, termed the *multiple imputation self-consistent (MISC) algorithm*, is a generic Monte Carlo based method. Because it uses Monte Carlo simulation, it avoids the analytic difficulty of solving the self-consistency equation (6). Consequently, it is generally applicable; indeed it can be easily modified to handle the L^p norm for general p. A necessary price one pays for this generality, much like the standard robustness and efficiency trade-off, is that it can be computationally very intensive, precisely because it relies on Monte Carlo simulation.

To describe MISC, we need some notation. Let $\boldsymbol{I}_{\rm obs}$ be the "observed data index set": $i \in \boldsymbol{I}_{\rm obs}$ if and only if the ith data point (x_i, y_i) in (5) is observed. Let $\boldsymbol{y}_{\rm com} = (y_1, \ldots, y_N)^T$ denote the complete responses, and let $\boldsymbol{y}_{\rm mis} = \{y_i : i \notin \boldsymbol{I}_{\rm obs}\}$ and $\boldsymbol{y}_{\rm obs} = \{y_i : i \in \boldsymbol{I}_{\rm obs}\}$, that is, the missing and observed portions of $\boldsymbol{y}_{\rm com}$, respectively. Let the covariance matrix of $\boldsymbol{y}$ be $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{\Sigma}$, which is *not* assumed to be diagonal, because we allow the error terms e_i's to be correlated, as long as $\boldsymbol{\Sigma}$ can be identified from $\boldsymbol{y}_{\rm obs}$.

Under this setup, assume that we have chosen a complete-data wavelet de-noising procedure, such as the SURE method[21], we then run the following iterative algorithm to compute the corresponding wavelet estimates based on the observed data. Let $\hat{f}^{(0)}$ and $\hat{\boldsymbol{\Sigma}}^{(0)}$ be our initial guesses of the signal f and noise level $\boldsymbol{\Sigma}$. Then each iteration of MISC consists of the following three steps for $t = 1, \ldots$:

Step 1 *Multiple Imputation*: For $m = 1, \ldots, M$, simulate $\boldsymbol{y}_{{\rm mis},m}$ independently from

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

Step 2 *Wavelet Shrinkage:* For $m = 1, \ldots, M$, apply the chosen complete-data wavelet shrinkage procedure to the *completed data* $\boldsymbol{y}_m = \{\boldsymbol{y}_{\rm obs}, \boldsymbol{y}_{{\rm mis},m}\}$ and obtain $\hat{f}_m(x_i)$, $i = 1, \ldots, N$.

Step 3 *Combining Estimates:* Compute the t-th iterative estimate of f as

$$\hat{f}^{(t)}(x_i) = \frac{1}{M} \sum_{m=1}^{M} \hat{f}_m(x_i), \quad i = 1, \ldots, N. \tag{16}$$

Also, use the residuals $\{y_i - \hat{f}^{(t)}(x_i) : i \in \boldsymbol{I}_{\rm obs}\}$ to obtain an efficient estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$, such as the MLE, of $\boldsymbol{\Sigma}$.

Clearly for Step 1, the larger the M is, the better results one would expect, but also the longer the computational time it would require. Our numerical experience so far suggests that M in the vicinity of 100 will often produce acceptable results, as long as the amount of missing data — or more precisely, the *fraction of missing information*[20]— is not too large. An approximate analysis shows that the number of iterations is roughly linear in $[-\log(r)]^{-1}$, where r is the missing percentage. Hence the computational load for MISC is roughly determined by $-M[\log(r)]^{-1}$, which goes to infinity as M or r becomes large.

3.2 A More Refined Algorithm

Although it is essentially impossible to avoid the potentially intensive computation if we insist on a single generic algorithm, such as MISC, to deal with all situations, it is entirely within our reach to design much more efficient algorithms if we restrict ourselves to specific classes of applications. One such class of algorithms was developed in our recent work,[7,8] by using a simple yet effective approximation in the case of either hard or soft thresholding. This class of algorithms was developed based on the observation that if we pretend that the noise variance $\mathbf{\Sigma}$ is known and diagonal, then we can obtain, under the normality assumption, a closed-form expression for the conditional expectation on the left-hand side of (6), thereby avoiding the need for multiple imputation (a.k.a., Monte Carlo simulation) in Step 1 of MISC.

To see this more clearly, consider a case with hard-thresholding at a fixed cut-off value c. We will use single-indexing w_l instead of the usual double-indexing w_{jk} notation to denote a wavelet coefficient. Because the complete-data wavelet estimator under the hard-thresholding is of the form

$$\hat{f}_{\text{com}} = \sum_l 1_{\{|w_l| \geq c\}} w_l \psi_l(t), \tag{17}$$

where $\{w_l; l = 1, \ldots\}$ are complete-data empirical wavelet coefficients and $\{\psi_l(t); l = 1, \ldots, \}$ are a wavelet basis from a chosen mother wavelet (e.g., Dauvechies V [22]), calculating the conditional expectation needed by the left-hand side of (6) at the t-th iteration amounts to calculating all

$$\tilde{w}_l^{(t)} \equiv E\left[1_{\{|w_l| \geq c\}} w_l \middle| \boldsymbol{y}_{\text{obs}}; f = \hat{f}^{(t-1)}\right], \quad \text{for } l = 1, \ldots \tag{18}$$

The real complication in computing (18) is that the thresholding value c is typically in the form $c = \hat{\sigma} h(N)$, where $\hat{\sigma}$ is an estimated standard error of w_l and $h(N)$ is a known function of the sample size, depending on the choice of the thresholding method; for example, universal thresholding uses $h(N) = \sqrt{2 \log N}$,[23] and a later improvement suggests $h(N) = \sqrt{2 \log N - \log(1 + 256 \log N)}$.[24] However, if we can ignore the uncertainty in $\hat{\sigma}$, which often is not a terrible assumption in the signal processing content (or putting it differently, one often has much more serious problems to worry about than the uncertainty in $\hat{\sigma}$), then the calculation of (18) can be done analytically under the assumption that the error terms in (5) are independent normal variables (i.e., white noise).

Using this approximation, we[7,8] obtained the following very effective algorithm, in terms of both statistical efficiency (e.g., reducing errors in de-noising estimation) and computational efficiency (e.g., all steps are carried out without any costly Monte Carlo simulation). This algorithm was labelled as "REF" (for "refined") because it is a refinement of an earlier crude approximation.[7,8]

Specifically, the "REF" algorithm has the following five steps at each iteration:

Step 1 For each i such that $i \notin \boldsymbol{I}_{\text{obs}}$, impute the corresponding missing y_i by $y_i^{(t)} = \hat{f}^{(t-1)}(x_i)$, which creates a completed-data set: $\boldsymbol{y}_{\text{com}}^{(t)} = \{y_i : i \in \boldsymbol{I}_{\text{obs}}\} \cup \{y_i^{(t)} : i \notin \boldsymbol{I}_{\text{obs}}\}$.

Step 2 Apply a Discrete Wavelet Transform (DWT) to $\boldsymbol{y}_{\text{com}}^{(t)}$ to obtain the empirical wavelet coefficients $\boldsymbol{w}^{(t)} = \boldsymbol{W} \boldsymbol{y}_{\text{com}}^{(t)}$.

Step 3 Obtain a robust estimate $\tilde{\sigma}^{(t)}$ of σ from $\boldsymbol{w}^{(t)}$, for example, the median absolute deviation method.[23] Inflate this error estimate to

$$\hat{\sigma}^{(t)} = \sqrt{\{\tilde{\sigma}^{(t)}\}^2 + r\{\hat{\sigma}^{(t-1)}\}^2}, \tag{19}$$

where $r = 1 - \frac{n}{N}$ is the fraction of missing observations.

Step 4 Compute

$$\tilde{w}_l^{(t)} = \alpha(w_l^{(t)}, r) + \beta(w_l^{(t)}, r) \times w_l^{(t)}, \tag{20}$$

degraded reconstructed

Figure 1. Experimenting on ourselves (Lee "choked" by Meng.) The left image was degraded with patches of missing pixels, obtained by masking with random clusters from a truncated Gaussian random field. The right image is a reconstruction by the REF algorithm using separable Daubechies-V wavelets, with primary resolution level 3.

where the α and β functions are given by

$$\alpha(w,\eta) = \frac{\eta\sigma}{\sqrt{2\pi}}\left\{e^{-\frac{1}{2}\left(\frac{c-w}{\eta\sigma}\right)^2} - e^{-\frac{1}{2}\left(\frac{c+w}{\eta\sigma}\right)^2}\right\} \quad \text{and} \quad \beta(w,\eta) = 2 - \Phi\left(\frac{c-w}{\eta\sigma}\right) - \Phi\left(\frac{c+w}{\eta\sigma}\right), \tag{21}$$

with Φ being the CDF function of $\mathcal{N}(0,1)$.

Step 5 Apply the inverse DWT to $\tilde{\boldsymbol{w}}^{(t)}$ and obtain the t-th iterative estimate $\hat{\boldsymbol{f}}^{(t)} = \boldsymbol{W}^T\tilde{\boldsymbol{w}}^{(t)}$.

Note that the inflation formula given in (19) is important, because it (approximately) corrects for the underestimation of the variance due to regression imputation. As emphasized in Section 1, without such a correction, because it puts all the missing data exactly on the regression line, the regression imputation method employed by Step 1 will lead to underestimation of σ^2, which in turn will lead to poorer quality of signal reconstruction.[7,8]

We also remark that Step 4 can be easily modified to handle *soft* thresholding: $1_{(|w_l|\geq c)}\text{sign}(w_l)\{|w_l| - c\}$. We can do this by adding a simple term to $\tilde{w}_l^{(t)}$ of (20), now relabelled as $\tilde{w}_{l,hard}^{(t)}$. That is, we replace (20) with

$$\tilde{w}_{l,soft}^{(t)} = \tilde{w}_{l,hard}^{(t)} + c\left\{\Phi\left(\frac{c-w_l^{(t)}}{r\sigma}\right) - \Phi\left(\frac{c+w_l^{(t)}}{r\sigma}\right)\right\}. \tag{22}$$

3.3 The Sky is Unlimited ...

To conclude this presentation, I want to emphasize that applications of self-consistent methods are unlimited. Even just for wavelet de-noising problems, the algorithms discussed above are by no means limited to one dimensional signal. It can easily be adapted to handle two dimensional data, especially when using separable wavelets. And, although the REF algorithm was derived under the assumption that the observed data are a random sample of the complete data, we can nevertheless apply it even when this assumption fails. As an illustration, Figure 1 shows a degraded picture by *clustered* missing pixel values, and its reconstruction via a 2-D version of our REF algorithm. Although the reconstructed photo still makes my white shirt too colorful, the great smiles — after seeing how well our algorithms worked — are much less distracted from the viewer by the black patches.

ACKNOWLEDGMENTS

My thanks go to Thomas Lee for introducing me to the fascinating world of wavelets and for all the great collaborations (and programming/implementaion!), to Zhan Li for many discussions on the theoretical results, and to Paul Baines for careful proofreading. I also want to thank the meeting organizers, Dimitri Van De Ville, Vivek Goyal, and Manos Papadakis, for their kind invitation, and the Editor Mary Starz for her generosity in granting me several extensions of deadline, which made it possible for me to prepare this written version of my presentation. The work reported here is support in part by a number of NSF grants.

REFERENCES

1. R. J. Little and D. B. Rubin, *Statistical Analysis With Missing Data (2nd Ed)*, John Wiley & Sons, New York, 2002.
2. A. Gelman and X.-L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, John Wiley & Sons, England, 2004.
3. D. B. Rubin, "Inference and missing data," *Biometrika* **63**, pp. 581–592, 1976.
4. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B, Methodological* **39**, pp. 1–37, 1977.
5. X.-L. Meng and D. A. van Dyk, "The EM algorithm – an old folk song sung to a fast new tune (with discussion)," *Journal of the Royal Statistical Society, Series B, Methodological* **59**, pp. 511–567, 1997.
6. D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, New York, 1987.
7. T. Lee and X.-L. Meng, "A self-consistent wavelet method for denoising images with missing pixels," *Proceedings of the 30th IEEE International Conference on Acoustics, Speech, and Signal Processing* , pp. 41–44, 2005.
8. T. Lee and X.-L. Meng, "Self-consistency: A general recipe for wavelet estimation with irregularly-spaced and/or incomplete data," *Under review by Statistical Science; Available at http://arxiv.org/abs/math.ST/0701196* , 2007.
9. K. Hoffmann and M. Schreiber, eds., *Computational Statistical Physics: From Billiards to Monte Carlo*, Springer, England, 2002.
10. M. A. Tanner and W. H. Wong, "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association* **82**, pp. 528–550, 1987.
11. X.-L. Meng and D. A. van Dyk, "Seeking efficient data augmentation schemes via conditional and marginal augmentation," *Biometrika* **86**, pp. 301–320, 1999.
12. D. A. van Dyk and X.-L. Meng, "The art of data augmentation (with discussion)," *Journal of Computational and Graphical Statistics* **10**, pp. 1–111, 2001.
13. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.
14. B. Efron, "The two sample problem with censored data," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, **IV**, pp. 831–851, Berkeley: University of California Press, 1967.
15. E. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association* **282**, pp. 457–481, 1958.
16. T. Tarpey and B. Flury, "Self–Consistency: A fundamental concept in statistics," *Statistical Science* **11**, pp. 229–243, 1996.
17. V. Bryant, *Metric Spaces: Iteration and Application*, Cambridge University Press, Cambridge, 1985.
18. T. S. Ferguson, *A Course in Large Sample Theory*, Chapman & Hall/CRC, London, 1996.
19. R. A. Fisher, "Theory of statistical estimation," *Proceedings of Cambridge Philosophical Society* **22**, pp. 700–725, 1925.
20. X.-L. Meng and D. B. Rubin, "Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm," *Journal of the American Statistical Association* **86**, pp. 899–909, 1991.
21. D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association* **90**, pp. 1200–1224, 1995.

22. I. Daubechies, *Ten Lectures on Wavelets*, Philadelphia: SIAM, 1992.
23. D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika* **81**, pp. 425–455, 1994.
24. A. Antoniadis and J. Fan, "Regularization of wavelet approximations (with discussion)," *Journal of the American Statistical Association* **96**, pp. 939–967, 2001.