Geophysical Journal International

Geophys. J. Int. (2012) **190,** 1423–1438 doi: 10.1111/j.1365-246X.2012.05569.x

Robust principal component analysis of electromagnetic arrays with missing data

M. Yu. Smirnov[1] and G. D. Egbert[2]

[1] *Department of Physics, University of Oulu, Oulu, Finland. E-mail: maxim.smirnov@oulu.fi*
[2] *College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, OR* 97330*, USA*

Accepted 2012 June 7. Received 2012 June 2; in original form 2011 September 29

SUMMARY

We describe a new algorithm for robust principal component analysis (PCA) of electromagnetic (EM) array data, extending previously developed multivariate methods to include arrays with large data gaps, and only partial overlap between site occupations. Our approach is based on a criss-cross regression scheme in which polarization parameters and spatial modes are alternately estimated with robust regression procedures. The basic scheme can be viewed as an expectation robust (ER) algorithm, of the sort that has been widely discussed in the statistical literature in the context of robust PCA, but with details of the scheme tailored to the physical specifics of EM array observations. We have tested our algorithm with synthetic and real data, including data denial experiments where we have created artificial gaps, and compared results obtained with full and incomplete data arrays. These tests reveal that for modest amounts of missing data (up to 20 per cent or so) the algorithm performs well, reproducing essentially the same dominant spatial modes that would be obtained from analysis of the complete array. The algorithm thus makes multivariate analysis practical for the first time for large heterogeneous arrays, as we illustrate by application to two different EM arrays.

Key words: Time series analysis; Magnetotelluric; Geomagnetic induction.

1 INTRODUCTION

With advances in electromagnetic (EM) geophysical instrumentation large synoptic digital EM array data sets are increasingly common, begging the development of practical analysis methods that can fully exploit the simultaneous character of these data. As discussed in Egbert (2002) one can apply multivariate statistical methods to EM array data to reduce bias, improve signal-to-noise ratios (SNR) and provide better control over source effects and coherent noise contamination in estimates of EM transfer functions (TFs). Furthermore, within a multisite array framework one is not limited to the classical uniform source magnetotelluric (MT) impedance and geomagnetic vertical field TFs, but can also include interstation horizontal magnetic TFs which can be used to map anomalous induced currents, as well as 'hybrid' impedance tensors between electric and magnetic fields from any pair of stations. The inclusion of such interstation TFs, which arise quite naturally and are easy to compute within a multivariate statistical framework, may provide potentially useful additional constraints on the Earth's conductivity.

Simultaneous arrays also provide a powerful tool for detecting and understanding effects of finite spatial scale sources in TF estimates, whether due to cultural noise (Larsen *et al.* 1996; Egbert 1997; Ritter & Banks 1998), or spatial complexity of natural sources (Egbert *et al.* 2000; Egbert 2002). At least in some cases this understanding can be translated to improved estimates of TFs (e.g. Qian & Pedersen 1991). At the same time, inhomogeneous components of the natural source field might be used to probe the Earth's conductivity structure in a different manner. A natural extension to the usual uniform source assumption implicit in the MT method allows for three curl-free magnetic gradient sources (Kuckes *et al.* 1985; Egbert & Booker 1989; Egbert 2002; Vozar & Semenov 2010). Quantitative interpretation of the response of a 3-D earth to these more spatially complex sources has the potential to provide valuable additional constraints on large-scale variations in crustal and upper-mantle conductivity. For example the horizontal spatial gradients (HSG) technique (Schmucker 2003) uses only magnetic field components to estimate a TE-mode impedance tensor, which should thus be free of galvanic distortions. Adding HSG responses as additional constraints might help to overcome problems due to aliasing of near surface distortion in 3-D inversion of widely spaced MT data, such as that collected in the US Earthscope project (Egbert *et al.* 2007; Patro & Egbert 2008).

Egbert & Booker (1989) and Egbert (1997) described the basic multivariate model and estimation methods that we consider here. However, in these early efforts only data from small (3–5 stations) EM arrays were available, and the small number of simultaneous sites (which in some cases were in a profile, rather than a 2-D array) severely limited opportunities for characterization

and separation of source components. Egbert (2002) applied these methods to data from the much larger EMSLAB magnetometer array, which consisted of 64 sites in a quasi-regular 2-D array covering much of the Pacific Northwest United States (Gough *et al.* 1989). However these data were collected with old analogue instruments, and only 60 hr of digitized data were available for the multivariate analysis. A number of large digital EM arrays are now available, including the EarthScope 'rolling' MT array (Patro & Egbert 2008), and the Scandinavian Baltic Electromagnetic Array Research (BEAR; Korja *et al.* 2002) and Electromagnetic Mini Array (EMMA; Smirnov *et al.* 2008) arrays, some of which we consider further. These data sets provide new opportunities to test and further develop multivariate techniques on this sort of large digital array, which can be expected to be ever more common in future.

Another potentially important application of the multivariate analysis approach considered here is to geomagnetic observatory data (e.g. Fujii & Schultz 2002). For many years the global network of geomagnetic observatories has been the backbone of studies concerning the sources of magnetic fields in the Earth's core, ionosphere and magnetosphere, as well as induction by external source variations. For these very long-period global induction studies there are few reliable measurements of the surface electric fields, and inferences of internal conductivity must be based solely on magnetic field data. This requires relatively strong assumptions about external source geometry (e.g. Banks 1969; Olsen 1998), limiting application to periods and locations where highly idealized external sources may plausibly be assumed, for example, a zonal ring current at periods longer than 5 d at geomagnetic mid-latitudes (Banks 1969; Fujii & Schultz 2002), or where site density is sufficient to use HSG methods (Olsen 1998). Multivariate methods may offer a path to better models of external sources, covering a broader range of periods and geographic locations. Recent studies of deep conductivity structure of the Earth (Kelbert *et al.* 2009), which suggest large lateral variations in transition zone conductivities associated with subduction of fluids, demonstrate the importance of improved source models for probing 3-D conductivity variations in the Earth.

Missing data, and more generally temporal and spatial heterogeneity of noise, present major challenges to processing EM data from a large array. The robust multivariate methods for EM data developed by Egbert (1997) do not explicitly allow for missing data, forcing the analysis to be restricted to times when all sites operate. This simple strategy becomes increasingly untenable as the size of the array increases. For example, if this approach is applied to the BEAR data set considered later, analysis would have to be restricted to a much reduced array, consisting of fewer than half the sites, operating for perhaps a week or so. Clearly, analysis of such large arrays requires a more sophisticated strategy for dealing with missing data.

Our focus here is on extending the approach for treating outliers developed in Egbert (1997) to allow estimation of principal components (i.e. array spatial modes) for large incomplete arrays. The diverse applications of principal component analysis (PCA) discussed earlier, for example, characterizing and modelling global source structure, novel TF or HSG applications, estimation of MT impedances, are beyond the scope of this paper. In the next section we review the statistical model, and outline the general approach to robust estimation for this model allowing for missing data. The estimation algorithm is presented in detail in Section 3, examples with synthetic and real data are presented in Section 4 and concluding remarks are given in Section 5.

2 THE STATISTICAL MODEL

The multivariate analysis scheme for synchronous EM data proposed by Egbert & Booker (1989) can be viewed as PCA of complex frequency domain data vectors, obtained from a windowed Fourier transform (WFT) of observed time-series from all channels/sites in the array. Robust approaches to PCA are by now rather well developed (e.g. Stanimirova *et al.* 2007; Serneels & Verdonck 2008; Frahm & Jaekel 2010). We briefly review a few of the key ideas, but with the notation and nomenclature used in the statistical literature modified to be consistent with the discussion in Egbert (1997), and with all variables taken to be complex. The idea of PCA is to derive a low-dimensional projection that contains the maximal amount of variation in the (high-dimension) data vectors. Mathematically the principal components that define this projection are obtained according to some variance maximization criterion. Let $\mathbf{X} \in \mathbf{C}^{N \times J}$ be the data matrix, consisting of J replicates of a multivariate frequency domain observation, each consisting of N complex variables, that is, distinct variables are in different rows and correspond to different data channels, while each column represents one instance (e.g. windowed time segment) of the multivariate observation. Then the principal components $\mathbf{U}_k$ are defined as linear combinations of the data

$$\mathbf{U}_k = \mathbf{X}\mathbf{a}_k^*, \tag{1}$$

that maximizes some sort of variance

$$\mathbf{a}_k = \arg\max_{\mathbf{p}} \ \mathrm{var}\{\mathbf{X}\mathbf{p}^*\} \tag{2}$$

under the constraints that $\|\mathbf{a}_k\| = 1$ and $\mathrm{cov}(\mathbf{U}_k, \mathbf{U}_{k'}) = 0$ for $k' \neq k$. Here, and subsequently, * denotes the Hermitian complex conjugate transpose. In classical PCA $\mathrm{var}(\mathbf{X}\mathbf{p}^*) = \|\mathbf{X}\mathbf{p}^*\|^2$, and exact maximization of (2) can be accomplished by the Lagrange multiplier method, leading to the conclusion that the principal components are the eigenvectors of the variance–covariance matrix $\mathbf{\Sigma} = \frac{1}{N}\mathbf{X}^*\mathbf{X}$ (Serneels & Verdonck 2008). Note that, consistent with the terminology of frequency domain time-series analysis, Egbert (1997) refers to $\mathbf{\Sigma}$ as the spectral density matrix, or SDM.

Modern robust PCA approaches are typically based on replacing the variance operator in (2) with some robust estimate of scale. Several distinct approaches have been proposed and extensively tested over the last years, including methods based on projection pursuit (e.g. Croux *et al.* 2007), methods using an eigenvector decomposition of a robust estimate of the covariance matrix (e.g. Hubert *et al.* 2005), or some combination of these (e.g. Verboven & Hubert 2005). In the projection pursuit approach, one projects data onto a lower dimensional space such that a robust measure of variance of the projected data is maximized. The affine invariant covariance estimate (Huber 1981), and the minimum covariance determinant (MCD) estimator (Verboven & Hubert 2010) are good examples of a robust covariance estimator which has been used for robust PCA. Both of these general approaches to robust PCA have been implemented in the Libra statistical library (Verboven & Hubert 2005), where projection pursuit is realized as RAPCA and the MCD estimator is used in ROBPCA. We consider application of these procedures to our problem later. Another approach based on the projection pursuit algorithm developed by Croux *et al.* (2007, refereed to as the C–R algorithm) is realized in the Matlab Toolbox TOMCAT (Daszykowski *et al.* 2007), and could also be readily accommodated within the framework of algorithms discussed here.

Egbert & Booker (1989) introduced use of PCA for EM array data, but also considered a closely related statistical model, the

multivariate-errors-in-variable (MEV) model (e.g. Gleser 1981)

$$X_{nj} = \sum_{k=1}^{K} U_{nk} a_{kj} + e_{nj}. \tag{3}$$

Here X_{nj} are frequency domain data for channels $n = 1, N$, and time windows $j = 1, J$; U_{nk} define the components (for channel n) of spatial modes $k = 1, \ldots, K$; a_{kj} are generalized 'polarization parameters' giving the coefficients of spatial mode k for window j. In the statistical literature on PCA the elements U_{nk} are generally referred to as the loadings, the vectors $\mathbf{U}_k$ as the principal components, and the coefficients a_{kj} are called the scores. The MEV model explicitly allows for noise, which we represent in (3) as e_{nj}. We assume this noise is incoherent (i.e. uncorrelated between channels, and between time segments), but allow for variations in amplitude between channels

$$\mathrm{cov}(e_{nj}, e_{n'j'}) = \sigma_n^2 \delta_{nn'} \delta_{jj'}. \tag{4}$$

Most PCA algorithms require data standardization, that is, shifting and normalizing each variable to zero mean and unit variance, using robust estimates of scale and location. In the context of the MEV model (3), the incoherent noise standard deviation σ_n should instead be used to normalize each variable to ensure optimal and unbiased results (Egbert 1997). A procedure for robustly estimating these additional parameters must thus be integrated into our analysis scheme. Because the complex Fourier coefficients intrinsically have randomized phases (in general there is no deterministic or systematic alignment of signals and the time windows used for the Fourier transform) data means are already effectively zero.

The MEV model of eq. (3) can be summarized in matrix notation as

$$\mathbf{X} = [\mathbf{X}_1 \cdots \mathbf{X}_J] = \mathbf{U}\,[\mathbf{a}_1 \cdots \mathbf{a}_J] + [\mathbf{e}_1 \cdots \mathbf{e}_J] = \mathbf{UA} + \mathbf{E}, \tag{5}$$

while (4) can be expressed as

$$\mathrm{cov}(\mathbf{e}_j, \mathbf{e}_{j'}) = \delta_{jj'} \mathbf{C_N}, \tag{6}$$

where $\mathbf{C_N}$ is the diagonal matrix of channel incoherent noise variances σ_n^2.

As discussed in Egbert & Booker (1989) the representation of (3) and (5) is not unique, since $\mathbf{UA}$ can be replaced by $\tilde{\mathbf{U}}\tilde{\mathbf{A}}$ where $\tilde{\mathbf{A}} = \mathbf{BA}$ and $\tilde{\mathbf{U}} = \mathbf{UB}^{-1}$, with $\mathbf{B}$ any non-singular $K \times K$ matrix. However, adopting the conventions of PCA makes the representation nearly unique: the columns of $\mathbf{U}$ can be taken to be the principal components or loadings, and then the polarization parameters in the matrix $\mathbf{A}$ are just the scores. Any representation of the form $\tilde{\mathbf{U}}\tilde{\mathbf{A}}$ can be converted to one satisfying the PCA convention by truncating the singular value decomposition (SVD) of the matrix $\tilde{\mathbf{U}}\tilde{\mathbf{A}} = \mathbf{USV}^*$ to include only the first K columns of $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{S}$ (corresponding to the K non-zero singular values of $\tilde{\mathbf{U}}\tilde{\mathbf{A}}$). Then $\mathbf{U}$ gives the spatial modes, which are orthonormal $\mathbf{U}^*\mathbf{U} = \mathbf{I}$, and the polarization parameters satisfy

$$\mathbf{A} = \mathbf{SV}^*, \qquad \mathbf{V}^*\mathbf{V} = \mathbf{I}, \qquad \mathbf{S} = \mathbf{diag}\,(\mathbf{s_1}, \ldots, \mathbf{s_K}), \tag{7}$$

with positive decreasing entries on the diagonal ($s_1 \geq \cdots \geq s_K$), as for the definition of the ordered principal components. Note that even with this convention the parametrization is not strictly unique, since the complex-valued columns $\mathbf{u}_k$ and $\mathbf{v}_k$ are still only defined up to an arbitrary phase factor: multiplying corresponding columns of $\mathbf{U}$ and $\mathbf{V}$ by unit magnitude complex numbers $e^{i\phi_k}$, $k = 1, \ldots, K$ would still result in the same product matrix $\mathbf{UA}$, if $\mathbf{A}$ is defined by (7).

Our principal focus here is on estimation of $\mathbf{U}$ and $\mathbf{A}$ in the model (3), when there is a significant amount of missing data. One approach to dealing with missing data in PCA is based on the expectation maximization algorithm (EMax; Dempster *et al.* 1977). The EMax algorithm is an iterative scheme which alternates between two steps: (i) missing elements are filled in by expected values, computed based on the most recent value for the parameter estimates (the expectation step or E-step) and (ii) the model parameters (in the case of PCA, the loadings and scores) are estimated using the filled-in data matrix (the maximization step or M-step). When a robust estimation technique is used for the second step this is sometimes called the R-step, and the overall iterative scheme is referred to as an ER algorithm. In either case, the true values of the missing elements are unknown, so the procedure must be initialized somehow, and then repeated until some convergence criterion is fulfilled. Both the EMax and ER algorithms have been applied to PCA with missing data (e.g. Serneels & Verdonck 2008) .

A slightly different approach to PCA for incomplete data matrices has been proposed in the context of image processing by De La Torre & Black (2003). This approach, which can itself be viewed as a special case of the EMax algorithm is based on 'criss-cross regression' (Gabriel & Zamir 1979). The basic idea here is that if one had a preliminary estimate of $\mathbf{U}$, then (3) is a linear model, and $\mathbf{A}$ can thus be estimated by linear regression. Conversely, once $\mathbf{A}$ has been estimated, $\mathbf{U}$ can be estimated in the same simple way. Thus, if some initial estimate of either set of parameters can be provided, these steps can be alternated to convergence. This scheme can be made robust by replacing the least-squares regression steps with a robust alternative (De La Torre & Black 2003). For example, Egbert (1997) used a regression M-estimate in an application of this iterative estimation scheme to EM array data. Because there is no obstacle to applying each of the linear problem steps with irregular data arrays, the criss-cross regression scheme readily generalizes to allow for missing data. This simple idea, augmented with some of the other methods for robust PCA discussed earlier, forms the basis for the robust multivariate analysis scheme we have developed for EM arrays, which we refer to as MsDEMPCA (missing data EM PCA).

3 ESTIMATION SCHEME

Our implementation of MsDEMPCA is based on the general ideas of robust PCA discussed earlier, tailored to characteristics specific to large EM arrays. We begin with an overview of the scheme, which is summarized in Fig. 1, before providing specific details on its component parts.

The first processing step, which we do not discuss in detail here, is to use a WFT (or equivalent scheme) to convert time-series for all channels at all sites into a set of complex frequency domain 'data matrices', each corresponding to a different frequency band. MsDEMPCA is applied to each frequency band independently, so we consider further only a single complex $N \times J$ data matrix $\mathbf{X}$. Each of the J columns of $\mathbf{X}$ is an N-dimensional complex data vector corresponding to Fourier coefficients for all channels for a fixed frequency/time window. For example, for an array of N_s five-channel MT sites $N = 5N_s$; if the frequency band consisted of two adjacent Fourier harmonics and the total number of windowed segments were N_w we would have $J = 2N_w$. Note that we typically use longer time windows (and decimated time-series) for lower frequency bands, as, for example, in Egbert (1997), so J typically

Compute WFT $\Rightarrow \mathbf{X}(N, J)$ = complex data matrix

Initialization

Select core sub-array $\Rightarrow \mathbf{X}(N', J')$, $N' < N$, $J' < J$ (Sec. 3.5)
Initial estimate of $\mathbf{U}$ for core array (Sec. 3.1)
Compute $\mathbf{C_N}$, clean data (Sec. 3.4)
Re-estimate $\mathbf{U}$ using cleaned data, $\mathbf{C_N}$ (Sec. 3.1)
Set U_Estimated = True

Outer Loop: Building up full array from core (Sec. 3.6)

While $N' < N$ or $J' < J$ (array not complete)
Execute Inner Loop
Increase # active segments J', keeping N' fixed
Execute Inner Loop
Increase # active channels N', keeping J' fixed
Set U_Estimated = False
Execute Inner Loop

Inner Loop (For fixed sub-array):

While not converged
If U_Estimated
Compute $\mathbf{C_N}$, clean data (Sec. 3.4)
Estimate $\mathbf{A}$, with $\mathbf{U}$ fixed (Sec. 3.3)
Estimate $\mathbf{U}$, with $\mathbf{A}$ fixed (Sec. 3.2)
Set U_Estimated = True

Figure 1. Summary of robust multivariate array analysis scheme. The inner loop implements the ER algorithm for a fixed array. The outer loop is used to build the full array up from an initial smaller core array, for which only a small fraction of data are missing. The inner loop can also be used alone when a small fraction of data are missing.

decreases with frequency. Note also that $N \times J$ represents the nominal size of the data matrix; many of the entries in the matrix will in general be missing.

The core of MsDEMPCA (inner loop in Fig. 1) is based on the criss-cross regression scheme, with alternating robust estimates of the polarization parameters $\mathbf{A}$ and the spatial modes $\mathbf{U}$, augmented with estimates of the channel incoherent noise variances $\mathbf{C}_\mathrm{N}$. These three steps are described in Sections 3.2–3.4. If the fraction of missing data is not too great (e.g. a few tens of per cent total, with no long gaps at any sites), initial estimates of $\mathbf{U}$ and/or $\mathbf{A}$ (required to start the criss-cross regression) can be obtained by setting any missing values to zero, and then applying the robust SVD scheme described in Section 3.1. However, this simple scheme may fail, for instance when there are individual sites with more extensive data gaps. To handle this case we first reduce to a 'core' array (Section 3.5) consisting of a subset of all sites, operating over a possibly shorter time interval, for which the fraction of missing data is small enough for the simple inner loop scheme to be effective. Once we have initial estimates of spatial modes and polarization vectors, restricted to the core sites and time intervals, the array can then be built up to ultimately include all sites and time segments, a process described in Section 3.6, and represented by the outer loop in Fig. 1.

Before proceeding two general comments are in order. First, the number of modes estimated, denoted by K, may not be known *a priori*. Obviously $L = \min(J, N)$ sets an upper bound on the number of estimable modes, and in fact we cannot expect statistically meaningful results for all modes unless $K \ll L$. Ideally, if the data set is large enough, we can set K to a large value, and then estimate the number of statistically significant coherent modes, for example, following ideas discussed in Egbert (1997). However, for very large arrays (such as the BEAR example considered later) the number of actual coherent modes may well exceed the number that may be reliably estimated, given the available number of segments J. Experiments with a range of values for K suggest that a set of dominant modes $\mathbf{U}_1, \ldots \mathbf{U}_k$ (associated with the k largest variances) are insensitive to the truncation level provided k is not comparable to K. Higher modes (especially as k approaches K), as well as estimates of incoherent noise variance, are more sensitive to the choice of K. For the arrays discussed later we have focused on obtaining estimates for five to six modes (which carry the bulk of the signal for these arrays), and set $K = 10$. For other applications, such as the global array of geomagnetic observatories, other values of K may be appropriate. Secondly, in our discussion later we frequently use the superscript (i) to denote inner-loop step or iteration number, with the convention that if i corresponds to a spatial mode ($\mathbf{U}$) estimation step, $i + 1$ will denote a polarization parameter ($\mathbf{A}$) step. Thus a full iteration of the inner loop increments i by 2.

The algorithm described here has been realized in MATLAB using object-oriented programming. The framework allows for easy modification, development and extensions of the code. Separate classes are coded for spectral analysis, PCA, plotting and subsequent post-processing of the principal components. With the object-oriented approach it is easy to modify or replace individual components of the overall scheme. For example, we have experimented with several approaches for robust regression, and for robust covariance estimation. In the following subsections we focus on our basic 'default' implementations of individual algorithmic steps, beginning with the robust covariance estimator used for initialization of the criss-cross regression, followed by a description of the alternating estimation steps for $\mathbf{U}$, $\mathbf{A}$ and $\mathbf{C}_\mathrm{N}$—all that would be required for an array with only a small amount of missing data (inner loop alone). In the last two subsections we consider extension to the general case: first describing our procedure for stripping down to a core array with minimal missing data, and then our approach to building the array back up to obtain estimates for the full set of sites, using as many time segments as possible.

3.1 Robust covariance estimation

If the data matrix is complete (or has been completed by setting a small fraction of missing data to zero) estimates of $\mathbf{U}$ and $\mathbf{A}$ can be obtained from the eigenvector decomposition of the normalized SDM or generalized SVD of $\mathbf{X}$. As in Egbert (1997) this estimate can be made more robust by applying the affinely invariant robust covariance matrix estimate of Huber (1981). Here we adopt a slightly modified approach and work directly with the SVD of the data matrix. In this form the estimator extends easily to the case where there are more channels than segments ($N > J$; so that the sample SDM is not of full rank). This extension is especially useful for the large arrays considered here.

Normalize: $\mathbf{Z} = \mathbf{C_N}^{-1/2}\mathbf{X}$

Compute initial SVD: $\mathbf{Z} = \mathbf{u}^{(0)}\mathbf{s}^{(0)}[\mathbf{v}^{(0)}]^*$

(with $L = min(N, J)$ defining size of $\mathbf{s}^{(0)}$)

Initialize weights: $\mathbf{W}^{(0)} = \mathbf{I}$

Loop over i:

$\mathbf{Y}^{(i)} = \left[\mathbf{y}_1^{(i)} \cdots \mathbf{y}_J^{(i)}\right] = [\mathbf{s}^{(i)}]^{-1}[\mathbf{u}^{(i)}]^*\mathbf{Z}$

$r_j^{(i)} = \left\|\mathbf{y}_j^{(i)}\right\| J/L$

$w_j^{(i)} = \begin{cases} 1 & \text{if } r_j^{(i)} < r_0 \\ r_0/r_j^{(i)} & \text{if } r_j^{(i)} \geq r_0 \end{cases}$

$\mathbf{W}^{(i)} = \mathbf{diag}(w_1, ..., w_J)$

compute SVD of $\mathbf{Y}^{(i)}\mathbf{W}^{(i)} = \mathbf{u}^{(i+1)}\mathbf{s}^{(i+1)}[\mathbf{v}^{(i+1)}]^*$

if $\left\|\mathbf{u}^{(i+1)}\mathbf{s}^{(i+1)}\mathbf{s}^{(i+1)}[\mathbf{u}^{(i+1)}]^* - \mathbf{I}\right\| < \epsilon$ break

Using final weights, compute SVD of $\mathbf{ZW} = \mathbf{USV}^*$

Figure 2. Affinely invariant robust covariance estimate of Huber (1981), slightly modified and recast in terms of the SVD, instead of an eigenvector decomposition. $L = \min(N, J)$ is the rank of the data matrix, and in the SVD $\mathbf{U}$ is $N \times L$, $\mathbf{S}$ is $L \times L$ diagonal and $\mathbf{V}$ is $J \times L$. The computation of $\mathbf{Y}$ can be stabilized by adding a small constant to the singular values on the diagonal of $\mathbf{s}^{(i)}$.

The basic idea is to estimate $\mathbf{U}$ from the generalized SVD of the scaled data matrix

$$\mathbf{C_N}^{-1/2}\mathbf{XW} = \mathbf{Z} = \tilde{\mathbf{U}}\mathbf{SV}^*. \tag{8}$$

Multiplication of $\mathbf{X}$ on the left-hand side by $\mathbf{C_N}^{-1/2}$ rescales each channel by the inverse of the estimated incoherent noise standard deviation, so that all channels are non-dimensional, and noise levels are uniform across channels. In (8), $\mathbf{W} = \mathbf{diag}(w_1, \ldots, w_J)$ is a diagonal matrix of segment weights ($0 < w_j < 1$). These are determined iteratively, as in the affinely invariant covariance estimate described in Egbert (1997), and as outlined in the pseudo-code in Fig. 2. Note that because the channels are scaled by incoherent noise variance, the components of $\tilde{\mathbf{U}}$ and the singular values are non-dimensional. The singular values can be interpreted as the SNR (amplitude), with values significantly above one indicative of coherent signal. The singular value spectrum can thus guide the choice of truncation level K. The transformation $\mathbf{U} = \mathbf{C_N}^{1/2}\tilde{\mathbf{U}}$ converts the spatial modes back to physical units, which is required for physical interpretation of the principal components.

For the initial call to the robust covariance estimator, we generally will not yet have estimates of $\mathbf{C_N}$. In this case we use the individual channel variances for scaling (so the initial principal components are based on the correlation matrix). Then, as outlined in Fig. 1, we compute estimates of incoherent noise variances (see Section 3.4), and repeat the robust covariance estimator using $\mathbf{C_N}$. Note also that we can obtain initial estimates of $\mathbf{A}$ from this step: essentially $\mathbf{A} = \mathbf{SV}^*$.

We have also experimented with alternative robust PCA algorithms from the Libra statistical library (Verboven & Hubert 2005). In particular, ROBPCA, which is based on a fast projection pursuit algorithm, appears to be a promising alternative to the affinely invariant covariance estimator.

3.2 Estimation of the spatial modes

We assume we have an estimate of the polarization parameters $a_{kj}^{(i-1)}$, $k = 1, K; j = 1, J$ (e.g. from iteration $i - 1$, or from the robust covariance estimate of the previous section). Treating these parameters as fixed and known, (3) represents N decoupled linear statistical models, one for each data channel. For each fixed n we can thus compute refined estimates $\hat{U}_{nk}^{(i)}$, $k = 1, K$ by applying a robust regression estimator to (3), omitting any segments j for which X_{nj} is missing. We have primarily used the regression M-estimate (RME; Huber 1981), but within our object-oriented framework alternative robust regression procedures [e.g. the repeated medians scheme used by Smirnov (2003)] could quite easily replace the M-estimate.

The RME can be viewed as an iterative weighted LS estimate, with the weights for segment j determined from the magnitude of the residuals $w_{nj} = w(|X_{nj} - \hat{X}_{nj}|)$, where the predicted data are given by

$$\hat{X}_{nj} = \sum_{k=1}^{K} U_{nk}^{(i)} a_{kj}^{(i-1)}. \tag{9}$$

In our implementation of the RME we have used a Huber loss function, for which the weights take the form

$$w(r) = \begin{cases} 1 & |r| < \hat{\sigma}_n r_0 \\ (\hat{\sigma}_n r_0)/|r| & |r| \geq \hat{\sigma}_n r_0 \end{cases}, \tag{10}$$

where $\hat{\sigma}_n^2$ is a robust estimate of the error variance for channel n. Because the weights depend on the residuals (and these depend, through the predicted data, on the weights) the weights must be computed through an iterative procedure.

The RME can also be viewed as an unweighted LS estimate, with the raw data 'cleaned' by pulling observations with large residuals towards the predicted data

$$\tilde{X}_{nj} = \hat{X}_{nj} + w_{nj}(X_{nj} - \hat{X}_{nj}). \tag{11}$$

The weights in (11) are also given by (10). The cleaned data computed in the iteration i update of the spatial mode estimates are denoted as $\tilde{\mathbf{X}}^{(i)}$, and are saved for use in subsequent processing steps. For observations X_{nj} that are missing, the prediction $\hat{X}_{nj}$ can be saved in the cleaned data array.

Note that the estimates for each channel are computed independently. Thus, estimates of the spatial mode vectors $\hat{\mathbf{U}}_k$ assembled from these estimated components will not in general be orthonormal. As noted earlier we can orthogonalize the spatial modes, and also ensure that the polarization parameters are consistent with the PCA convention of (7), by computing the SVD of the predicted data matrix

$$\hat{\mathbf{X}}^{(i)} = \hat{\mathbf{U}}^{(i)}\mathbf{A}^{(i-1)} = \mathbf{U}^{(i)}\,[\mathbf{SV}] = \mathbf{U}^{(i)}\mathbf{A}^{(i)}. \tag{12}$$

Note that the spans of the columns of $\mathbf{U}^{(i)}$ and $\hat{\mathbf{U}}^{(i)}$ are identical.

3.3 Estimation of polarization parameters

For this step we assume we have an estimate of the spatial modes $U_{kn}^{(i)}$, $k = 1, K; n = 1, N$, which are now to be treated as fixed and known. Then (3) represents J decoupled linear statistical models, that is, for each time segment j we have

$$\mathbf{P}_j\mathbf{X}_j = \mathbf{P}_j\mathbf{U}^{(i)}\mathbf{a}_j + \mathbf{P}_j\mathbf{e}_j, \tag{13}$$

where $\mathbf{P}_j$ is the $N_j \times N$ matrix which selects the $N_j \leq N$ channels which are available for this segment. If no data are missing for segment j, $\mathbf{P}_j$ is the identity, the design matrix for the linear model of (13) reduces to the unitary matrix $\mathbf{U}^{(i)}$, and the least-squares estimate will simply be

$$\mathbf{a}_j^{(i+1)} = [\mathbf{U}^{(i)}]^*\mathbf{X}_j. \tag{14}$$

In the more general case with missing data the design matrix will not be unitary, and, if the fraction of missing data is large, it may even be poorly conditioned. Moreover, if the number of channels exceeds the available number of data segments the problem would be undetermined. We thus use a regularized or damped LS approach to obtain stable estimates of the polarization vectors $\mathbf{a}_j$.

In fact, within the framework of our iterative scheme, we will generally have reasonably good prior information about the statistics of both the data errors and the model parameters in (13), and these can be used to justify a sensible regularization. In the notation used earlier, the covariance of data errors can be estimated (using the incoherent noise variance estimates) as $\mathbf{C}_d = \mathbf{P}_j \mathbf{C}_\mathbf{N} \mathbf{P}_j^*$, while the covariance of model parameters ($\mathbf{a}_j$) can be estimated from the statistics of all polarization parameters estimated from the previous iteration (or initially from the robust covariance estimate). Because the spatial modes $\mathbf{U}$ are orthonormal this covariance is simply $\mathbf{C}_m = \mathbf{S}^2/J$, where $\mathbf{S}$ is the diagonal matrix of singular values of the predicted data matrix $\hat{\mathbf{X}}$, and J the number of available time segments. For simplicity in the following we use the notation $\Psi_j = \mathbf{P}_j \mathbf{U}^{(i)}$ and $\Xi_j = \mathbf{P}_j \mathbf{X}_j$, and, in the usual way for weighted damped LS (Menke 1989), we minimize the objective function

$$\Phi = (\mathbf{\Xi}_j - \mathbf{\Psi}_j \mathbf{a}_j)^* \mathbf{C}_d^{-1} (\Xi_j - \mathbf{\Psi}_j \mathbf{a}_j) + \alpha \mathbf{a}_j^* \mathbf{C}_m^{-1} \mathbf{a}_j \tag{15}$$

to obtain

$$\hat{\mathbf{a}}_j^{(i+1)} = (\mathbf{\Psi}_j^* \mathbf{C}_d^{-1} \mathbf{\Psi}_j + \alpha \mathbf{C}_m^{-1})^{-1} \mathbf{\Psi}_j^* \mathbf{C}_d^{-1} \mathbf{\Xi}_j. \tag{16}$$

This estimate, with $\alpha = 1$, is the minimum variance linear unbiased estimate under the covariance assumptions given earlier, and is also the Bayesian estimate of $\mathbf{a}_j$ when prior distributions for both the model parameter and data errors are Gaussian, with the covariances given (e.g. Menke 1989).

The estimate (16) can also be computed using the SVD of the appropriate submatrix for segment j, that is, if

$$\Psi_j = \check{\mathbf{U}} \mathbf{\Lambda} \check{\mathbf{V}}^*, \tag{17}$$

then the estimate of (16) can be written as

$$\hat{\mathbf{a}}_j = \mathbf{C}_m^{1/2} \check{\mathbf{V}} \mathbf{\Lambda} \left(\mathbf{\Lambda}^2 + \mathbf{I} \right)^{-1} \check{\mathbf{U}}^* \mathbf{C}_d^{-1/2} \mathbf{\Xi}_j. \tag{18}$$

Note that we have again taken $\alpha = 1$ since the covariance of model parameters and data errors have already been independently estimated, and that $\mathbf{\Lambda} \left(\mathbf{\Lambda}^2 + \mathbf{I} \right)^{-1} = \mathbf{diag} \left[\frac{\lambda_i}{\lambda_i^2 + 1} \right]$ is easily computed from the singular values of Ψ_j.

To reduce the effect of outliers one could in principal also use a robust regression estimator. Instead, we replace the raw data $\mathbf{X}_j$ by the cleaned data $\tilde{\mathbf{X}}_j$ (see 11) in the definition of $\mathbf{\Xi}_j$. Because the polarization parameters are estimated independently for each segment, additional computations are required (as for the spatial mode step) to maintain the orthogonality/normalization conditions (7) for the new estimates at step i. Here this is accomplished by computing the SVD of the matrix $\hat{\mathbf{A}}$ assembled from all of the estimated polarization vectors (i.e. $\hat{\mathbf{a}}_j$ from 18 for $j = 1, \ldots, J$)

$$\hat{\mathbf{A}} = [\hat{\mathbf{a}}_1 \cdots \hat{\mathbf{a}}_J] = \check{\mathbf{U}} \mathbf{\Lambda} \check{\mathbf{V}}^*. \tag{19}$$

The estimates for step $i + 1$ are then

$$\mathbf{A}^{(i+1)} = \mathbf{\Lambda} \check{\mathbf{V}}^*, \qquad \mathbf{U}^{(i+1)} = \mathbf{U}^{(i)} \check{\mathbf{U}}^*. \tag{20}$$

3.4 Estimation of incoherent noise variances

Estimates of the channel incoherent noise variances (4), which are used to define relative channel weights for the robust covariance estimate of Section 3.1, and for the estimates of $\mathbf{A}$ in Section 3.3, can be derived from residuals in the fit of each channel to a set of predicting variables, for example, from the residual variances from the spatial mode estimation step of Section 3.2. However, this simple approach may fail when noise heavily contaminates a single site. In this case, one or more of the modes (which by definition explain maximal variance in the overall array) can become dominated by the noisy site. This noise would then be very easy to fit (essentially predicted by itself), leading to the erroneous conclusion that this site has high SNR. This causes the noisy site to be more heavily weighted in subsequent estimation steps, further increasing contamination of the modes.

We thus follow the approach advocated by Egbert (1997), with each channel in turn excluded from predicting itself. More specifically, residual variances for the N_s channels at a single site (denoted by index s) are constructed as follows. First, we use the matrix of cleaned data from all other sites, with any missing data filled in by predictions, as computed from (9). Call this matrix of $N - N_s$-dimensional data vectors $\tilde{\mathbf{X}}_{(s)}^{(i)}$. These data vectors are then 'projected' onto the current estimate of the spatial modes via

$$\tilde{\mathbf{A}}_{(s)} = \mathbf{U}_{(s)}^{(i)*} \tilde{\mathbf{X}}_{(s)}^{(i)}, \tag{21}$$

where $\mathbf{U}_{(s)}^{(i)}$ is the current estimate of the spatial modes with the channels from site s excluded; this submatrix can be reorthonormalized but this is not necessary. As our notation suggests, this process results in something very like an estimate of the polarization parameters, but with data from site s excluded. We then fit the N_s-dimensional site s data vectors, not filled in or cleaned, and denoted as $\mathbf{X}_s$, using the RME. This produces a robustly estimated TF $\tilde{\mathbf{T}}_{(s)}$, which relates channels at site s to columns of $\tilde{\mathbf{A}}_{(s)}$ as

$$\mathbf{X}_s = \tilde{\mathbf{T}}_{(s)}^* \tilde{\mathbf{A}}_{(s)}, \tag{22}$$

together with robust estimates of residual variances for the N_s predicted components. Note that $\mathbf{T}_s = \tilde{\mathbf{T}}_{(s)} [\mathbf{U}_{(s)}^{(i)}]^*$ provides the corresponding TF, which predicts data at site s using all other sites in the array.

The whole process outlined earlier is repeated for each site, with the results assembled into a vector of N residual variances (one for each channel) and an $N \times N$ TF matrix $\mathbf{T}$ (assembled from the individual site TFs $\mathbf{T}_s$, and with zeros on diagonal blocks. The variances and $\mathbf{T}$ can then be used in the scheme given in Egbert (1997) to allow for effects of noise in the predicting channels, and thus to obtain approximately unbiased estimates of the incoherent noise parameters σ_n^2, $n = 1, N$. Calls to the RME in the incoherent noise estimation step can also be used to generate a cleaned data matrix $\tilde{\mathbf{X}}$. Indeed, we prefer this computation of $\tilde{\mathbf{X}}$ to that resulting from the RME in the spatial mode estimation step, since only variables from other sites are used to compute the predicted data, thereby minimizing the potential influence of large-amplitude noise at isolated sites.

3.5 Choosing a 'core array'

When the amount of missing data are not too great the steps outlined already suffice for robust PCA: obtain initial estimates of $\mathbf{U}$ and $\mathbf{A}$ by setting missing values to zero and using the robust covariance estimation scheme of Section 3.1, and then iterate the steps described in Sections 3.2–3.4 to refine these estimates (inner loop in Fig. 1). In the simplified synthetic data tests discussed later this scheme is successful with up to a few tens of per cent of randomly distributed missing data. If the fraction of missing data is greater,

and for more realistic patterns of missing data (e.g. some sites with long gaps), we find that the iterative estimation process cannot be reliably initiated by just setting missing observations to zero. In general we thus adopt a more complex strategy, first reducing the full array to a 'core' subarray (fewer sites, and segments) with a relatively small amount of missing data (e.g. 5–10 per cent overall; no sites with a very high fraction of missing data). After obtaining estimates of **U** and **A** for this subarray, we then extend estimates to the full array, following the approach described in the next section.

Our scheme for choosing the core array is relatively simple, compromising between the number of sites retained and the number of segments analysed. We first eliminate all segments for which the fraction of available channels falls below some threshold p_{Ch1}, set to a relatively low value to eliminate those segments with few functioning channels (e.g. at the very beginning and end of the array deployment). Then, among this reduced set of segments, all channels for which the available fraction of segments is not at least p_{seg} are eliminated. Finally, using this reduced set of channels, all segments are eliminated for which a fraction of the remaining channels available falls below $p_{\mathrm{Ch2}} > p_{\mathrm{Ch1}}$. Appropriate values for the parameters will depend on array-specific details, including the number of sites, duration of occupation and the pattern of missing data, as discussed further in the context of specific examples. An example of the core selection process is given in Figs 3(a) and (c), for an array with 13 five-channel MT sites (so $N = 65$ channels) and with the three parameters set to $p_{\mathrm{Ch1}} = 0.5$, $p_{\mathrm{seg}} = 0.75$ and $p_{\mathrm{Ch2}} = 1$, respectively.

3.6 Outer loop

Once the core array is chosen, we use the methods of Section 3.1 to obtain estimates of the spatial modes **U**, and polarization parameters **A**. We can then estimate the incoherent noise variances $\mathbf{C_N} = \mathbf{diag}(\sigma_1^2, \ldots, \sigma_N^2)$, using the procedure described in Section 3.4. As discussed there, as a by-product of estimating $\mathbf{C_N}$, outliers in individual data channels are pulled towards predicted values, resulting in a cleaned data matrix $\tilde{\mathbf{X}}$. We thus repeat this step two to three times to further refine the data cleaning, and enhance robustness of incoherent noise variance estimates. We then alternately estimate **A** (following Section 3.3) and **U** (following Section 3.2), requiring typically only a few iterations of this inner loop to obtain stable estimates.

The next step (assuming that the core is a subset of the full array) is to extend the analysis to a larger number of segments, but still restrict to the core subset of data channels for which components of **U** have already been estimated (Fig. 3d). In particular, we use all segments j for which $N_j \geq f_{\mathrm{seg}} K$, where N_j is the number of channels, among those in the core array, which are available for segment j. The parameter $f_{\mathrm{seg}} \geq 1$ is an adjustable scale factor, used to ensure that there are sufficient channels available for the

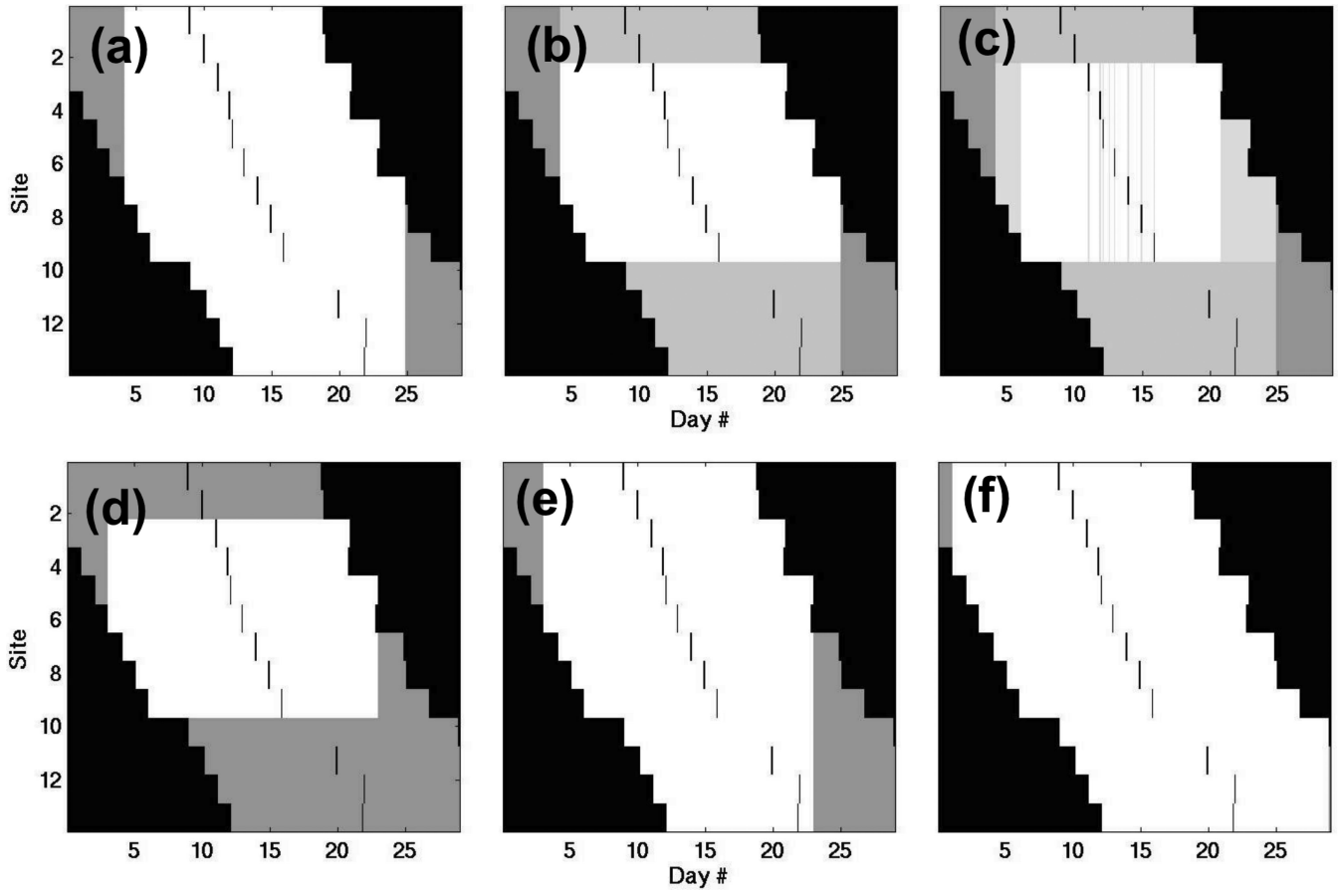

Figure 3. Upper panel: Selection of the core array. Three steps are shown with segments, sites and the last segments omitted in grey (slightly lighter for each successive step). Black is missing data. The remaining white area at the end is the core array (in this case $p_{\mathrm{Ch2}} = 1$, so there are no missing data in the core). Lower panel: shows how the array is built back up, after estimating for the core. Now data not yet used are a single colour of grey. The first illustrates adding segments using the PCs estimated for the core segments, the second extends to additional sites (all in fact), and the third shows the final extension to include all segments with the minimum number of sites (four in this case, corresponding to f_{seg} K = 20 channels).

segment to estimate K polarization parameters; in the applications discussed later (and in particular in Fig. 3) we have set $f_{\rm seg} = 2$. Using this larger set of segments, the incoherent noise estimate is updated, and the alternating sequence of **A** and **U** estimation (i.e. the inner loop) is iterated to convergence. This step provides estimates of the polarization parameters for more segments, and incorporates additional data into the spatial mode estimates.

Next, the number of data channels is increased (Fig. 3e), now using all channels n for which $J_n \geq f_{\rm ch}K$, where J_n is the number of segments (among those used in the previous step) available for channel n, and $f_{\rm ch} \geq 1$ is again a scale factor (which we have again taken to be 2). For this larger array estimates of the polarization parameters are available for all segments, but spatial modes have not yet been estimated for the additional channels. Thus, it is necessary to begin the alternating estimation sequence by estimating **U**. This is followed by estimation of $\mathbf{C}_{\rm N}$ (incoherent noise variances must be estimated for the new channels), and then by a refined estimate of **A**. The sequence is again iterated to convergence. Finally, (Fig. 3f) the time interval used is extended to include all segments with a sufficient number of channels (i.e. $f_{\rm seg}K$), and estimates of **A**, **U** and $\mathbf{C}_{\rm N}$ are again iteratively refined. Depending on the pattern of missing data and the parameters used, additional steps to increase first the number of sites and then the analysis interval may be required. The process is terminated when no segments remain with at least $f_{\rm seg}K$ channels.

4 EXAMPLES

In the following we describe a range of tests we have performed with MsDEMPCA. We keep our focus here on the initial multivariate analysis phase, testing the ability of the algorithm to cope with missing data and noise when estimating spatial modes.

4.1 Synthetic data tests

As a first test of MsDEMPCA we consider a relatively simple synthetic data set, generated to be qualitatively similar to what might be expected from a modest-sized array of MT sites. Here we aim to examine the ability of the algorithm to cope with missing data and noise. For these initial tests we do not reduce to a core array, but rather apply only the inner loop steps to the full array, to explore limits of this simple approach.

To generate the synthetic data two vectors (**U**), representing the idealized plane-wave response for 10 sites (a total of $K = 50$ channels) were generated, with electric channels $\mathbf{E}_x$ and $\mathbf{E}_y$ assigned random amplitudes 5–10 times larger than the horizontal magnetic components $\mathbf{H}_x$ and $\mathbf{H}_y$, which were scaled to unity. Vertical $\mathbf{H}_z$ components were in turn given amplitudes a factor of 2–10 times smaller than the horizontal magnetic components. Then, pairs of random polarization parameters (**a**) for 1000 segments were drawn from a $N(0, 1)$ distribution and used to generate the synthetic data matrix $\mathbf{X} = \mathbf{U}\mathbf{a}$, of size 50×1000.

These synthetic data were contaminated by multiple components of noise, and randomly selected segments were deleted. First, normally distributed random noise with constant amplitude, on average 10 per cent of the horizontal magnetic signal, were added to all channels to simulate incoherent instrumental/site noise. Note that this homogeneous component of noise is relatively smaller for the electric channels, which have significantly larger absolute amplitudes. To simulate outliers we then added large amplitude random errors drawn from an exponential distribution, contaminating a variable fraction (1–10 per cent) of segments, which was chosen randomly for each site, and for each synthetic data realization. To further model intersite variations in data quality, we subdivided nine of the sites between two groups, with different fractions/patterns of missing data. The first group was allowed to have at maximum 90 per cent missing segments, while the second group had at most 30 per cent of the segments deleted. One site out of 10 was always kept complete, with no missing data. We ran Monte Carlo simulations to generate 1000 realizations of synthetic array data following this general template, but with randomly varying fractions of missing and severely contaminated data. It should be noted that we did not introduce any noise which was coherent between sites, so the coherence dimension of the data always remained equal to two.

To compare true and estimated PCs we need a measure of the distance between the subspaces spanned by the columns of the orthonormal matrices **U** (two PCs used to generate the synthetic data) and $\tilde{\mathbf{U}}$ (the two PCs estimated from noisy and incomplete data). This measure must allow for the indeterminacy in the phases of individual PCs, as well as possible mixing between modes (which is not relevant to the distance between the subspaces). Thus we allow the two matrices to be related through multiplication by an arbitrary complex matrix

$$\mathbf{U} = \tilde{\mathbf{U}}\mathbf{C} + \mathbf{e}, \tag{23}$$

where **e** represents the error matrix, associated with differences in the column spaces of **U** and $\hat{\mathbf{U}}$. Taking into account that the columns of both matrices are unit vectors, the LS estimate for **C** would be $\mathbf{C} = \tilde{\mathbf{U}}^*\mathbf{U}$ so that the error matrix can be expressed as $\mathbf{e} = (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^*)\mathbf{U}$. The size of the relative error can then be summarized as $\epsilon = \sqrt{\mathbf{Tr}(\mathbf{e}^*\mathbf{e})/K}$, which we refer to in the following as the subspace distance. Note that for two orthogonal spaces $\epsilon = 1$, while for two coincident spaces $\epsilon = 0$. For small deviations the subspace distance can be taken as an average relative error in the PC components, ignoring mixing between modes.

We first tested a direct SVD approach on clean data sets generated with no missing data and no outliers. This simple approach (and of course also MsDEMPCA) always reproduces the true PC modes with accuracy (as measured by the subspace distance ϵ) to better than a fraction of a per cent. Results for synthetic realizations with missing data (and outliers) are presented in Fig. 4, where ϵ is plotted as a function of the missing data fraction for three different cases, which are distinguished by colour.

The direct SVD applied to the data matrix with missing values set to zero (red dots) does not reproduce the original PCs reliably or accurately, even for missing data fractions as low as 1–3 per cent. MsDEMPCA (green dots) does a much better job, recovering the true PCs within a few per cent in almost all cases, often even with very high fractions (up to 60–70 per cent) of missing data. We also used a variant of the basic MsDEMPCA algorithm outlined earlier, with ROBPCA (Verboven & Hubert 2005) from the Libra statistical library used instead of the affinely invariant robust covariance estimator of Section 3.1 to compute the initial spatial modes estimate (blue dots). Both variants produce results of similar quality, that is, generally quite good, but with failure in some cases with large fractions ($\approx$30 per cent or more) missing data. Interestingly, the estimates break down for different specific cases, which indicates that different estimators have different protection from outliers and different sensitivity to missing data pattern. Because the inner loop algorithm begins to break down when the fraction of missing data is too large we use the outer loop scheme in such cases, first reducing to a core array. This more complex estimation scheme is also applied to the real data cases

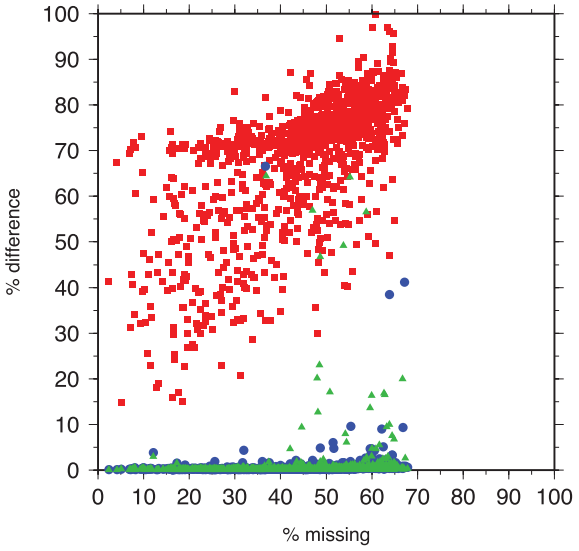

Figure 4. Synthetic test showing the dependence of the subspace distance ϵ (multiplied by 100, and expressed as per cent) between spaces defined by the two PC modes from the amount of missing data in per cent. Red squares show the error of reconstructed PCs using direct SVD of data matrix with missing elements filled with zeroes, green triangles and blue circles show the error of PCs reconstruction using MsDEMPCA basic algorithm and with ROBPCA modification correspondingly. With amount of up to 30 per cent of missing data the algorithm is able to recover PCs within a few per cent, independent on missing data patten.

discussed in the following subsections, where signal and noise are both more complex, and patterns of missing data often more challenging.

4.2 EMScope 'rolling arrays'

As a second test case we consider a set of long-period EMScope MT sites occupied in 2007 in Eastern Washington State, USA, as part of the EarthScope USArray project (Patro & Egbert 2008). In the EMScope campaigns up to 20 MT systems are deployed simultaneously, but in a 'rolling array'. After an instrument has operated for roughly 3 weeks at a site that instrument is moved to a new location, resulting in an array configuration that changes almost daily. We consider a set of 13 consecutively installed sites, located on the quasi-regular USArray grid, with roughly 70 km between sites (Fig. 5). Because of the timing of site occupations, illustrated in Fig. 3, there is a trade-off between the number of sites that may be included in a fully synoptic array and the length of the time window that may be analysed. If the full set of 13 sites are included, only $\approx$5 d could be used for analysis without the extensions developed here to treat missing data.

Before considering the larger set of sites with staggered deployment times, we test the effects of artificially created gaps in a smaller subarray of six sites (filled symbols in Fig. 5). These all operated together for roughly a 2-week period, allowing analysis with more classical (but still robust) methods such as the multivariate scheme originally described in Egbert (1997). Starting from this small array, we deleted data segments randomly, as illustrated in Fig. 6, applied MsDEMPCA, and then compared results to those obtained with the complete data set. In this first run we again test the core part of the algorithm (inner loop).

For each site a random number (between 0 and 6) of gaps were created, with random timing, and with random length distributed so that the expected fraction of missing data (averaged over all sites) was p_{miss}. Note that with this procedure the amount of data missing is quite variable over sites, with some sites having few or no gaps (see Fig. 6). Note also that the same gaps were created for all channels at a site, and for all periods processed. The impact of the gaps becomes slightly greater at the longest periods, because with fewer available segments, the fraction of windowed time-series segments affected by gaps becomes larger. To assess the impact of these gaps on estimates of the modes we use the subspace distance ϵ of Section 4.1, now measuring the distance between the spaces defined by the two dominant PCs estimated with and without the artificial gaps. Relative errors (expressed as 'epsilon') are plotted as a function of period in Fig. 7 for three missing data fractions ($p_{\mathrm{miss}} = 5$ per cent, 10 per cent and 20 per cent).

For reference, in the same figure we also plot an estimate of errors in the two leading PCs, computed with the bootstrap (e.g. Efron &

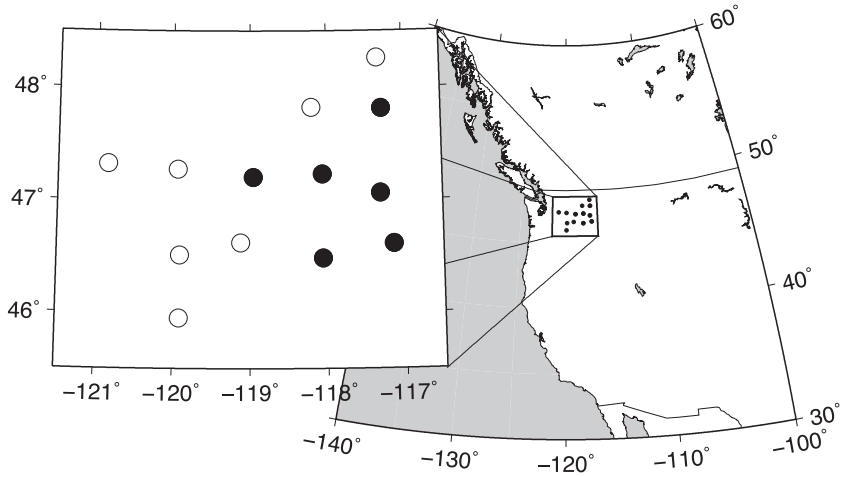

Figure 5. Location of EarthScope sites used for test arrays. The full set of 13 sites have overlapping occupation times (see Fig. 3), but there were only about 5 d when all were operational. MsDEMPCA allows us to analyse data from the full time window shown in Fig. 3. The six solid dots denote sites which were occupied successively in the middle of the time interval, with sufficient overlap in all occupation intervals (roughly 2 weeks) for analysis with a standard (no missing data) approach. We use this subarray to test the effectiveness of MsDEMPCA in recovering these standard estimates when data gaps are introduced artificially.

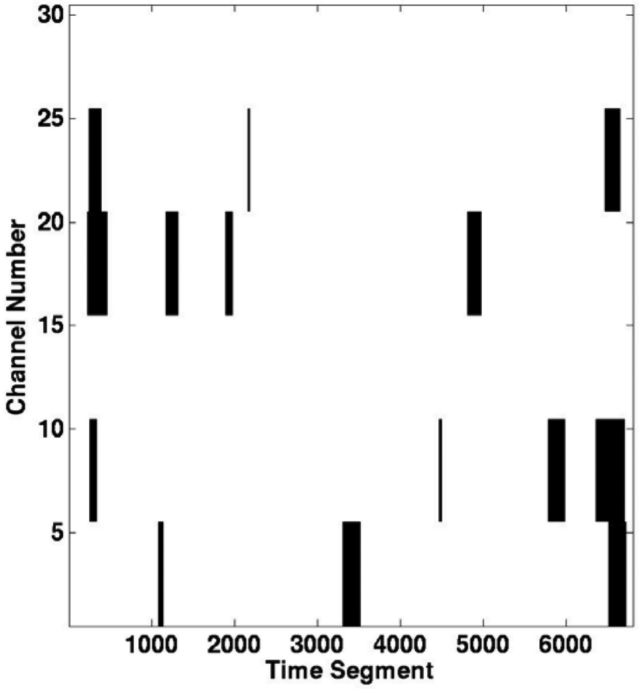

Figure 6. Example of random gaps introduced into the six-site array for testing the missing data scheme. Here the expected fraction missing is 5 per cent.

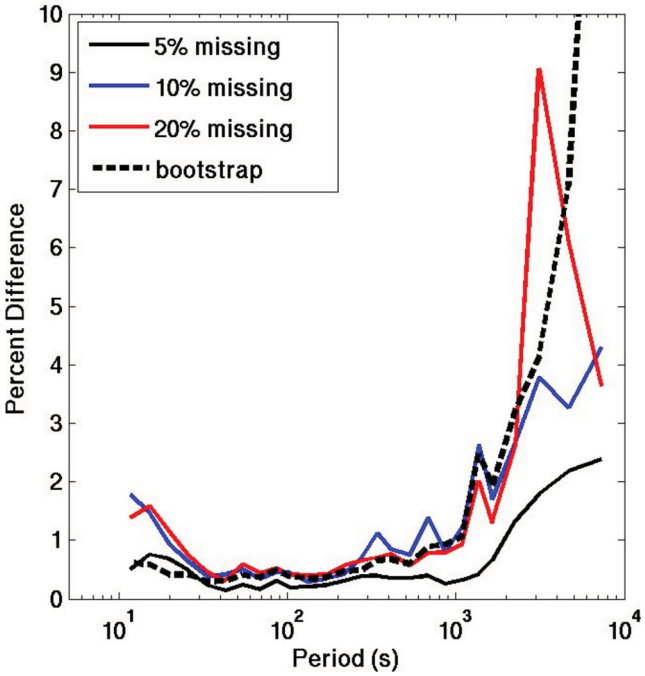

Figure 7. Subspace distance ϵ between spaces defined by the two dominant PC modes, computed with and without artificial gaps in the time-series for the six-site EarthScope array defined by solid dots in Fig. 5. The three solid curves give ϵ as a function of period for three different fractions of missing data, 0.05, 0.10 and 0.20. The heavy dashed line gives a bootstrap estimate of statistical errors in the full data set PC estimates, also expressed as subspace distance.

Tibshirani 1993). The bootstrap errors, which are also summarized with the subspace distance ϵ, were computed by sampling individual array vectors (i.e. all components at all sites together) with replacement from the full set of J such vectors, and then applying the robust PC estimation scheme to the resulting (full, no missing data) arrays. A total of 20 bootstrap replicas were computed for each period band. Then, we computed the subspace distance ϵ (for the two leading PCs) between the population mean (i.e. PCs computed with all data) and each bootstrap replicate. The dashed line in Fig. 7 gives the rms of ϵ over the 20 replicates, a summary estimate of the statistical precision of the PC subspaces determined from the full data set.

In the middle of the period range (≈30–1000 s) the effect of 5–20 per cent missing data is minimal, with relative errors a small fraction of a per cent, and the change in estimates comparable to or smaller than statistical errors inferred from the bootstrap. The effects of missing data are somewhat greater at the shortest and longest periods, especially as the fraction of missing data increases. In particular, for 10–20 per cent missing data changes in the estimates approach 2 per cent at a period of 10 s, well above statistical error levels. These short periods are essentially in the 'dead band' where the SNR for the fluxgate magnetometers used for the survey is very low, typically below 1. Although SNRs for the electric channels are generally higher, overall signal levels are weak, apparently increasing the negative impact of missing data. At long periods (beyond 1000 s) changes resulting from missing data become much larger, approaching 10 per cent at the longest periods (10 000 s) for the case of 20 per cent missing data. However, the bootstrap error estimates are comparable in magnitude suggesting that sample sizes (with 2 weeks of data) are too small for reliable estimation of the PCs, with or without missing data. Indeed, at the longest periods, the estimates of statistical precision are themselves quite variable, and probably not very meaningful. Overall these results suggest that as long as the SNR is not too low, and the fraction of missing data not too great (which we estimate to be about 15–20 per cent), our estimator performs reliably, reproducing estimates that would be obtained with a complete data set, at least to within the intrinsic statistical uncertainty.

We next consider application of the complete estimator, including reduction to a core array and execution of the outer loop, to the full array of 13 sites from Fig. 5. To set the total time interval shown in figure Fig. 3, we required the minimum number of sites to be three, ensuring that the minimum number of channels available exceeded the number of spatial modes estimated ($K = 10$). Earlier and later times, when only two or fewer of the 13 sites were operating are excluded from our analysis. Construction of the 'core' for the initial estimates is illustrated for this array in (already discussed) Figs 3(a–c). Parameters used for selection of the core were in this case (see Section 3.5) $p_{\rm seg} = 0.75$, $p_{\rm Ch1} = 0.5$ and $p_{\rm Ch2} = 1.0$, resulting in a core array with seven sites (35 channels total) running for approximately 15 d. Stages in the expansion from the core to the full array (see Section 3) are illustrated in the bottom three panels of Figs 3(d–f).

The non-dimensional eigenvalues of the normalized SDM (i.e. expressed in SNR units) are plotted as a function of period in Fig. 8. The curves are smooth functions of period, suggesting statistically stable and reliable estimates. There are two clearly dominant modes, with SNR about 20 dB above the next most significant mode; these indeed correspond (at least approximately) to spatially uniform horizontal magnetic fields (Figs 9a and c). The corresponding electric components exhibit substantially more variability, reflecting site-to-site variations in conductivity, including very localized near-surface distortion. In Fig. 9 we have reduced this variability by plotting electric field components scaled by the inverse of the (frequency-dependent) rms channel amplitude. With this scaling electric field

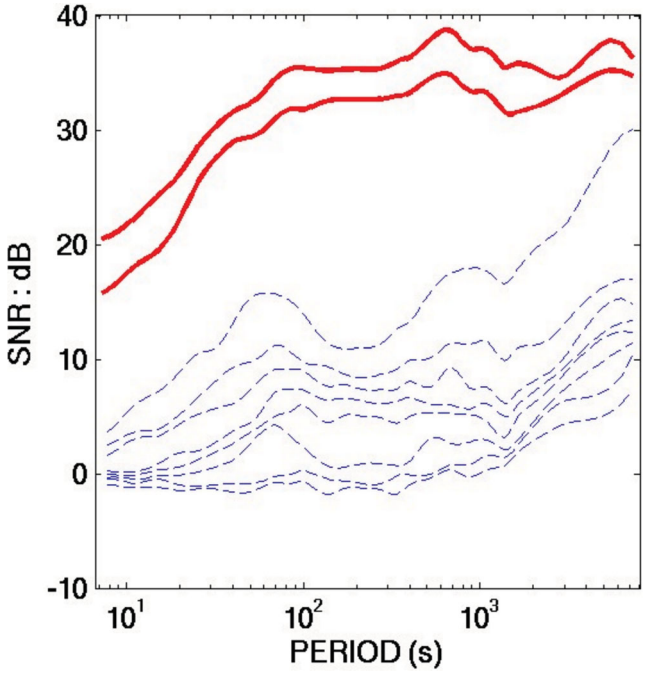

Figure 8. Spectrum of eigenvalues of the normalized SDM estimated for the 13 site array of Fig. 5, obtained using MsDEMPCA. Eigenvalues give SNR of each of the modes (PCs). As expected for MT data from a small array at geomagnetic mid-latitudes, there are two dominant modes, corresponding roughly to two polarizations of uniform magnetic sources.

components for the first two modes (Figs 9b and d) have similar amplitudes, directions and phases, revealing current flow in the Earth that is at least qualitatively consistent with the quasi-uniform magnetic fields. However, site-to-site variations are still quite notable.

The general character of the ordered eigenvalue spectrum is quite similar to that seen previously using the methods of Egbert (1997) on smaller EarthScope arrays with no missing data (Egbert 2008). For example, the amplitude of the third mode increases at long periods; this mode is associated with 'normal Z', that is, H_z that is coherent across the array, associated with horizontal magnetic field gradients (i.e. the signal for an HSG approach). There is also a peak in the mode 3 curve near a period of 50 s, at this period range corresponding approximately to N–S gradients of the N–S magnetic field component (Fig. 9e). At the geomagnetic latitude of the array (53.7°) the field line resonance period for PC3 pulsations is roughly 50 s (Samson *et al.* 1982). A steep N–S gradient of the N–S magnetic field is exactly what would be expected near this resonance period (e.g. Bransky *et al.* 1985). The pattern of scaled electric fields (Fig. 9f) is also consistent with the induced or image currents associated with the ionospheric current systems resulting from field line resonance.

Other secondary modes at this and other periods (not shown) have strong large-scale gradient components in both magnetic and electric components, consistent with expectations from simple models (Egbert 1989) and previous experience with PC analysis of geomagnetic arrays (Egbert 2002, 2008). There is some indication in Fig. 9 for a component of spatially uniform horizontal magnetic fields mixed with the gradients of modes 3–5 (e.g. in Fig. 9g the average magnetic field points to the northeast, implying a uniform component in this direction). As discussed in Egbert (2002) this implies that the gradient components also contaminate the (nominally uniform source) modes 1–2, even though this may not be readily apparent from the figure. Further analysis and processing would be required to allow quantitative interpretation, as discussed in Egbert (2002). Electric components again are roughly consistent with the corresponding magnetics, but site-to-site variation and distortion is even more apparent than for the dominant (first two) quasi-uniform modes.

4.3 The BEAR array

As a first application of MsDEMPCA to a truly large (and rather heterogeneous) EM array we consider the BEAR data (Korja *et al.* 2002) plotted in Fig. 10.

A representative picture of the pattern of available data is shown with the plot of $\lg|\mathbf{X}|$ for 44 sites, for a single period, in Fig. 11. As this figure shows, the BEAR array has a substantial amount of missing data, significantly exceeding 20 per cent, with quality quite variable from site to site. Furthermore, the array is at high geomagnetic latitude (up to 65°), extending into the auroral zone, so we can expect quite strong source field effects. This array thus represents a challenging test case for MsDEMPCA. Initial runs using data from all 44 sites resulted in noisy and inconsistent results, with modes sometimes dominated by a few apparently very noisy sites.

Closer examination of plots, such as Fig. 11, revealed that most of the troublesome sites had either very short recording times (e.g. just a few days of acceptable data), or channels that were dead most of the time or showed little or no coherence with the rest of the array. Consequently we eliminated eight stations (indicated with red boxes in Fig. 11) with limited or poor-quality data, reducing the array to a total of 36 sites. Furthermore we excluded from analysis segments which had less than 10 simultaneous sites running, as well as sections of data from individual sites where local multiple coherences between electric and magnetic channels fell below 0.1. Pre-processing to eliminate obviously contaminated data, as well as parts of recordings where only a few sites were running simultaneously, is apparently required before applying MsDEMPCA to a real data set.

For the reduced 36 site array the parameters of Section 3.5 were set at $p_{\mathrm{Ch1}} = 0.5$, $p_{\mathrm{Ch2}} = 0.9$ and $p_{\mathrm{seg}} = 0.8$, resulting in a core array with 19 sites. Note that because there are many more sites than in the case of the EMScope example, smaller values of p_{Ch1} and p_{Ch2} can be used, and the core array still has enough sites for reliable initial estimates for $K = 10$ spatial modes. In Fig. 12 we show the final eigenvalue spectrum resulting from application of MsDEMPCA. In contrast to the smaller mid-latitude EarthScope array discussed in Section 4.2, for which there were two dominant modes corresponding to nearly spatially uniform magnetic sources, for BEAR there are a number of modes with comparable signal power. Clearly the source fields in this larger high-latitude array are complex, and require significantly more than two source modes for accurate characterization. Given this source complexity, the challenges encountered in estimation of plane-wave TFs for these data (Varentsov *et al.* 2003) are hardly surprising.

Although the horizontal magnetic field components in the first two modes exhibit very similar polarization at all sites (Fig. 13), there are significant (albeit smooth) variations in direction and amplitude across the array. For example, in the first mode amplitudes of the mostly south-pointing magnetic fields increase by almost a factor of three from south to north. All five of the modes plotted in Fig. 13 show smoothly varying horizontal magnetic fields across

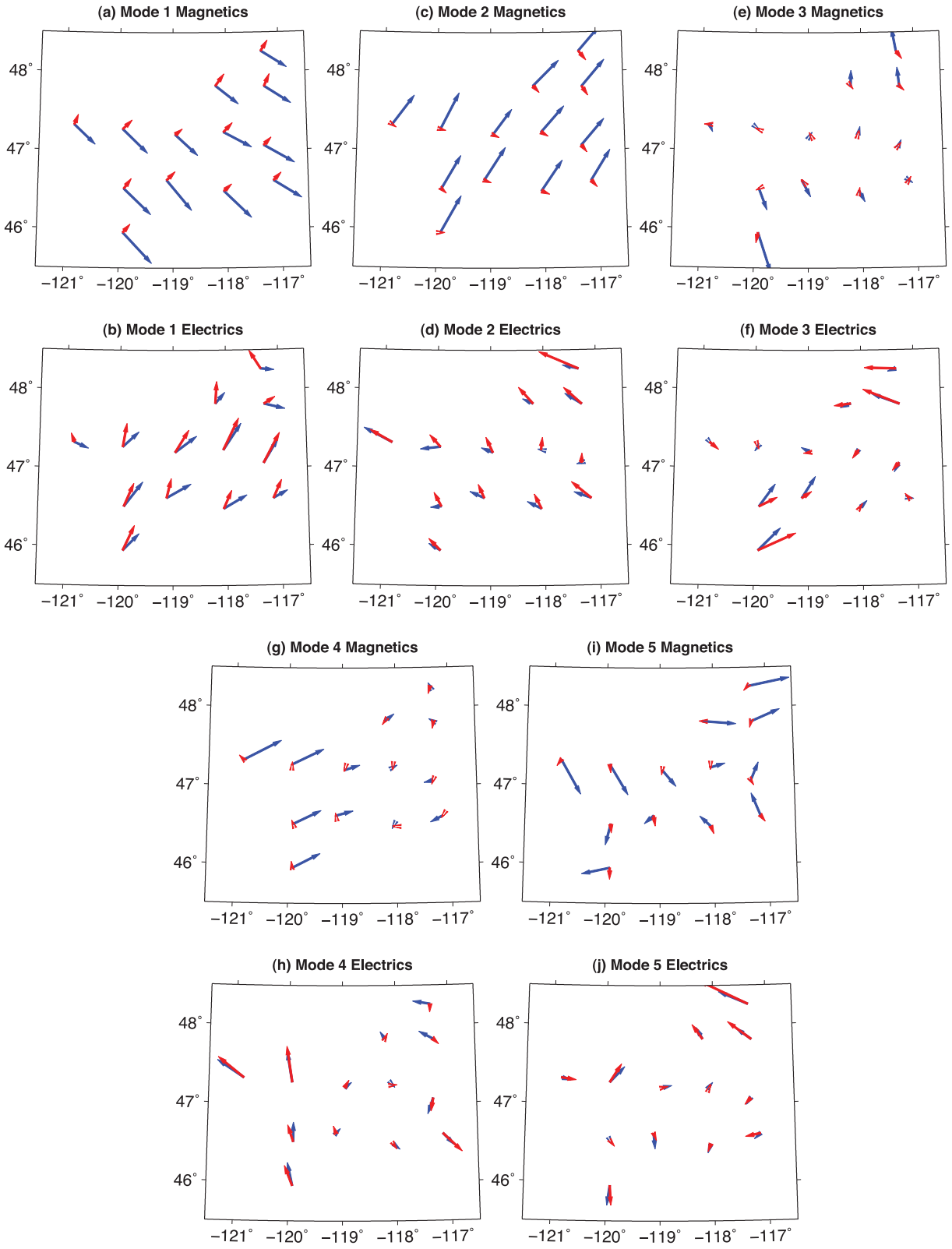

Figure 9. Horizontal magnetic and electric components of the first five spatial modes for the 13 site EarthScope array at a period of 50 s, with blue (red) arrows denoting real (imaginary) components of the horizontal field vectors. Note that the electric components have been scaled by the inverse rms channel amplitude, to approximately remove variability due to near-surface conductive heterogeneity. Horizontal magnetic fields for the first two modes are roughly uniform, with higher modes dominated by gradients; there is also indication for mixing between these two source types. Electric components are physically consistent with the observed magnetic variations, but even after rescaling exhibit greater site-to-site variability. Mode 3, which exhibits N–S gradients of the N–S magnetic component (and of the E–W electric component) is likely associated with PC3 pulsations.

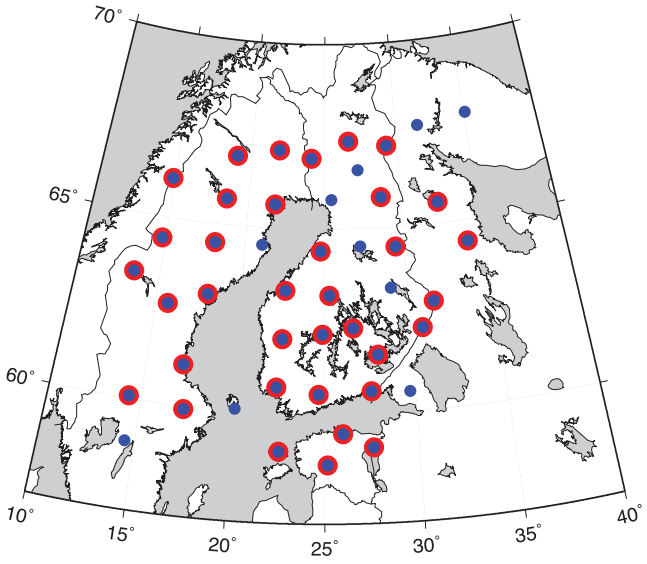

Figure 10. The map of BEAR array with 36 sites indicated by red circles used for multivariate processing. The other 10 sites were omitted due to short recording duration or poor data quality.

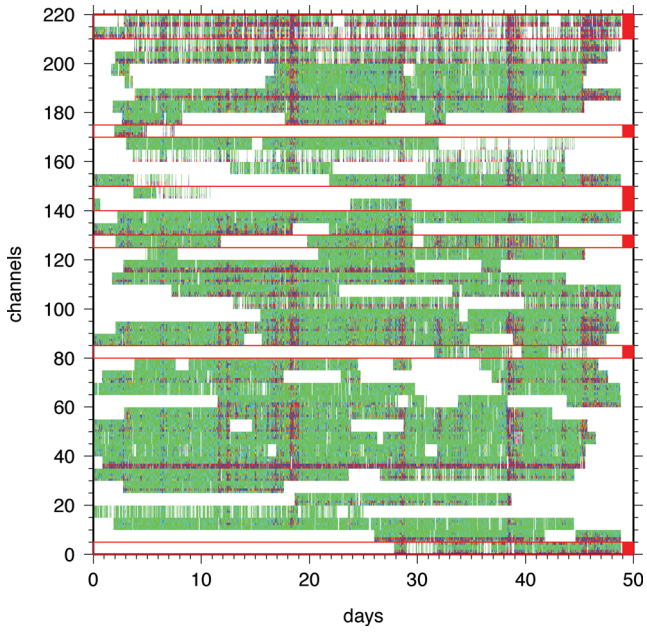

Figure 11. Data pattern at a period of 512 s for the BEAR array, consisting of 44 magnetotelluric sites, recorded simultaneously for about 2 months. The colour indicates $lg|\mathbf{X}|$, the amplitude of the signal. Sites that were excluded from analysis are marked with red boxes.

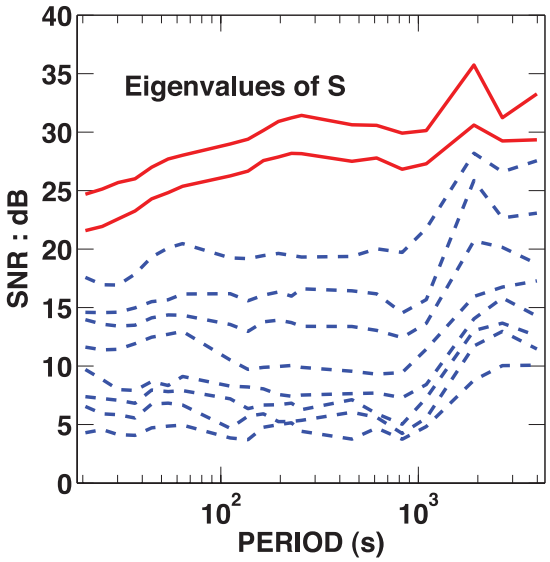

Figure 12. Eigenvalues of the scaled SDM for the BEAR data, giving SNR for the estimated spatial modes. Here all 10 modes are significantly above estimated noise levels, demonstrating that sources are not reasonably approximated by plane wave plus gradient terms.

the array. There are clear signatures of large-scale gradient fields, as we would expect, but there are also more complex (but still smoothly varying), spatial structures evident, especially near the auroral zone. As in Fig. 9 we have scaled electric components based on overall channel rms. Consistent with the southward-pointing magnetic fields in the first mode in Fig. 13(a) the corresponding electric components in panel (b) exhibit dominantly east–west current flow, with larger amplitudes in the north. However, there is still substantial spatial variability in the scaled vectors, consistent with twisting and distortion of electric fields across the array. This is even more pronounced in the second mode, for which the horizontal magnetic components are if anything more uniform. The third mode for the BEAR array is particularly interesting, with the magnetic vectors pointing outwards from the centre of the array, while the scaled electric vectors suggest currents flowing in a circular fashion around this point. This pattern is consistent with the primary gradient component, that associated with normal H_z over a layered (1-D) earth (Egbert 2002).

The PCA modes contain information about both external sources and conductivity variations within the Earth. We briefly consider one approach to separating these components, following the procedure described in (Egbert 2002). Considering that all modes contain a mixture of plane-wave and spatially more complex sources, we seek linear combinations of modes which most closely approximate the form expected for idealized plane-wave sources, that is, spatially uniform and linearly polarized N–S and E–W. Estimates of anomalous horizontal magnetic components for these two idealized sources, displayed as equivalent internal current sheet vectors, are displayed in Fig. 14 for a period of 1097 s. Here the estimated uniform source modes were formed as a sum of the leading eight modes, with coefficients chosen to minimize large-scale gradients; see Egbert (2002) for details. For the NS (EW) mode the horizontal magnetic components have unit magnitude, zero phase and point in the NS (EW) direction. The anomalous horizontal magnetic fields represent deviations from this average—that is, we effectively assume that the average field is 'normal', rather than designating a specific site to define this. Before plotting the anomalous horizontal fields were rotated 90° counter-clockwise to represent equivalent internal currents, and a small component of the normal current (0.25 in Fig. 14) was added to enhance clarity of the image of total electric current flow in the heterogeneous Earth, as discussed in Egbert & Booker (1993) and Egbert (2002).

The anomalous magnetic fields obtained by this procedure are of reasonable magnitude (of the order of 0.1–0.3), and vary smoothly with period, with amplitudes generally reduced at longer periods (not shown). The mapped internal current systems are spatially coherent, and at least qualitatively consistent with conductivity variations determined from previous studies of the area (Fennoscandian conductivity map; Korja *et al.* 2002). There are virtually

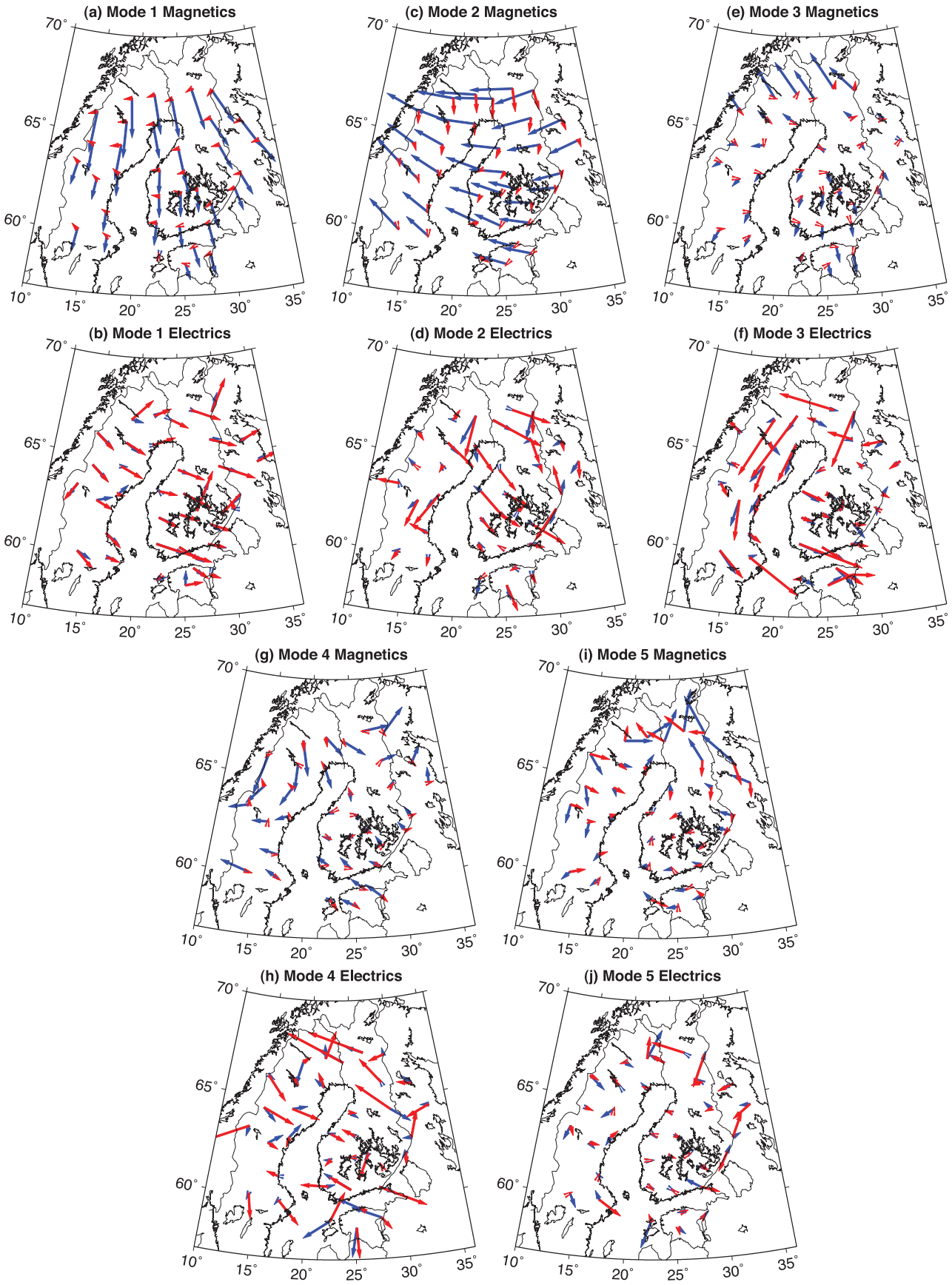

Figure 13. Horizontal magnetic and electric field components for the first five spacial modes at period 1097 s. As in Fig. 9, electric components are scaled by the inverse rms channel amplitude, and blue and red arrows indicate real and imaginary parts correspondingly.

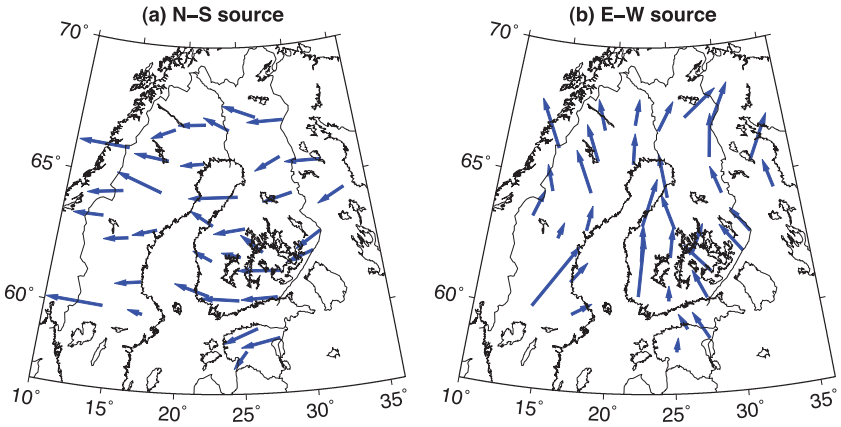

Figure 14. In-phase equivalent internal current sheet vectors, representing horizontal anomalous magnetic fields at period of 1097 s for two different source polarizations.

no large-scale gradients evident for the N–S magnetic field case (Fig. 14a), but for the E–W polarization (Fig. 14b) there is at least some suggestion of residual contamination. The latter can be seen in the tendency for converging current flow in the south, with divergence in the north. The more significant source biases for E–W magnetic fields may result from the dominance for this polarization of field aligned currents with relatively fixed geometry, while the auroral electrojet currents largely responsible for N–S polarized fields exhibit greater variations, and as a result can be more readily superposed to form uniform sources. In contrast to the results of Fig. 14, anomalous magnetic fields obtained from standard interstation TF analysis, for example (Varentsov *et al.* 2003) exhibit very large N–S gradients and are virtually uninterpretable.

5 DISCUSSION AND CONCLUSIONS

In this paper we have extended the multivariate EM array processing methods suggested by Egbert & Booker (1989) and Egbert (1997) to allow for analysis of arrays with large data gaps, and only partial overlaps between site occupations. Our approach is based on a criss-cross regression scheme in which generalized source polarization parameters (characterizing temporal variation of the array signal) and array mode parameters (characterizing spatial structure across the array) are alternately estimated with robust regression procedures. This idea, which was used previously by Egbert (1997) to make the array analysis robust to outliers at individual sites, extends readily to the sort of irregular data patterns frequently encountered with real arrays. The basic scheme can be viewed as a variant on the ER algorithm that has been widely discussed in the statistical literature in the context of robust PCA. However, the MsDEMPCA algorithm has been tailored to physical specifics of the EM array analysis problem, with a number of additional features (e.g. estimation of individual component noise levels, and damped LS estimation of polarization parameters) added to improve robustness and stability.

We have tested MsDEMPCA with synthetic and real data, including data denial experiments where we have created artificial gaps, and compared results obtained with full and incomplete data arrays. These tests reveal that for modest amounts of missing data (up to 20 per cent or so) MsDEMPCA performs well, reproducing essentially the same dominant spatial modes (i.e. those with good SNR) that would be obtained from analysis of the complete array. MsDEMPCA thus makes multivariate analysis practical for the first time for large heterogeneous arrays such as BEAR, where large gaps are common, and a naive analysis based on sections without gaps is all but impossible. MsDEMPCA also allows synoptic analysis of larger numbers of sites collected using practical deployment strategies such as the 'rolling array' used by EMScope.

Our tests have also revealed some limitations to MsDEMPCA. For example, if we increase the number of sites in the EMScope example to around 20, so that there is no temporal overlap between the first and last sites installed, the estimator begins to break down, with results that are notably less smooth with period, and spatially. We also encountered some difficulties with a few sites with poor quality and/or minimal data in our analysis of the BEAR array. One would hope that the robust methods used for the component steps of MsDEMPCA would already protect against anomalous or noisy sites, but we still found it necessary to intervene and remove the worst quality sites by hand. Further study of how these breakdowns occur may allow us to make MsDEMPCA more robust to these sorts of challenging sites, and perhaps to treat arrays with even less temporal overlap. We view MsDEMPCA as a step towards practical routine processing of EM array data, but further refinements to the algorithms are certainly possible, and warranted. We also anticipate that with further experience applying MsDEMPCA to real data sets, performance may be improved. In particular, appropriate choices for the modifiable parameters discussed earlier (e.g. for choosing a core array), are expected to depend on the array size and configuration. In our initial tests we have used a trial-and-error approach to set these parameters. A more systematic study of this issue could result in some generally useful guidelines for optimizing performance.

Another important issue that we have not discussed explicitly here is characterization and estimation of statistical uncertainty of the spatial modes. One fairly straightforward approach would be to apply a bootstrap scheme, such as that used in Section 4.2 to estimate the magnitude of the subspace distance ϵ expected due to errors in PC estimates for a complete array. The same approach could obviously be applied to MsDEMPCA to estimate uncertainties in modes computed from an incomplete data array, or other quantities (such as TF components) derived from these. Error estimates based on asymptotics for the MEV model, as discussed in Egbert (1997) could also be adapted to the missing data case, although doing this in a rigorous manner would probably not be so simple.

Finally, we note that we have focused here on development of the robust multivariate analysis of EM data, a tool to extract the

coherent signal (and noise) from an array. Development of methods for separating PCA modes into source components that can be usefully interpreted, and application of these methods to large arrays such as BEAR for source and geological studies, are beyond the scope of this paper. Our examples here do however illustrate the value of the methods we have developed. MsDEMPCA makes robust PCA practical for a very heterogeneous large array such as BEAR, ultimately enabling new approaches to analysis which will enhance our understanding of external sources, induction, and Earth structure.

ACKNOWLEDGMENTS

This work was partially supported by grants from NASA (NNX08AG04G) and NSF (EAR-0739111) to GDE, and by Academy of Finland (136345) to MYuS.

REFERENCES

Banks, R., 1969. Geomagnetic variations and the electrical conductivity of the upper mantle, *Geophys. J. R. astr. Soc.,* **17,** 457–487.

Bransky, L.N., Borokov, J.E., Gohkberg, M.B., Krylov, S.M. & Trotskaya, V., 1985. High resolution method of direct measurement of magnetic field lines eigenfrequencies, *Planet. Space Sci.,* **33,** 1369–1375.

Croux, C., Filzmoser, P. & Oliveira, M.R., 2007. Algorithms for projection pursuit robust principal component analysis, *Chemometr. Intell. Lab. Syst.,* **87,** 218–225.

Daszykowski, M., Serneels, S., Kaczmarek, K., Espen, P.V., Croux, C. & Walczak, B., 2007. TOMCAT: a MATLAB toolbox for multivariate calibration techniques, *Chemometr. Intell. Lab. Syst.,* **85,** 269–277.

De La Torre, F. & Black, M.J., 2003. A framework for robust subspace learning, *Int. J. Comput. Vision,* **54,** 117–142.

Dempster, A.P., Laird, N.M. & Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm, *J. R. Statist. Soc. Series B (Methodological),* **39,** 1–38.

Efron, B. & Tibshirani, R., 1993. *An Introduction to the Bootstrap*, Chapman & Hall, Boca Raton, FL.

Egbert, G.D., 1989. Multivariate analysis of geomagnetic array data 2. Random source models, *J. geophys. Res.,* **94**(B10), 14 249–14 265.

Egbert, G.D., 1997. Robust multiple-station magnetotelluric data processing, *Geophys. J. Int.,* **130,** 475–496, doi:10.1111/j.1365-246X.1997.tb05663.x.

Egbert, G.D., 2002. Processing and interpretation of electromagnetic induction array data, *Surv. Geophys.,* **23,** 207–249, doi:10.1023/A:1015012821040.

Egbert, G.D., 2008. Source complications in earthscope magnetotelluric data: challenges and opportunities, in *The 19th International Workshop on EM Induction in The Earth*, Beijing, China, 2008 October 23–29.

Egbert, G.D. & Booker, J.R., 1989. Multivariate analysis of geomagnetic array data 1. The response space, *J. geophys. Res.,* **94**(B10), 14 227–14 247.

Egbert, G.D. & Booker, J.R., 1993. Imaging crustal structure in southwestern Washington with small magnetometer arrays, *J. geophys. Res.,* **98**(B9), 15 967–15 985.

Egbert, G.D., Eisel, M., Boyd, O.S. & Morrison, H.F., 2000. DC trains and PC3s: source effects in mid-latitude geomagnetic transfer functions, *Geophys. Res. Lett.,* **27**(1), 25–28.

Egbert, G. *et al.*, 2007. EmScope 150, electromagnetic component of EarthScope backbone and transportable array experiments 2006-2008, *EOS, Trans. Am. geophys. Un.,* **88**(52), Fall Meet. Suppl., Abstract GP32A-01.

Frahm, G. & Jaekel, U., 2010. A generalization of Tyler's M-estimators to the case of incomplete data, *Comput. Stat. Data Anal.,* **54**(2), 374–393.

Fujii, I. & Schultz, A., 2002. The 3D electromagnetic response of the earth to ring current and auroral oval excitation, *Geophys. J. Int.,* **151**(3), 689–709.

Gabriel, K.R. & Zamir, S., 1979. Lower rank approximation of matrices by least-squares with any choice of weights, *Technometrics,* **21**(4), 489–498.

Gleser, L.J., 1981. Estimation in a multivariate "errors in variables" regression model: large sample results, *Ann. Statist.,* **9**(1), 24–44.

Gough, D.I., McKirdy, D.M., Woods, D.V. & Geiger, H., 1989. Conductive structures and tectonics beneath the EMSLAB land array, *J. geophys. Res.,* **94,** B10, doi:10.1029/JB094iB10p14099.

Huber, P., 1981. *Robust Statistics*, Wiley, New York, NY.

Hubert, M., Rousseeuw, P.J. & Vanden Branden, K., 2005. ROBPCA: a new approach to robust principal component analysis, *Technometrics,* **47**(1), 64–79.

Kelbert, A., Schultz, A. & Egbert, G.D., 2009. Global electromagnetic induction constraints on transition-zone water content variations, *Nature,* **460**(7258), 1003–1006.

Korja, T. *et al.*, 2002. Crustal conductivity in Fennoscandia: a compilation of a database on crustal conductance in the Fennoscandian shield, *Earth Planets Space,* **54**(5), 535–558.

Kuckes, A.F., Nekut, A.G. & Thompson, B.G., 1985. A geomagnetic scattering theory for evaluation of earth structure, *Geophys. J. R. astr. Soc.,* **83**(2), 319–330.

Larsen, J.C., Mackie, R.L., Manzella, A., Fiordelisi, A. & Rieven, S., 1996. Robust smooth magnetotelluric transfer functions, *Geophys. J. Int.,* **124**(3), 801–819.

Menke, W., 1989. *Geophysical Data Analysis: Discrete Inverse Theory*, International Geophysics Series Vol. 45, Academic Press, London.

Olsen, N., 1998. The electrical conductivity of the mantle beneath europe derived from C-responses from 3 to 720 hr, *Geophys. J. Int.,* **133**(2), 298–308.

Patro, P.K. & Egbert, G.D., 2008. Regional conductivity structure of Cascadia: preliminary results from 3D inversion of USArray transportable array magnetotelluric data, *Geophys. Res. Lett.,* **35**(20), L20311, doi:10.1029/2008GL035326.

Qian, W. & Pedersen, L.B., 1991. Industrial interference magnetotellurics: an example from the Tangshan area, China, *Geophysics,* **56(2),** 265–273.

Ritter, P. & Banks, R.J., 1998. Separation of local and regional information in distorted GDS response functions by hypothetical event analysis, *Geophys. J. Int.,* **135,** 923–942, doi:10.1046/j.1365-246X.1998.t01-1-00674.x.

Samson, J., Jacobs, J. & Rostoker, G., 1982. Latitude dependent characteristics of long-period geomagnetic micropulsations, *J. geophys. Res.,* **76,** 3675–3683.

Schmucker, U., 2003. Horizontal spatial gradient sounding and geomagnetic depth sounding in the period range of daily variations, in *Protokoll 20 Kolloquium Elektromagnetische Tiefenforschung in Konigstein/Elbsandsteingeb*, pp. 306–317, Hördt, A. & Stoll, J., eds, Deutsche Geophysikalische Gesellschaft.

Serneels, S. & Verdonck, T., 2008. Principal component analysis for data containing outliers and missing elements, *Comput. Stat. Data Anal.,* **52**(3), 1712–1727.

Smirnov, M.Yu., 2003. Magnetotelluric data processing with a robust statistical procedure having a high breakdown point, *Geophys. J. Int.,* **152**(1), 1–7.

Smirnov, M.Yu., Korj, T.J. & Pedersen, L.B., 2008. Electrical conductivity of the Archaean lithosphere in Fennoscandia: results from two magnetotelluric mini arrays (EMMA project), in *The 19th International Workshop on EM Induction in The Earth*, Beijing, China, 2008 October 23–29.

Stanimirova, I., Daszykowski, M. & Walczak, B., 2007. Dealing with missing values and outliers in principal component analysis, *Talanta,* **72**(1), 172–178.

Varentsov, I.M., Sokolova, E.Y. & Grp, B.W., 2003. Diagnostics and suppression of auroral distortions in the transfer operators of the electromagnetic field in the BEAR experiment, *Izvestiya Phys. Sol. Earth,* **39**(4), 283–307.

Verboven, S. & Hubert, M., 2005. LIBRA: a MATLAB library for robust analysis, *Chemometr. Intell. Lab. Syst.,* **75**(2), 127–136.

Verboven, S. & Hubert, M., 2010. MATLAB library LIBRA, *Wiley Interdisciplin. Rev. Computat. Statist.,* **2**(4), 509–515.

Vozar, J. & Semenov, V.Y., 2010. Compatibility of induction methods for mantle soundings, *J. geophys. Res.,* **115**(B3), B03101, doi:10.1029/2009JB006390.