

# Unbiased reconstruction of the large-scale structure

Saleem Zaroubi<sup>★</sup>

*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild Str. 1, D-85740, Garching, Germany*

Accepted 2001 December 5. Received 2001 November 30; in original form 2000 November 3

## ABSTRACT

We present a new unbiased minimal variance (UMV) estimator for the purpose of reconstructing the large-scale structure of the Universe from noisy, sparse and incomplete data. Similar to the Wiener filter (WF), the UMV estimator is derived by requiring the linear minimal variance solution given the data and an assumed a priori model specifying the underlying field covariance matrix. However, unlike the WF, the minimization is carried out with the added constraint of an unbiased reconstructed mean field. The new estimator does not necessitate a noise model to estimate the underlying field; however, such a model is required for evaluating the errors at each point in space. The general application of the UMV estimator is to predict the values of the reconstructed field in unsampled regions of space (e.g. interpolation in the unobserved Zone of Avoidance), and to dynamically transform from one measured field to another (e.g. inversion of radial peculiar velocities to over-densities). Here, we provide two very simple applications of the method. The first is to recover a 1D signal from noisy, convolved data with gaps, for example CMB time-ordered data. The second application is a reconstruction of the density and 3D peculiar velocity fields from mock SEcat galaxy peculiar velocity catalogues.

**Key words:** methods: statistical – cosmology: theory – large-scale structure of Universe.

## 1 INTRODUCTION

The mapping of the distribution of galaxies and their peculiar velocity field constitutes a major research area in modern astronomy, setting both the observational and theoretical foundations of cosmology and, in particular, of large-scale structure (LSS). Within the framework of gravitational instability, the large-scale galaxy distribution offers a probe of the nature of the primordial perturbation field, and can be used to set strong constraints on the values of cosmological parameters. However, since astronomical observations give only incomplete and noisy information on the real Universe, the recovery of the underlying signal from these observations can be a non-trivial task, forcing one to resort to regularization methods, for example Wiener filtering (Wiener 1949; Rybicki & Press 1992; Zaroubi et al. 1995) or the Maximum Entropy algorithm (Gull 1989).

In particular, the WF has been widely applied to galaxy surveys (Lahav et al. 1994; Fisher et al. 1995; Schmoldt et al. 2000), CMB studies (Bunn et al. 1994; Tegmark & Efstathiou 1996, Bouchet & Gispert 1999), and galaxy peculiar velocity catalogues (Hoffman et al. 2000; Zaroubi 2000; Zaroubi et al. 1999, 2001a). The WF provides an optimal estimator of the underlying field in the sense of a minimum-variance solution given the data and an assumed a priori model (Wiener 1949; Press et al. 1992). The model defines

the data auto-correlation and the data-field cross-correlation matrices. In the case where the data is drawn from a random Gaussian field, the WF estimator coincides with the conditional mean field and with the most probable configuration given the data (see Zaroubi et al. 1995).

Although the application of the WF is very simple and has proved to be useful for many purposes, it is easy to show that the estimator is intrinsically biased, often in a scale-dependent manner. The main cause of this bias stems from the modulation introduced by the signal/(signal + noise) weighting it invokes. This drawback has prevented the use of the Wiener reconstructed maps in many areas, for example power spectrum estimations, bias parameter extraction from galaxy peculiar velocity data comparison with galaxy survey data, etc. To account for this bias in the power spectrum estimation a correction factor is often applied to the Wiener reconstructed signal (Rybicki & Press 1992; Tegmark & Efstathiou 1996).

In this paper, we propose a new linear unbiased minimal variance (UMV) estimator that is designed to avoid the intrinsic bias that exists in the WF. This is achieved by solving the minimization equation subject to the constraint of an unbiased mean underlying field. To test the UMV estimator we apply it (1) to recover a 1D time series from convolved and noisy data with gaps, and (2) to a mock SEcat peculiar velocity data set, a combination of the SFI (Giovanelli et al. 1998) and the ENEAR (da Costa et al. 2000) galaxy peculiar velocity catalogues, for which we

<sup>★</sup>E-mail: saleem@mpa-garching.mpg.de

reconstruct the distribution of the true density and velocity fields from which the mock catalogue is constructed.

The outline of this paper is as follows. The method is presented in Section 2. In Section 3 two simple extra regularization methods are discussed. Section 4 shows how constrained realizations could be produced within the UMV framework. The method is tested using artificial data based on simulations in Section 5. The results are discussed and the conclusions are summarized in Section 6.

## 2 THEORY

### 2.1 Derivation of the UMV estimator

Consider the case of a set of observations, or measurements, performed on an underlying field  $s = \{s_\alpha\} (\alpha = 1, \dots, N)$ , or on any field linearly related to  $s$ , which yields a set of data points  $d = \{d_i\} (i = 1, \dots, M)$ . In particular, we are interested in measurements that can be modelled mathematically as a linear convolution or mapping of the underlying field,

$$d = o + \epsilon = \mathbf{R}s + \epsilon, \quad (1)$$

where  $o = \mathbf{R}s$ ,  $\mathbf{R}$  is an  $M \times N$  matrix which represents the response or point spread function (hereafter RF), and  $\epsilon$  is the noise vector associated with the data. The RF will be treated as any function that linearly connects the underlying signal to the data, be it blurring or smoothing introduced by the measurement, some theoretical relationship between two fields, or any other linear transformation on the underlying signal.

In principle, the reconstruction of  $s$  can be accomplished by inverting equation (1). However, two main obstacles usually prevent one from pursuing this approach. First, the number of independent data points is usually much smaller than the number of underlying degrees of freedom. Secondly, the presence of noise can render such a direct inversion unstable and the obtained results meaningless. As a result of these potential difficulties, one is often forced to resort to some statistical regularization techniques (e.g. WF) in order to solve equation (1).

Similar to in the derivation of the WF, we assume prior knowledge of the signal correlation function

$$\mathbf{S} = \langle ss^+ \rangle, \quad (2)$$

where  $s^+$  is the complex conjugate of the transpose of the underlying signal, and angled brackets denote the signal ensemble average. Notice that there have been no assumptions made regarding the actual probability distribution function from which the field  $s$  is drawn. We define the unbiased minimal variance estimator  $s^{\text{UMV}} = \mathbf{H}d$ , where  $\mathbf{H}$  is the  $N \times M$  matrix that minimizes the variance of the residual  $r = s - s^{\text{UMV}}$ , while satisfying the constraint

$$[s^{\text{UMV}}]_{\mathcal{N}} = [\mathbf{H}d]_{\mathcal{N}} = s. \quad (3)$$

The ensemble average,  $[\dots]_{\mathcal{N}}$ , is an ensemble average over noise realizations and is very different from the ensemble average,  $\langle \dots \rangle$ , which denotes an ensemble average over signal realizations. This distinction between the two ensemble averages will be used only in this section; in the rest of the paper we use only the signal ensemble average.

In short, we are seeking  $\mathbf{H}$  that minimizes

$$\langle [rr^+ + \lambda \mathbf{H}d]_{\mathcal{N}} \rangle = \langle [(s - s^{\text{UMV}})(s - s^{\text{UMV}})^+ + \lambda \mathbf{H}d]_{\mathcal{N}} \rangle, \quad (4)$$

where  $\lambda$  is a Lagrange multiplier. The ensemble average over noise realizations precedes the one over signal realization. The order of

averaging over ensembles is very important since if it is reversed the contribution of the term  $\mathbf{H}d$  will be null.

The constraint, introduced in equation (3), assumes that the data are unbiased, namely that the errors are random, and therefore requires that the estimator does not alter the value of the measured data points, but rather forces the field to retain its measured values at the appropriate locations. However, this requirement does not guarantee an unbiased variance of the reconstruction; on the contrary, one expects that the variance of the reconstructed field is some compromise between the (signal + noise) variance of the data points and the assumed variance of the underlying signal.

Carrying out the minimization of equation (4) with respect to  $\mathbf{H}$ , one obtains an equation that together with equation (3) is used to solve for  $\mathbf{H}$  and  $\lambda$ . The solution yields the UMV estimator,

$$s^{\text{UMV}} = \langle so^+ \rangle \langle oo^+ \rangle^{-1} d. \quad (5)$$

The Lagrange multiplier,  $\lambda$ , is roughly proportional to the noise/(signal + noise), making it dominant when the noise is dominant and small when the signal/noise  $\gg 1$ .

In the absence of a response function that operates on the signal, namely  $\mathbf{R} = \mathbf{I}$  (where  $\mathbf{I}$  is the unity matrix), the reconstructed signal at the location of the data points is identical to the data measured values, which is consistent with the constraint given in equation (3). In the rest of space the degrees of freedom are recovered by interpolating the data points in a manner consistent with the correlation assumed in the underlying theory.

Mathematically, the difference between the WF and the current estimator is that in the former the term  $\langle oo^+ \rangle$  is replaced by  $\langle dd^+ \rangle$ , a matrix that includes the noise correlations, an addition that accounts for the signal suppression which renders the WF mean field biased.

The variance of the field estimated in equation (5) is

$$\langle s^{\text{UMV}} s^{\text{UMV}+} \rangle = \langle so^+ \rangle \langle oo^+ \rangle^{-1} (\langle oo^+ \rangle + \mathbf{N}) \langle oo^+ \rangle^{-1} \langle os^+ \rangle, \quad (6)$$

where  $\mathbf{N}$  is the noise correlation matrix. The variance of the residual is

$$\begin{aligned} \langle rr^+ \rangle &= \langle ss^+ \rangle - \langle so^+ \rangle \langle oo^+ \rangle^{-1} \langle os^+ \rangle \\ &\quad + \langle so^+ \rangle \langle oo^+ \rangle^{-1} \mathbf{N} \langle oo^+ \rangle^{-1} \langle os^+ \rangle. \end{aligned} \quad (7)$$

As a simple example, assume that  $\mathbf{R} = \mathbf{I}$  and that the field is estimated at the exact locations of the data points: the variance in equation (6) simply reduces to  $\langle ss^+ \rangle + \mathbf{N}$ , recovering the power spectrum of the signal + noise at those points. The variance of the residual (equation 7) at the location of the data points is simply reduced to the correlation matrix of the noise,  $\mathbf{N}$ . In addition, when the data points are uncorrelated with the rest of the underlying degrees of freedom, the reconstructed values, at locations different from those of the data points, are zero (equation 6), and the variance of the residual at those same locations, as obtained from equation (7), is simply the underlying prior correlation.

Substitution of  $o = \mathbf{R}s$  in equation (5) yields

$$s^{\text{UMV}} = \langle ss^+ \mathbf{R}^+ \rangle \langle \mathbf{R}ss^+ \mathbf{R}^+ \rangle^{-1} d, \quad (8)$$

which one could be tempted to simplify further to obtain  $s^{\text{UMV}} = \langle ss^+ \rangle \langle ss^+ \rangle^{-1} \mathbf{R}^{-1} d = \mathbf{R}^{-1} d$ . Normally, however, the matrix  $\mathbf{R}$  is not square but rather has a larger number of rows than columns, its inverse is not unique, and the inverse of its transpose,  $\mathbf{R}^+$ , does not exist at all. Therefore, in most cases carrying the simplification further is mathematically incorrect. This simplification is possible if the inverses of both  $\mathbf{R}$  and  $\mathbf{R}^+$

exist, which is only true if the number of data points is identical to the number of degrees of freedom in the underlying signal, or when the signal correlation function is a Dirac delta function.

However, when the simplification can be performed the UMV estimator is equivalent to direct inversion. Since direct inversion in the presence of noise could produce very large uncertainties in the reconstructed signal one should resort to extra regularization that reduces the variance to a manageable size. An example of such an extra regularization is given in Section 3.

To summarize, the regularization strength of the UMV estimator stems from the cross-talk between the underlying signal and the data points at different locations. The stronger the correlation is, the more advantageous the use of the UMV. With the absence of such a cross-talk, the number of degrees of freedom that one can reconstruct is identical to the number of data points, and the UMV estimator is reduced in this case back to  $\mathbf{R}^{-1}$ ; that is, it is equivalent to direct inversion, a case that requires extra regularization in order to reduce the, often huge, uncertainty introduced by the noise.

### 3 EXTRA REGULARIZATION

As shown in the previous section, the UMV could be equivalent to direct inversion; therefore, while it produces an unbiased estimator of the underlying field, it gives a ‘minimum variance’ that is too large to be useful. In contrast, the WF filter produces a manageable variance but a large degree of bias. Hence, the UMV alone is often insufficient to stabilize the deconvolution and consequently it must either be modified, be supplemented with extra regularization, or replaced with another method.

Of course, to solve the inversion problem fully one might want to use more sophisticated non-linear methods such as maximum entropy (Gull 1989), or Pixons image restoration algorithm (Piña & Puetter 1993), etc. These methods, however more advanced, are much more computationally expensive and complicated to apply.

In this paper we choose to modify the UMV inversion method with extra regularization that is both simple and easy to integrate into the algorithm. In Subsections 3.1 and 3.2, two simple extra regularization methods are introduced, the singular value decomposition (SVD) algorithm (e.g. Press et al. 1992), and a WF-like regularization. These methods are basically applied to find the vector  $\mathbf{a}$ , defined as  $\mathbf{a} \equiv \mathbf{O}^{-1}\mathbf{d} \equiv \langle \mathbf{oo}^+ \rangle^{-1}\mathbf{d}$  (see equation 5). The difficulty in finding  $\mathbf{a}$  could be due to the inversion instability of  $\mathbf{O}$  or due to the noise in vector  $\mathbf{d}$ . In both methods suggested here this difficulty is solved by, in effect, adding elements to the diagonal of  $\mathbf{O}$  or to its eigenvalues, so that the inversion of  $\mathbf{O}$  is stabilized and at the same time the noise contribution of  $\mathbf{d}$  to the vector  $\mathbf{a}$  is suppressed.

#### 3.1 The singular value decomposition algorithm as a regularizer

The following presentation follows the one in Zaroubi et al. (1995). Mathematically, the solution of equation (5) involves the solution of the equation

$$\mathbf{O}\mathbf{a} = \mathbf{d}, \quad (9)$$

for the unknown vector  $\mathbf{a}$ .

The SVD algorithm basically decomposes the positive-definite matrix<sup>1</sup>  $\mathbf{O}^*$  into a multiplication of three matrices,  $\mathbf{O} = \mathbf{U}\text{diag}\{w_i\}$

<sup>1</sup>The matrix  $\mathbf{O}$  is an auto-correlation matrix, therefore it is a square positive-definite matrix.

$\mathbf{U}^+$ , where the set  $\{w_i\}$  is referred to as the collection of singular values, or the eigenmodes, and the columns of the matrix  $\mathbf{U}$  are orthogonal and proportional to the eigenvectors of  $\mathbf{O}$ . The inversion, after decomposition, is straightforward and gives  $\mathbf{O}^{-1} = \mathbf{U}\text{diag}\{1/w_i\}\mathbf{U}^+$ . Formally speaking, equation (9) has a unique solution if and only if  $\mathbf{O}$  is a non-singular matrix, namely if  $w_i \neq 0$  for all  $i$ . However, a meaningful solution can be obtained even in the case where  $\mathbf{O}$  is singular, by requiring the solution to minimize the norm of the residuals,  $|\mathbf{O}\mathbf{a} - \mathbf{d}|$ . Such a solution is obtained by substituting  $1/w_i = 0$  in the expression for the inverse for any  $w_i = 0$  (Press et al. 1992).

In any given problem, the lower limit of the singular values below which the inverse values are set to zero must be set. In general, the singular values measure the amount of ‘information’ carried by each mode in the problem (Press et al. 1992); in other words, the small singular values do not contribute significantly to the reconstruction, but they can destabilize the inversion. As an extension of the ideal case of  $w_i = 0$ , we impose a cut-off on the small singular values in order to maintain stability. Often, the structure of the sorted spectrum of singular values contains a very sharp drop normally appearing as a ‘knee’. The location of the ‘knee’ usually determines the singular values that contain sufficient information. A simple cut-off at the location of the ‘knee’ normally does the trick and stabilizes the inversion.

Another possible criterion for the choice of the cut-off would be by expanding the matrix  $\mathbf{O}$  in terms of signal-to-noise eigenmodes (Bond 1995). This method allows a simultaneous diagonalization of the matrix  $\mathbf{O}$  and the noise matrix  $\mathbf{N}$ , which in turn allows the usage of some signal-to-noise ratio threshold (typically 1) in order to set the cut-off. For a specific application see the 1D example in Section 5.1, especially the expansion shown in Fig. 2 (see also the example discussed in Zaroubi et al. 1995).

#### 3.2 Wiener-filter-like regularizer

The WF by itself is a very robust and efficient regularizer. The main reason for this stems from the relatively high values of the diagonal terms in the matrix  $\langle \mathbf{d}\mathbf{d}^+ \rangle$ . These high values usually come from the noise contribution, especially if it is uncorrelated (e.g. white noise). This aspect is, of course, also responsible for the suppression of the Wiener reconstructed signal.

From the point of view of the SVD algorithm the stabilization effect caused by the noise diagonal elements is quite obvious, especially in the light of the following example. Consider the simplest case of white Gaussian noise where the noise correlation matrix is proportional to the unity matrix,  $\mathbf{I}$ . With the application of the SVD algorithm, the noise contribution is still proportional to the unit matrix and is the same for every singular value. This means that the singular values can be as small as the noise and none of them has a value of ‘zero’; therefore, the matrix is naturally stable and no extra stabilization is normally needed.

In order to utilize the stabilization aspect of the WF on the one hand, while avoiding the signal suppression aspect on the other, we add a very small white noise-like contribution to the correlation matrix  $\mathbf{O}$ . Naturally, one should seek the smallest possible addition in order that the recovered signal is suppressed as little as possible. The choice of the amplitude of the additional noise term could be guided by the same signal-to-noise cut-off criterion as discussed in Section 3.1. This is, of course, problem-dependent and should be carried out with caution.

An example of the application of this approach is discussed in Section 5.1.

#### 4 CONSTRAINED REALIZATIONS

Hoffman & Ribak (1991) showed that, within the framework of Gaussian random fields and a given prior, any realization can be split into two parts: the mean field, or the WF field, which is determined by the constraints (data); and the residual field, which is a Gaussian random field. Thus, one can make a constrained realization of the underlying field given the assumed prior model and the data.

With the UMV reconstructed field one can produce constrained noisy realizations at every point in space only if the noise in the observed quantity is uncorrelated. The formulation here is quite similar to the constrained realizations procedure of Hoffman & Ribak (1991); one produces a random realization of the noisy underlying field and noise,  $\tilde{s}^{\text{Noisy}} (= \tilde{s} + \tilde{\sigma})$ , then samples it in the same way as the actual data are obtained:

$$\tilde{d} = \mathbf{R}\tilde{s} + \tilde{\epsilon}. \quad (10)$$

Here,  $\tilde{s}$  and  $\tilde{\epsilon}$  are random realizations of the underlying field and the statistical uncertainties, respectively. The noise in the random realization of the data  $\tilde{\epsilon}$  is related to the noise in the random realization of the underlying noisy field  $\tilde{\sigma}$  through the response function,

$$\tilde{\epsilon} = \mathbf{R}\tilde{\sigma}. \quad (11)$$

A constrained realization of the field given the data is given by

$$s^{\text{Noisy}} = \tilde{s}^{\text{Noisy}} + \mathbf{H}(d - \tilde{d}). \quad (12)$$

Here  $s^{\text{Noisy}}$  is the constrained noisy realization.

Note that if the noise,  $\tilde{\epsilon}$ , is correlated, one cannot use this formalism as the UMV takes into account only the correlation in the underlying signal. In such a case, an alternative approach is with the noise correlation included in the estimator; here one can expand the UMV estimator to have the form

$$s^{\text{UMV}} = \langle s\mathbf{o}^+ + \sigma\epsilon^+ \rangle \langle \mathbf{o}\mathbf{o}^+ + \epsilon\epsilon^+ \rangle^{-1} d. \quad (13)$$

This kind of noisy constrained realization could be very useful for cases when one has gaps in the data that one would like to fill, not only in a manner consistent with the data but also with the same power spectrum as the signal + noise.

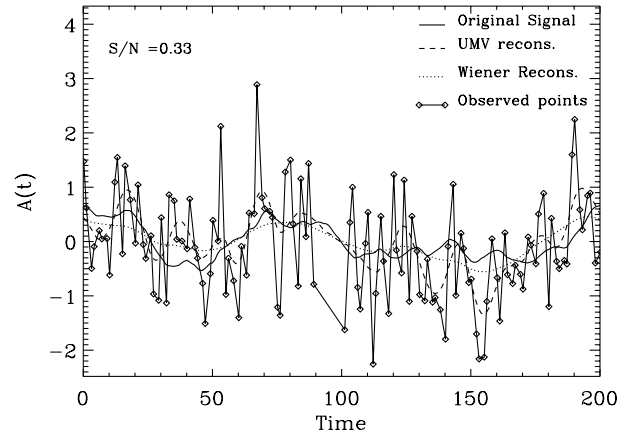
#### 5 APPLICATIONS

The UMV estimator can be applied to many areas of LSS and CMB for (1) interpolating between data points, for example filling in gaps in the time-ordered data to help map-making, bridging the Zone of Avoidance and other uncovered areas in nearly full sky catalogues, etc.; and (2) stabilizing inversions used to transform one dynamical field to another, for example from radial peculiar velocity field to density field, or from redshift to real density.

Here we present two examples: the first is a reconstruction from a 1D time series of noisy convolved signal with gaps, for example CMB time-ordered data; the second shows how this method could be applied to reconstruct the 3D density and peculiar velocity fields from observed galaxy radial peculiar velocity catalogues.

##### 5.1 Time series example: deconvolution of noisy data with gaps

To test the performance of the UMV estimator we proceed with a simple 1D example. Let  $s$  be a random Gaussian time series, with a



**Figure 1.** A 1D reconstruction example. The heavy solid line shows the underlying signal  $A(t)$  as a function of time; both time and amplitude have arbitrary units. The underlying signal is convolved with a Gaussian window function with a width of 5 time units; the convolved signal is then uniformly sampled with the exception of a gap in the time range 90–100. A random noise is then added to produce the ‘data’ points shown with the diamond-shaped connected points; the signal-to-noise ratio in this example is 3. The heavy dashed line shows the UMV reconstructed signal, while the heavy dotted line shows the Wiener reconstructed signal.

known correlation function, which we would like to measure in the time range [0–200] (the signal and time units are arbitrary). The measurement involves a convolution of the signal with a Gaussian window of 5 time units width, and the convolution is described by the matrix  $\mathbf{C}$ . The measurement procedure uniformly samples the signal at about 100 positions, except for the time range of [90–100], where there is a gap in the data. An instrumental white noise,  $\epsilon$ , with standard deviation three times larger than the signal standard deviation is added. Mathematically, the data are connected to the underlying signal by  $d = \mathbf{C}s + \epsilon$ . The heavy solid line in Fig. 1 shows the underlying signal and the connected diamonds represent the measured data.

This specific example is typical of what one obtains in CMB time-ordered data types of measurements where the matrix  $\mathbf{C}$  is the instrument’s point spread function, the noise level reflects the detectors sensitivity, and the gaps are drop-outs in the data stream.

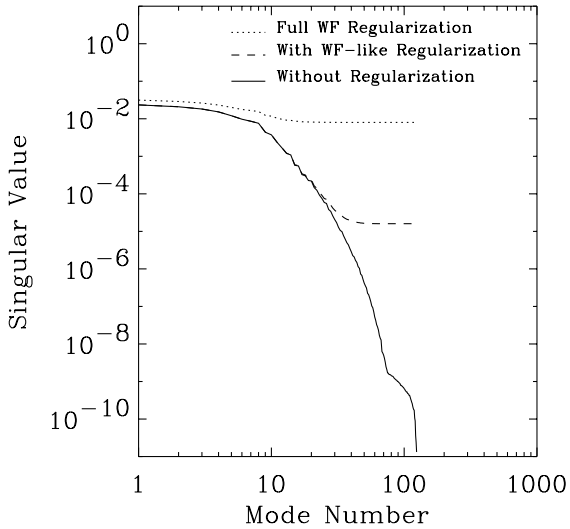
The UMV and WF estimators for this case are

$$s^{\text{UMV}} = \langle ss^+ + \mathbf{C}^+ \rangle \langle \mathbf{C}ss^+ + \mathbf{C}^+ \rangle^{-1} d, \quad (14)$$

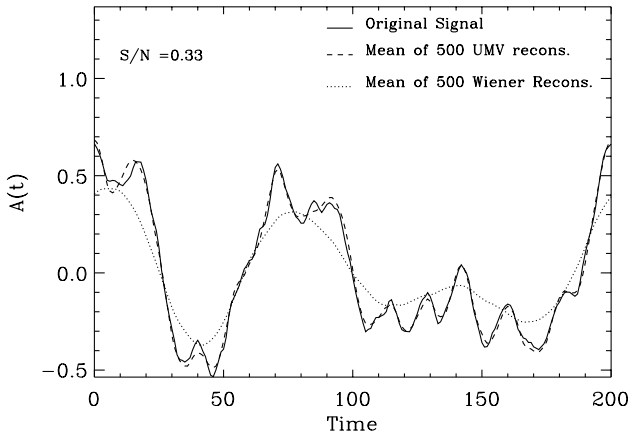
and,

$$s^{\text{WF}} = \langle ss^+ + \mathbf{C}^+ \rangle \langle \mathbf{C}ss^+ + \mathbf{C}^+ + \epsilon^2 \mathbf{I} \rangle^{-1} d. \quad (15)$$

Owing to the large correlation length in this example, the matrix  $\mathbf{O} \equiv \langle \mathbf{C}ss^+ + \mathbf{C}^+ \rangle$  in equation (14) is unstable for inversion, and therefore, as previously discussed, one has to apply an extra regularization scheme. The heavy solid line in Fig. 2 shows the sorted spectrum of the singular values ( $w_i$ ) of the matrix  $\mathbf{O}$  versus the mode number. The very abrupt drop of the singular values around mode number 10 indicates that the information content of the rest of the singular values is very small and that they are essentially responsible for the inversion instability. To stabilize the inversion, we adopt the regularization method described in Section 3.2, and add a diagonal constant with 0.005 of the noise contribution. The heavy dashed line in Fig. 2 shows the sorted spectrum of singular values of the regularized  $\mathbf{O}$ . The two lines are identical for large singular values and depart at small singular



**Figure 2.** Sorted spectrum of singular values of the matrix  $\mathbf{O}$ . The heavy solid line shows the singular values of  $\mathbf{O}$  without regularization. The heavy dashed line shows the singular values of  $\mathbf{O}$  after adding a diagonal constant noise matrix: the constant is 0.005 of the noise rms. The ‘knee’ around singular values of  $10^{-9}$  is caused by numerical noise. For comparison, the dotted line shows the singular values of the full WF regularization.



**Figure 3.** The solid line shows the same underlying signal as in Fig. 1. To this signal we add 500 noise realizations to produce Monte-Carlos of the ‘observed’ data. The dashed line shows the mean of the 500 UMV reconstructions of these realization. The dotted line shows the mean of the Wiener reconstruction of the 500 data realizations.

values. For comparison, the dotted line in Fig. 2 shows the singular values of the matrix used in the WF application, namely  $\mathbf{O} + \mathbf{N}$ .

The heavy dashed line in Fig. 1 shows the UMV reconstruction of the signal that follows directly from calculating  $\mathbf{s}^{\text{UMV}}$  from equation (14). The dotted line shows the WF reconstructed signal as obtained from equation (15). The UMV reconstruction follows the data points while taking into account the deconvolution and the prior for the signal temporal correlations. The WF reconstruction is much smoother and has smaller variance than the underlying signal.

To demonstrate the differences between the two reconstructions, Fig. 3 shows an average of 500 reconstructions of data realizations with the same underlying signal but different noise Monte Carlo. The unbiased nature of the UMV and the biased nature of the WF reconstructions are evident. In each UMV reconstruction we have applied the aforementioned WF-like regularization. The amount of

bias introduced by this procedure is negligible, while the bias for the full WF application is not.

## 5.2 Reconstruction from peculiar radial velocity data

Next, we present an example of an application of the UMV estimator to reconstruct the density and 3D velocity fields from galaxy radial peculiar velocity data. The data set used here is a mock catalogue that mimics the SEcat galaxy peculiar velocity catalogue (Zaroubi 2000), which is a combination of the grouped and Malmquist-bias-corrected SFI (Giovannelli et al. 1998) and ENEAR (da Costa et al. 2000) galaxy peculiar velocity catalogues. The catalogue consists of about 2050 objects ( $\approx 1300$  from SFI and  $\approx 750$  from ENEAR), for which it provides radial velocities and inferred distances with errors, of the order of  $\approx 19$  per cent of the distance per galaxy. The sampling is reasonably homogeneous and covers the whole sky outside the Zone of Avoidance; the radial selection function is uniform out to  $\approx 5500 \text{ km s}^{-1}$ . The Zone of Avoidance is about  $\pm 15^\circ$ . The SEcat catalogue contains distances and peculiar velocities of both late-type and early-type galaxies, and therefore has the advantage of sampling both high- and low-density environments, minimizing possible biases that may affect reconstruction from catalogues based on a single population of galaxies.

The mock catalogue is produced as follows. We first generate a random linear Gaussian realization of density and velocity with a  $\Lambda$ CDM power spectrum (with  $\Omega_m = 0.3$  and  $\Lambda = 0.7$ ). Then the mock catalogue is produced so that the locations of the data points in the mock catalogue are the same as for those in the real SEcat catalogue; however, the values of the radial peculiar velocities are taken from a random Gaussian realization of the underlying fields. The original density field is used for comparison with the reconstructed one.

In this example the data points are given as a set of observed radial peculiar velocities  $u_i^o$  sampled at positions  $\mathbf{r}_i$  with estimated errors  $\epsilon_i$ , assumed to be uncorrelated. The observed velocities are thus related to the true underlying velocity field  $\mathbf{v}(\mathbf{r})$ , or its radial component  $u_i$  at  $\mathbf{r}_i$ , via

$$u_i^o = \mathbf{v}(\mathbf{r}_i) \cdot \hat{\mathbf{r}}_i + \epsilon_i \equiv u_i + \epsilon_i. \quad (16)$$

We assume that the peculiar velocity field  $\mathbf{v}(\mathbf{r})$  and the density fluctuation field  $\delta(\mathbf{r})$  are related via linear gravitational–instability theory:  $\delta = f(\Omega)^{-1} \nabla \cdot \mathbf{v}$ , where  $f(\Omega) \approx \Omega^{0.6}$  and  $\Omega$  is the mean universal density parameter. Under the assumption of a specific theoretical prior for the power spectrum  $P(k)$  of the underlying density field, we can write the UMV estimator of the fields as

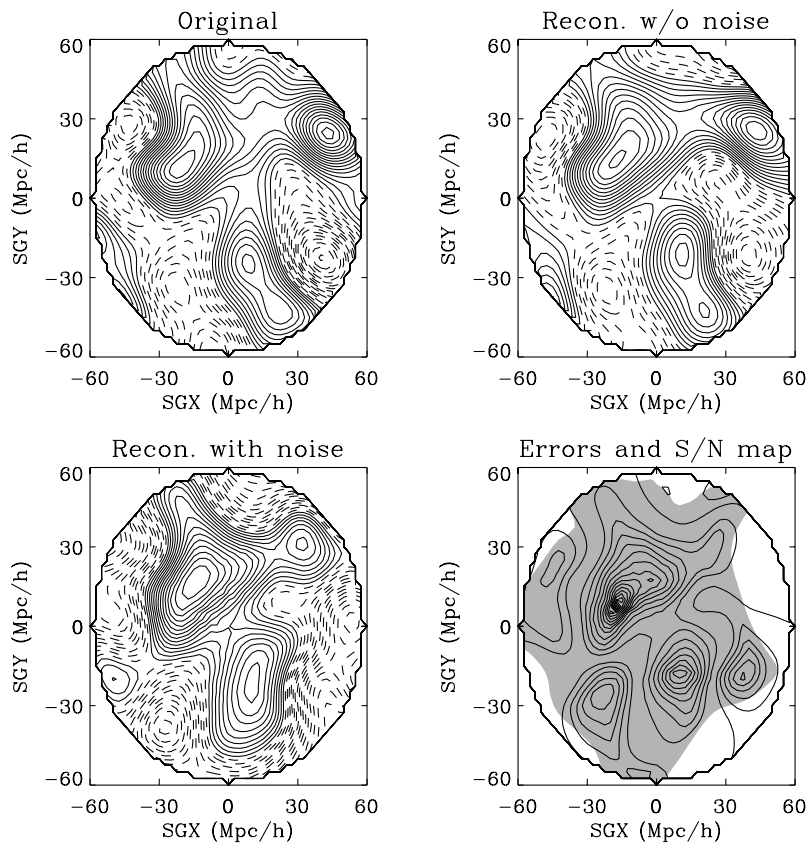
$$\mathbf{v}^{\text{UMV}}(\mathbf{r}) = \langle \mathbf{v}(\mathbf{r}) u_i \rangle \langle u_i u_j \rangle^{-1} u_j^o, \quad (17)$$

and,

$$\delta^{\text{UMV}}(\mathbf{r}) = \langle \delta(\mathbf{r}) u_i \rangle \langle u_i u_j \rangle^{-1} u_j^o. \quad (18)$$

Assuming linear theory and that the velocities are drawn from a Gaussian random field, the two-point velocity–velocity and density–velocity correlation tensors (bracketed quantities in equations 17 and 18) are readily calculated. The calculation of these matrices is discussed elsewhere (Górski 1988; Zaroubi et al. 1995, 1999).

We wish to test two aspects of the reconstruction. First, whether the coverage within the assigned area is good enough for a faithful recovery of the underlying signal. Second, whether the noise level allows a reasonable (high signal/noise) reconstruction within the



**Figure 4.** Testing the method with mock SEcat data. Shown are maps of density in the Supergalactic plane, smoothed with a Gaussian window of  $900 \text{ km s}^{-1}$ , G9. The density contour spacing is 0.1; the mean  $\delta = 0$  contour is heavy; positive contours are solid and negative contours are dashed. The top-left panel shows the original mock Supergalactic plane. The top-right panel is the reconstruction from a noiseless mock catalogue, which shows the uniformity of the sampling and the quality of the interpolation. The bottom-left panel shows a typical UMV reconstruction from noisy data. The bottom-right panel shows the signal-to-noise ratio with contour spacing 1. The heavy solid line indicates a signal-to-noise ratio of unity. The shading indicates regions where the error is less than 0.45.

$60 h^{-1} \text{ Mpc}$  sphere. The success of the reconstruction is demonstrated in Fig. 4. The top-left panel shows the density of the underlying field, used to construct the mock SEcat catalogue, at the  $Z = 0 h^{-1} \text{ Mpc}$  plane, smoothed with a  $9 h^{-1} \text{ Mpc}$  Gaussian, within a sphere of radius  $60 h^{-1} \text{ Mpc}$ ; this is the target map which we are attempting to recover.

In order to test the quality of the coverage we construct a noise-free mock SEcat catalogue which has accurate radial velocities at the locations of the data points. The top-right panel in Fig. 4 shows the density reconstruction from the noise-free catalogue; this map tests mainly the uniformity of the catalogue within a sphere of  $60 h^{-1} \text{ Mpc}$ . The very good agreement between this map and the target map shows that the coverage of the SEcat catalogue is excellent.

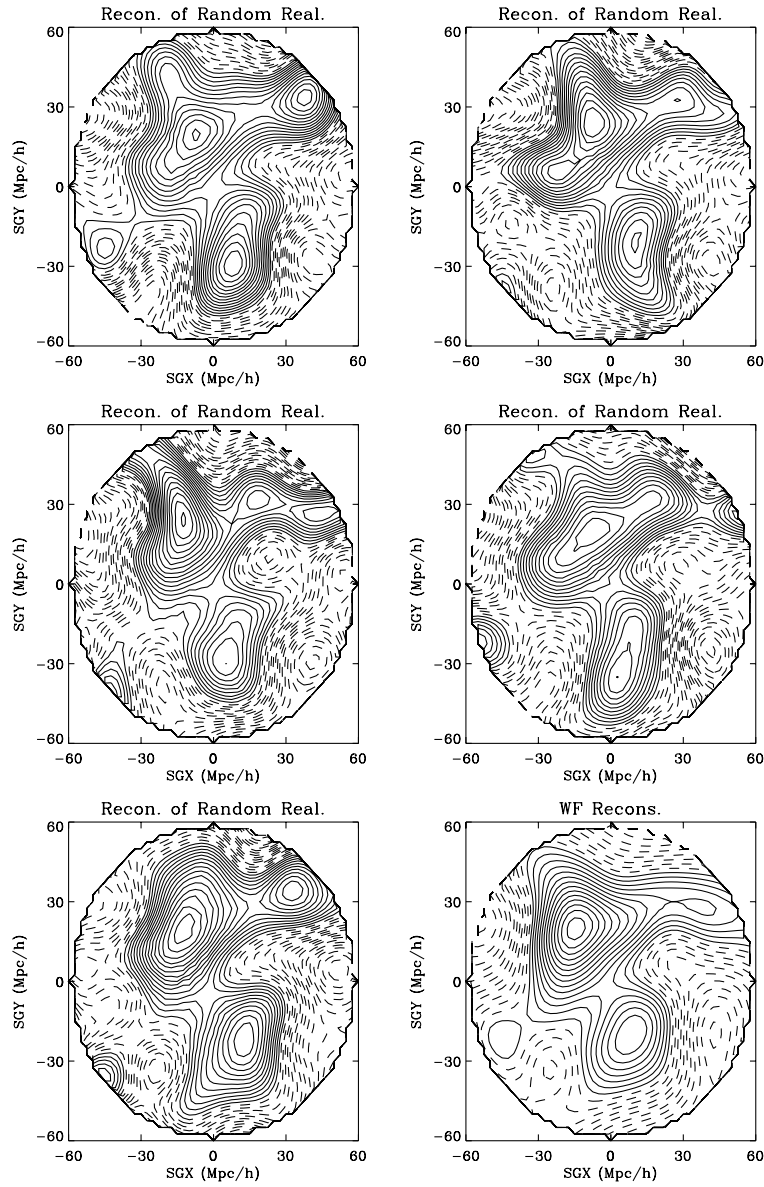
Next, we test the effect of noise on the reconstruction; here we construct a noisy mock SEcat catalogue where the added noise is realistic and corresponds to the quoted Tully–Fisher and  $D_n - \sigma$  errors in the real catalogue. The bottom-left panel shows the density reconstruction from a noisy mock realization of the SEcat catalogue. The agreement between this map and the target map is also good (see Fig. 6), although there are some areas where the recovered signal is not very satisfactory, especially towards the edge of the sphere.

In order to test where the recovered signal is reliable we reconstruct the density field from 30 mock catalogues, with the same underlying field but with different noise realizations, and compare them with the target map. The bottom-right panel shows

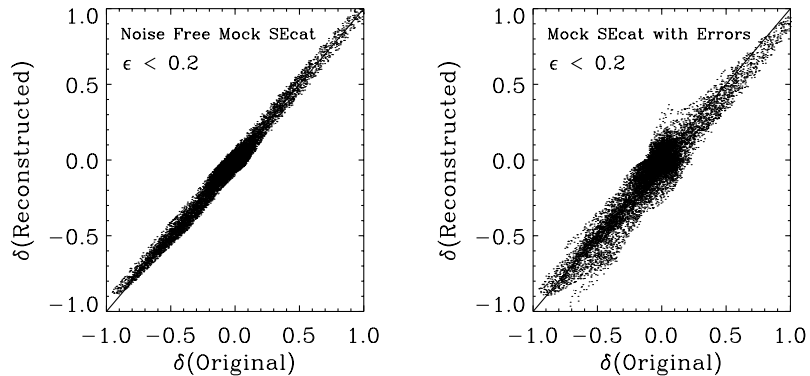
contours of the signal-to-noise ratio, with a spacing of 1 and with the heavy solid contour denoting a signal-to-noise ratio of unity. The signal-to-noise ratio in some areas in the map can be up to 15, and in most of the map the signal-to-noise ratio is greater than a few. However, if the underlying density is of order zero the signal-to-noise ratio gives a misleading impression about the quality of the reconstruction, as in this case it will be always of order zero. Therefore, the lower-right panel also shows the area (shaded) within which the error in the density contrast is less than 0.45.

To demonstrate the stability of the inversion, all the panels, with the exception of the bottom-right panel, in Fig. 5 show reconstructions similar to the ones shown in the bottom-left panel of Fig. 4 but with different error realizations. For comparison, the bottom-right panel in Fig. 5 shows the Wiener reconstructed density field from one of the noisy mock catalogues (the one used in the top left panel). Note that here the WF reconstruction roughly recovers the features in the target map; however, the amplitude of the density is suppressed throughout the plane, reflecting the biased nature of the Wiener reconstruction.

Fig. 6 shows a scatter plot of the original versus reconstructed densities within the whole reconstructed sphere. The densities are chosen from areas within which the errors, estimated from Monte Carlos of noisy mock SEcat catalogues, are less than 0.2. The left panel shows the quality of the reconstruction from a noise-free catalogue, while the right panel shows the reconstruction quality from a noisy catalogue. As expected, the scatter in the right panel is larger. The measured slope is slightly smaller than 1



**Figure 5.** Maps of density in the Supergalactic plane reconstructed from various Monte Carlo realizations of the errors. The smoothing window is a Gaussian of radius  $900 \text{ km s}^{-1}$ ,  $G9$ . The density contour spacing is 0.1; the mean  $\delta = 0$  contour is heavy; positive contours are solid and negative contours are dashed. The bottom-right panel shows the Supergalactic plane WF reconstruction from one of the Monte Carlo realizations.



**Figure 6.** A scatter plot of the underlying density versus the reconstructed density. The densities chosen are from areas within which the reconstruction error is  $< 0.2$ ; this error is determined using 30 mock SEcat catalogues. The left panel shows the quality of the reconstruction from a noise-free mock catalogue. The right panel shows the quality of the reconstruction from a noisy mock SEcat catalogue.

( $\approx 0.98 \pm 0.03$ , where the quoted uncertainty is the  $1\sigma$  error); nevertheless, the agreement between the original and reconstructed densities is excellent.

## 6 DISCUSSION AND SUMMARY

A general framework of linear estimation and prediction by minimal variance subject to a linear constraint in the data has been introduced. The solution of the minimization problem yields the UMV estimator, which is shown to be a very useful tool for reconstructing the large-scale structure of the Universe from incomplete, noisy and sparse data. The UMV estimator has been designed to overcome one of the main drawbacks of the WF, which is that it predicts the null field in the absence of good data; that is, in the limit of very poor signal-to-noise data the perturbation field is estimated to be zero. In contrast, the UMV estimator does not alter the values of the reconstructed field at the locations of the data points, because it lacks the filtering aspect of the WF; instead, it keeps the values of the measured data at the locations of the data points and interpolates between them in accordance with the correlation function assumed in the model. Like the WF, the new estimator can be used for dynamical reconstruction; that is, to reconstruct one dynamical field, for example over-density, from another measured field, for example radial peculiar velocity. These two properties make the UMV estimator a very appealing tool for various applications in LSS and CMB problems.

The regularization strength of the new estimator stems from the cross-correlation between the signal and the data points at different locations, which allows a reconstruction that is consistent with the data points and the correlations among them. Lacking such a correlation, the UMV estimator is equivalent to direct inversion. In this case, an additional regularization is required. This issue could be viewed as follows. The UMV produces an unbiased estimate of the underlying signal subject to the minimal variance requirement. This minimal variance could sometimes be too large to be useful. In comparison, the WF produces a *truly* minimal variance but with a biased mean. Therefore, when applying the UMV estimator, sometimes an extra regularization is needed in order to, on the one hand, reduce the variance in the UMV reconstruction to a manageable value, while keeping the reconstructed underlying signal as unbiased as possible, on the other. In this paper two extra regularization methods, which are both simple and easily integrated into the UMV method, have been discussed.

Constrained realizations of the underlying field (Hoffman & Ribak 1991) will not be possible with the UMV estimator. However, constrained realizations of a noisy underlying field are possible. Such noisy constrained realizations could be very useful if one wished to fill gaps in the data, for example from balloon-born CMB measurements, with a realization that has the same assumed properties as the signal + noise.

An apparent difficulty arises from the fact that the current UMV reconstruction assumes linear gravitational instability, yet it is applied to a universe that is *non-linear* on scales smaller than a few Megaparsecs. To obtain a non-linear reconstruction of the underlying field,  $s_{nl}$ , one can include the non-linear correlation by substituting  $\langle s_{nl}\theta^+ \rangle$  in equation (5). This of course only takes into account the contribution of the variance and ignores the contribution of higher moments, but in many cases this will do.

In one of the examples presented here, we have applied the method to galaxy radial peculiar velocity data and shown that the reconstruction is unbiased and trustworthy within a very large region of the volume covered by the data. In this context the UMV

reconstruction could be viewed as a compromise between the POTENT algorithm (Bertschinger & Dekel 1989; Dekel, Bertschinger & Faber 1990), which assumes no regularization, and the WF, which relies too heavily on it.

The UMV reconstruction of the over-density and the 3D velocity field from galaxy peculiar velocity catalogues is suitable for bias parameter extraction from a comparison with the respective fields obtained from redshift galaxy catalogues such as the PSCz (Saunders et al. 2000; Branchini et al. 2000). The current estimator allows density–density and velocity–velocity comparisons to be carried out using the same reconstruction technique (Zaroubi et al. 2001b).

## ACKNOWLEDGMENTS

I would like to thank A. J. Banday, E. Branchini, Y. Hoffman and H. Mo for detailed comments on the manuscript, and L. N. da Costa, O. Lahav, T. Theuns, and A. Nusser for discussions. I also thank the anonymous referee whose excellent comments led to significant improvements in the paper.

## REFERENCES

- Bertschinger E., Dekel A., 1989, ApJ, 336, L5  
 Bond J. R., 1995, Phys. Rev. Lett., 74, 4369  
 Bouchet F. R., Gispert R., 1999, New Astron., 4, 443  
 Branchini E. et al., 2000, MNRAS, 308, 1  
 Bunn E., Fisher K. B., Hoffman Y., Lahav O., Silk J., Zaroubi S., 1994, ApJ, 432, L75  
 da Costa L. N., Bernardi M., Alonso M. V., Wegner G., Willmer C. N. A., Pellegrini P. S., Rit  C., Maia M. A. G., 2000, AJ, 120, 95  
 Dekel A., Bertschinger E., Faber S. M., 1990, ApJ, 364, 349  
 Fisher K. B., Lahav O., Hoffman Y., Lynden-Bell D., Zaroubi S., 1995, MNRAS, 272, 885  
 Giovanelli R., Haynes M. P., Freudling W., da Costa L. N., Salzer J. J., Wegner G., 1998, ApJ, 505, L91  
 G rski K. M., 1988, ApJ, 332, L7  
 Gull S. F., 1989, in Skilling J., ed., Maximum Entropy and Bayesian Methods. Kluwer, Dordrecht, p. 53  
 Hoffman Y., Eldar A., Zaroubi S., Dekel A., 2000, ApJ, submitted  
 Hoffman Y., Ribak E., 1991, ApJ, 380, L5  
 Lahav O., Fisher K. B., Hoffman Y., Scharf C. A., Zaroubi S., 1994, ApJ, 423, L93  
 Pi a R. K., Puetter R. C., 1993, PASP, 105, 630  
 Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes, 2nd edn. Cambridge Univ. Press, Cambridge  
 Rybicki G. B., Press W. H., 1992, ApJ, 398, 169  
 Saunders W. et al., 2000, MNRAS, 317, 55  
 Schmoltdt I. M. et al., 2000, AJ, 118, 1146  
 Tegmark M., Efstathiou G., 1996, 281, 1297  
 Wiener N., 1949, Extrapolation and Smoothing of Stationary Time Series. Wiley, New York  
 Zaroubi S., 2000, in Kraan-Korteweg R. C., Henning P. A., Andernach H., eds, ASP Conf. Ser. Vol. 218, Proc. of Mapping the Hidden Universe: The Universe Behind the Milky Way – The Universe in H I. Astron. Soc. Pac., San Francisco, p. 173  
 Zaroubi S., Hoffman Y., Fisher K. B., Lahav O., 1995, ApJ, 449, 446  
 Zaroubi S., Hoffman Y., Dekel A., 1999, ApJ, 520, 413  
 Zaroubi S., Bernardi M., da Costa L. N., Hoffman Y., Alonso M. V., Wegner G., Willmer C. N. A., Pellegrini P. S., 2001a, MNRAS, 326, 375  
 Zaroubi S., Branchini E., Hoffman Y., da Costa L. N., 2001b, MNRAS, submitted

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.