Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values

TAPIO SCHNEIDER

Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey

ABSTRACT

Estimating the mean and the covariance matrix of an incomplete dataset and filling in missing values with imputed values is generally a nonlinear problem, which must be solved iteratively. The expectation maximization (EM) algorithm for Gaussian data, an iterative method both for the estimation of mean values and covariance matrices from incomplete datasets and for the imputation of missing values, is taken as the point of departure for the development of a regularized EM algorithm. In contrast to the conventional EM algorithm, the regularized EM algorithm is applicable to sets of climate data, in which the number of variables typically exceeds the sample size. The regularized EM algorithm is based on iterated analyses of linear regressions of variables with missing values on variables with available values, with regression coefficients estimated by ridge regression, a regularized regression method in which a continuous regularization parameter controls the filtering of the noise in the data. The regularization parameter is determined by generalized cross-validation, such as to minimize, approximately, the expected mean-squared error of the imputed values. The regularized EM algorithm can estimate, and exploit for the imputation of missing values, both synchronic and diachronic covariance matrices, which may contain information on spatial covariability, stationary temporal covariability, or cyclostationary temporal covariability. A test of the regularized EM algorithm with simulated surface temperature data demonstrates that the algorithm is applicable to typical sets of climate data and that it leads to more accurate estimates of the missing values than a conventional noniterative imputation technique.

---

1. Introduction

Because the availability of climatic measurements varies spatially and temporally, sets of climate data are usually incomplete. This circumstance complicates multivariate analyses of climate data. Already the estimation of mean values and covariance matrices, the fundamental statistics from which every multivariate analysis issues, becomes difficult when a dataset is incomplete. For example, mean values and covariance matrices of the earth's surface temperature are needed to assess whether climate models simulate the spatial and temporal temperature variability adequately. If a complete dataset of surface temperatures were available, the sample mean and the sample covariance matrix of the dataset—provided that measurement errors are negligible—would represent consistent and unbiased estimators of the mean and covariance matrix of the surface temperature in the region and period encompassed by the dataset. But if only an incomplete dataset of surface temperatures is available, a direct estimation of the mean

and of the covariance matrix from the available data usually is not admissible. For instance, the sample mean of the available data can be an inaccurate estimate of the mean of the data. And if one would estimate a covariance matrix from all data available in an incomplete dataset, leaving out in the sample covariance matrix the terms involving missing values, the estimated covariance matrix would not necessarily be positive semidefinite and might have negative eigenvalues. But a covariance matrix estimate with negative eigenvalues might lead to erratic results in analyses that, like the principal component analysis, rest upon eigendecompositions of covariance matrices. Moreover, projections of multivariate data onto subspaces—projections onto the empirical orthogonal functions (EOFs) of a principal component analysis, for example—are not well-defined when values of variables are missing.

One might circumvent the difficulties that incomplete data cause in a multivariate analysis by excluding from the analysis all variables for which values are missing. For example, if the dataset under consideration contains yearly records of monthly mean surface temperatures, with each variable of the dataset representing the temperature at one point of a global grid, one could exclude from the analysis all variables, or all grid points, for which at least one temperature value is missing. Having

Corresponding author address: Tapio Schneider, Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012.
E-mail: tapio@cims.nyu.edu

thus restricted the analysis to a complete subset of the data, one could estimate mean values and covariance matrices as sample means and sample covariance matrices, and projections of the reduced dataset onto subspaces would be well-defined. However, excluding variables from the analysis for which only a few values are missing would mean using the available information inefficiently. Methods are therefore needed for estimating mean values and covariance matrices reliably from all information available in an incomplete dataset. Since low-dimensional projections of multivariate data, for example, in the form of spatial averages or principal components, play an important role in the analysis of climate data, methods for the estimation of mean values and covariance matrices should also fill in missing values with plausible imputed values, such that projections of the completed dataset onto subspaces are good approximations of the corresponding projections of the unknown complete dataset.

The estimation of mean values and covariance matrices from incomplete data and the imputation of missing values are closely linked problems. When an estimate of the mean and a positive definite estimate of the covariance matrix of a dataset are available, the missing values in the dataset can be filled in with their conditional expectation values given the available values in the dataset (Buck 1960). Conversely, the mean and the covariance matrix can be estimated from a completed dataset with imputed values filled in for missing values, provided that an estimate of the covariance matrix of the error of the imputed values is also available. The covariance matrix of the error of the imputed values is required because the sample covariance matrix of the completed dataset underestimates the variances and covariances of the data if, as is the case when imputing conditional expectation values, the imputed values come exclusively from the center of the conditional distribution of the missing values given the available values. The expected variances and covariances of the deviations of the missing values from the imputed values, or the expected variances and covariances of the imputation error, must be taken into account in estimating the covariance matrix of the data (Little and Rubin 1987, chapter 3.4).

Since estimates of the mean and of the covariance matrix of an incomplete dataset depend on the unknown missing values, and since, conversely, estimates of the missing values depend on the unknown statistics of the data, estimating the mean and the covariance matrix of an incomplete dataset and imputing missing values generally is a nonlinear problem, which must be solved iteratively. In what follows, an iterative method both for the estimation of mean values and covariance matrices and for the imputation of missing values will be presented. The expectation maximization (EM) algorithm (Dempster et al. 1977) is taken as the point of departure for the development of a regularized EM algorithm that is applicable to incomplete sets of climate data, in which the number of variables typically exceeds the number of records. What will result is an algebraic framework within which some conventional techniques for the imputation of missing values in climate data—for example, the techniques described by Smith et al. (1996) and by Mann et al. (1998)—can be interpreted as approximations to regularized EM algorithms.

With the EM algorithm, the maximum likelihood estimates of the parameters of any probability distribution can be computed from incomplete data [see Little and Rubin (1987, chapter 7) for a review]. For Gaussian data, whose probability distribution can be parameterized by the mean and the covariance matrix, the EM algorithm starts with initial guesses for the mean and the covariance matrix and then cycles through the alternating steps of imputing missing values and re-estimating the mean and the covariance matrix from the completed dataset and from an estimate of the covariance matrix of the imputation error. In the imputation step, the missing values of the dataset are filled in with their conditional expectation values given the available values, and the covariance matrix of the error of the thus imputed values is estimated. The expectations both of the missing values and of the covariance matrix of the imputation error are computed from the estimates of the mean and of the covariance matrix and hence are conditional expectations given these estimates. In the estimation step, the mean and the covariance matrix are re-estimated, whereby the contribution of the conditional imputation error to the covariance matrix is taken into account. The imputation step and the estimation step are iterated until the imputed values and the estimated mean and covariance matrix stop changing from one iteration to the next.

In order to make the EM algorithm applicable to typical sets of climate data, it must be modified. In the EM algorithm for Gaussian data, the conditional expectations of the missing values and of the covariance matrix of the imputation error follow, for each record with missing values, from an analysis of the linear regression of the variables with missing values on the variables with available values, which means that the EM algorithm for Gaussian data is based on iterated linear regression analyses (see, e.g., Little and Rubin 1987, chapter 8). Yet typical sets of climate data, containing thousands of variables but at most a few hundred records from which the statistics of the data can be estimated, are rank-deficient, so that the parameters of the regression models in the EM algorithm, and thus the conditional expectations of the missing values given the available values, are underdetermined. Such underdetermined, or ill-posed, problems can be solved with regularization methods, which impose additional constraints on the solution to render it unique, for example, by requiring smoothness of the completed dataset with imputed values filled in for the missing values [see Hansen (1997) for a survey of regularization methods]. To make the EM algorithm applicable to sets of climate data, the

ill-posed problem of estimating regression models from rank-deficient data must be regularized with some such method.

Ill-posed problems in climate research are often regularized by performing multivariate analyses in a truncated principal component basis (see, e.g., Smith et al. 1996; Kaplan et al. 1997; Mann et al. 1998). If a problem is regularized by truncating a principal component analysis, high-frequency or small-scale components of the solution, represented by higher-order principal components, are filtered out. The truncation parameter, specifying the degree of regularization, or, for spatial data, the degree of smoothness, is a discrete regularization parameter, which is often determined with ad hoc techniques (see, e.g., Kaplan et al. 1997).

In the regularized EM algorithm that will be presented, the regression parameters are not estimated in a truncated principal component basis, but regularized regression parameters are computed with a method known to statisticians as ridge regression and to applied mathematicians as Tikhonov regularization (Hoerl and Kennard 1970a,b; Tikhonov and Arsenin 1977). In ridge regression, a continuous regularization parameter controls the degree of regularization imposed on the regression coefficients. High-frequency or small-scale components in the regression coefficients are filtered out not by truncating a principal component analysis, but by gradually damping the amplitudes of higher-order principal components (Hansen 1997, chapter 4.2). In the regularized EM algorithm, the regularization parameter that controls the filtering is determined by generalized cross-validation (Golub et al. 1979), in such a way as to minimize, approximately, the expected mean-squared error of the imputed values. With simulated surface temperature data, it will be demonstrated that the resulting algorithm leads to more accurate estimates of the missing values than a noniterative imputation technique (Smith et al. 1996) that is based on a truncated principal component analysis.

In section 2, a review of the EM algorithm for Gaussian data of full rank introduces the concepts and the notation used throughout this paper. Section 3 takes the EM algorithm for Gaussian data of full rank as the point of departure to develop a regularized variant of the EM algorithm in which ridge regression with generalized cross-validation replaces the potentially ill-posed maximum likelihood estimation of the regression parameters in the conventional EM algorithm. This regularized EM algorithm is applicable to rank-deficient data. The discussion in sections 2 and 3 is abstract; no particular spatial or temporal interpretation is assigned to the variables in a dataset. The variables may represent spatial or temporal or mixed spatiotemporal data. In section 4, it is shown how, by arranging the variables in a dataset into records in different ways, the regularized EM algorithm can estimate, and exploit for the imputation of missing values, spatial covariance matrices and mixed spatiotemporal covariance matrices. In section 5, the

regularized EM algorithm is compared with conventional techniques for the imputation of missing values in climate data. Section 6 describes results of a test of the regularized EM algorithm with simulated surface temperature data in which values were deleted in a manner characteristic for observational data. Section 7 summarizes the conclusions and discusses potential applications of the regularized EM algorithm, for example, to the construction of historic surface temperature datasets. The appendix contains a note on estimating the imputation error.

2. The expectation maximization algorithm

With the EM algorithm, the parameters of a probability distribution are estimated from incomplete data by maximizing iteratively the likelihood of the available data, the likelihood of the available data being viewed as a function of the parameters (Dempster et al. 1977). The EM algorithm, like all methods for incomplete data that ignore the mechanism causing the gaps in the dataset, rests on the assumption that the missing values in the dataset are missing at random, in the sense that the probability that a value is missing does not depend on the missing value (Rubin 1976). For example, in a dataset with monthly mean surface temperatures on a spatial grid, the missing values are missing at random if correlations between anthropogenic temperature changes and the availability of data are negligible, for then the availability of a temperature measurement usually does not depend on the temperature to be measured. As a contrasting example, the availability of in situ measurements of the windspeeds in hurricanes does depend on the windspeed to be measured, so it would not be justified to assume that the missing values are missing at random. The EM algorithm and the methods that will be derived from it in subsequent sections are only applicable to datasets in which the missing values are missing at random.

The probability distribution of multivariate Gaussian data can be parameterized by the mean and the covariance matrix (i.e., the mean and the covariance matrix are sufficient statistics of the Gaussian distribution). In an iteration of the EM algorithm for Gaussian data, estimates of the mean and of the covariance matrix are revised in three steps. First, for each record with missing values, the regression parameters of the variables with missing values on the variables with available values are computed from the estimates of the mean and of the covariance matrix. Second, the missing values in a record are filled in with their conditional expectation values given the available values and the estimates of the mean and of the covariance matrix, the conditional expectation values being the product of the available values and the estimated regression coefficients. Third, the mean and the covariance matrix are re-estimated, the mean as the sample mean of the completed dataset and the covariance matrix as the sum of the sample co-

variance matrix of the completed dataset and the contributions of the conditional covariance matrices of the imputation errors in the records with imputed values (see, e.g., Little and Rubin 1987, chapter 8). The EM algorithm starts with initial estimates of the mean and of the covariance matrix and cycles through these steps until the imputed values and the estimates of the mean and of the covariance matrix stop changing appreciably from one iteration to the next.

For the following formal description of the EM algorithm, let $\mathbf{X} \in \Re^{n \times p}$ be a data matrix with n records consisting of p variables, with the values of some of the variables missing in some records. The p variables might represent a geophysical field at p different locations, and the n records might represent incomplete measurements of the p variables at n different times. In the conventional EM algorithm, the number n of records is assumed to be much greater than the number p of variables, so that the sample covariance matrix of the dataset completed with imputed values is positive definite.

From the incomplete dataset, the mean $\mu \in \Re^{1 \times p}$ of the records and the covariance matrix $\mathbf{\Sigma} \in \Re^{p \times p}$ of the variables are to be estimated. For a given record $\mathbf{x} = \mathbf{X}_{i:}$ with missing values,[1] let the vector $\mathbf{x}_a \in \Re^{1 \times p_a}$ consist of the p_a variables for which, in the given record, the values are available, and let the vector $\mathbf{x}_m \in \Re^{1 \times p_m}$ consist of the remaining p_m variables for which, in the given record, the values are missing. Let the mean μ be partitioned correspondingly into a part $\mu_a \in \Re^{1 \times p_a}$ with the mean values of the variables for which, in the given record, the values are available, and a part $\mu_m \in \Re^{1 \times p_m}$ with the mean values of the variables for which, in the given record, the values are missing. For each record $\mathbf{x} = \mathbf{X}_{i:}$ ($i = 1, \ldots, n$) with missing values, the relationship between the variables with missing values and the variables with available values is modeled by a linear regression model

$$\mathbf{x}_m = \mu_m + (\mathbf{x}_a - \mu_a)\mathbf{B} + \mathbf{e}. \qquad (1)$$

The matrix $\mathbf{B} \in \Re^{p_a \times p_m}$ is a matrix of regression coefficients, and the residual $\mathbf{e} \in \Re^{1 \times p_m}$ is assumed to be a random vector with mean zero and unknown covariance matrix $\mathbf{C} \in \Re^{p_m \times p_m}$. In each iteration of the EM algorithm, estimates of the mean μ and of the covariance matrix $\mathbf{\Sigma}$ are taken as given, and from these estimates, the conditional maximum likelihood estimates of the matrix of regression coefficients $\mathbf{B}$ and of the covariance matrix $\mathbf{C}$ of the residual are computed for each record with missing values. With the estimated regression model for each record, the missing values are then filled in with imputed values, and new estimates of the mean μ and of the covariance matrix $\mathbf{\Sigma}$ are computed from the

completed dataset and from the estimates of the residual covariance matrices $\mathbf{C}$.[2]

Let $\hat{\mu}^{(t)}$ and $\hat{\mathbf{\Sigma}}^{(t)}$ denote the estimates of the mean and of the covariance matrix in the tth iteration of the EM algorithm. (The hat accent $\hat{\mathbf{A}}$ designates an estimate of a quantity $\mathbf{A}$.) The estimates of the mean and of the covariance matrix are either the result of the preceding EM iteration or, in the first EM iteration, they may be the sample mean and the sample covariance matrix of the dataset with initial guesses filled in for the missing values. For a given record $\mathbf{x} = \mathbf{X}_{i:}$ with missing values, let the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ be partitioned corresponding to the partitioning of the given record into variables with available values and variables with missing values: let the submatrix $\hat{\mathbf{\Sigma}}_{aa}$ of the estimated covariance matrix $\hat{\mathbf{\Sigma}}^{(t)}$ consist of the estimated variances and covariances of the variables for which, in the given record, the values are available; let the submatrix $\hat{\mathbf{\Sigma}}_{mm}$ consist of the estimated variances and covariances of the variables for which, in the given record, the values are missing; and let the two submatrices $\hat{\mathbf{\Sigma}}_{am}$ and $\hat{\mathbf{\Sigma}}_{ma}$ with $\hat{\mathbf{\Sigma}}_{am} = \hat{\mathbf{\Sigma}}_{ma}^{\mathrm{T}}$ consist of the estimated cross-covariances of the variables for which, in the given record, the values are available with the variables for which, in the given record, the values are missing.[3] Given the partitioned estimate of the covariance matrix $\hat{\mathbf{\Sigma}}^{(t)}$, the conditional maximum likelihood estimate of the regression coefficients can be written as

$$\hat{\mathbf{B}} = \hat{\mathbf{\Sigma}}_{aa}^{-1}\hat{\mathbf{\Sigma}}_{am} \qquad (2)$$

(cf. Mardia et al. 1979, chapter 6.2). From the structure of the regression model (1) follows that, given an estimate $\hat{\mathbf{B}}$ of the regression coefficients and the partitioned estimate of the covariance matrix $\hat{\mathbf{\Sigma}}^{(t)}$, an estimate of the residual covariance matrix takes the generic form

$$\hat{\mathbf{C}} = \hat{\mathbf{\Sigma}}_{mm} + \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{aa}\hat{\mathbf{B}} - \hat{\mathbf{B}}^{\mathrm{T}}\hat{\mathbf{\Sigma}}_{am} - \hat{\mathbf{\Sigma}}_{ma}\hat{\mathbf{B}}. \qquad (3)$$

Upon substitution of the conditional maximum likelihood estimate (2) of the regression coefficients, the conditional maximum likelihood estimate of the residual covariance matrix turns out to be the Schur complement

$$\hat{\mathbf{C}} = \hat{\mathbf{\Sigma}}_{mm} - \hat{\mathbf{\Sigma}}_{ma}\hat{\mathbf{\Sigma}}_{aa}^{-1}\hat{\mathbf{\Sigma}}_{am}$$

of the submatrix $\hat{\mathbf{\Sigma}}_{aa}$ in the covariance matrix estimate

---

[1] $\mathbf{A}_{i:}$ denotes the ith row and $\mathbf{A}_{:j}$ the jth column of a matrix $\mathbf{A}$. The index i has been omitted from the symbols whose affiliation to a specific record $\mathbf{X}_{i:}$ can be inferred from the context.

[2] In principle, it suffices to estimate one regression model for each pattern of missing values in a dataset, instead of estimating one regression model for each record. However, in sets of climate data with many variables, it is rare that two or more records have the same pattern of missing values, and so the computational effort of finding matching patterns of missing values will often exceed the computational savings that result from having to estimate fewer regression models.

[3] The submatrices of the covariance matrix estimate, as well as the estimates of other quantities appearing in what follows, depend on the EM iteration t and on the record $\mathbf{X}_{i:}$ under consideration. Nevertheless, the indexes t and i have been omitted from the symbols whose affiliation to a specific iteration and to a specific record can be inferred from the context.

$\hat{\mathbf{\Sigma}}^{(t)}$ (cf. Mardia et al. 1979, chapter 6.2). As a Schur complement of a positive definite matrix $\hat{\mathbf{\Sigma}}^{(t)}$, the residual covariance matrix $\hat{\mathbf{C}}$ is assured to be positive definite (Horn and John 1985, p. 472).

The conditional expectation value $\hat{\mathbf{x}}_m \equiv E(\mathbf{x}_m | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)})$ of the missing values in a given record follows from the estimated regression coefficients $\hat{\mathbf{B}}$ and the available values $\mathbf{x}_a$ as

$$\hat{\mathbf{x}}_m = \hat{\mu}_m + (\mathbf{x}_a - \hat{\mu}_a)\hat{\mathbf{B}}, \tag{4}$$

where the vector $\hat{\mu}_a$ is that part of the mean estimate $\hat{\mu}^{(t)}$ that belongs to the variables for which, in the given record, the values are available, and the vector $\hat{\mu}_m$ is that part of the mean estimate $\hat{\mu}^{(t)}$ that belongs to the variables for which, in the given record, the values are missing.

After the missing values in all records $\mathbf{x} = \mathbf{X}_{i:}$ ($i = 1, \ldots, n$) have thus been filled in with imputed values $\hat{\mathbf{x}}_m$, the sample mean

$$\hat{\mu}^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{i:} \tag{5}$$

of the completed dataset is a new estimate of the mean of the records. A new estimate of the covariance matrix follows from the conditional expectation of the cross-products $\hat{\mathbf{S}}_i^{(t)} \equiv E[\mathbf{X}_{i:}^{\mathsf{T}}\mathbf{X}_{i:} | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}]$ as

$$\hat{\mathbf{\Sigma}}^{(t+1)} = \frac{1}{\tilde{n}} \sum_{i=1}^{n} \{\hat{\mathbf{S}}_i^{(t)} - [\hat{\mu}^{(t+1)}]^{\mathsf{T}}\hat{\mu}^{(t+1)}\}, \tag{6}$$

where, for each record $\mathbf{x} = \mathbf{X}_{i:}$, the conditional expectation $\hat{\mathbf{S}}_i^{(t)}$ of the cross-products is composed of three parts. The two parts that involve the available values in the record,

$$E(\mathbf{x}_a^{\mathsf{T}}\mathbf{x}_a | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}) = \mathbf{x}_a^{\mathsf{T}}\mathbf{x}_a \quad \text{and} \tag{7}$$

$$E(\mathbf{x}_a^{\mathsf{T}}\mathbf{x}_m | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}) = \mathbf{x}_a^{\mathsf{T}}\hat{\mathbf{x}}_m, \tag{8}$$

are sample cross-products of values in the completed record. The part that involves exclusively the imputed values in the record,

$$E(\mathbf{x}_m^{\mathsf{T}}\mathbf{x}_m | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)}) = \hat{\mathbf{x}}_m^{\mathsf{T}}\hat{\mathbf{x}}_m + \hat{\mathbf{C}}, \tag{9}$$

is the sum of the cross-product of the imputed values and the residual covariance matrix $\hat{\mathbf{C}} = \mathrm{Cov}(\mathbf{x}_m, \mathbf{x}_m | \mathbf{x}_a; \hat{\mu}^{(t)}, \hat{\mathbf{\Sigma}}^{(t)})$, the conditional covariance matrix of the imputation error. The normalization constant $\tilde{n}$ of the covariance matrix estimate (6) is the number of degrees of freedom of the sample covariance matrix of the completed dataset. If, as above, one mean vector μ is estimated, the number of degrees of freedom is $\tilde{n} = n - 1$. If, as will be described in section 4c, S mean vectors of S groups of records (for example, S seasonal mean vectors) are estimated, the number of degrees of freedom is $\tilde{n} = n - S$. The covariance matrix (6) is computed with the factor $1/\tilde{n}$ in place of the factor $1/n$ with which a maximum likelihood estimate would be computed, in order to correct the bias of the maximum likelihood estimate in a manner that parallels the bias-cor-

rection in the case of a complete dataset (cf. Beale and Little 1975). Thus, the new estimate (6) of the covariance matrix is computed in the same way as the sample covariance matrix of the completed dataset, except that, for each record with missing values, the estimated residual covariance matrix $\hat{\mathbf{C}}$ is added to the cross-products $\hat{\mathbf{x}}_m^{\mathsf{T}}\hat{\mathbf{x}}_m$ of the imputed values (cf. Little and Rubin 1987, chapter 8).

The next iteration of the EM algorithm is carried out with the updated estimates $\hat{\mu}^{(t+1)}$ and $\hat{\mathbf{\Sigma}}^{(t+1)}$ of the mean and of the covariance matrix. The iterations are stopped when the algorithm has converged, that is, when the estimates $\hat{\mu}^{(t)}$ and $\hat{\mathbf{\Sigma}}^{(t)}$ and the imputed values $\hat{\mathbf{x}}_m$ stop changing appreciably. The EM algorithm converges monotonically in that the likelihood of the available data increases monotonically from iteration to iteration. However, the EM algorithm converges only linearly, with a rate of convergence that depends on the fraction of values that are missing in the dataset, and so it may need many iterations to converge [see Little and Rubin (1987, chapters 7 and 8) for a more rigorous derivation and properties of the EM algorithm].

If, for any record, the number p_a of variables with available values is greater than the number $\tilde{n}$ of degrees of freedom available for the estimation of the covariance matrix, the submatrix $\hat{\mathbf{\Sigma}}_{aa}$ of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ is singular and the conditional maximum likelihood estimate (2) of the matrix of regression coefficients $\mathbf{B}$ is not defined. The submatrix $\hat{\mathbf{\Sigma}}_{aa}$ of the covariance matrix estimate may already be poorly conditioned if the number $\tilde{n}$ of degrees of freedom only marginally exceeds the number p_a of available values in a record. In such ill-posed or ill-conditioned cases, it is necessary to regularize the estimate (2) of the regression coefficients.

3. The regularized EM algorithm

The regularized EM algorithm consists of the same steps as the EM algorithm, with the exception that, in each iteration and for each record with missing values, the inverse matrix $\hat{\mathbf{\Sigma}}_{aa}^{-1}$ in the estimate (2) of the regression coefficients is replaced with a regularized inverse

$$\hat{\mathbf{\Sigma}}_{aa}^{-1} \leftarrow (\hat{\mathbf{\Sigma}}_{aa} + h^2\hat{\mathbf{D}})^{-1}, \tag{10}$$

where $\hat{\mathbf{D}} = \mathrm{Diag}(\hat{\mathbf{\Sigma}}_{aa})$ is the diagonal matrix consisting of the diagonal elements of the covariance matrix $\hat{\mathbf{\Sigma}}_{aa}$ and the scalar h is a regularization parameter. That is, the ill-defined or ill-conditioned inverse $\hat{\mathbf{\Sigma}}_{aa}^{-1}$ is replaced with the inverse of the matrix that results from the covariance matrix $\hat{\mathbf{\Sigma}}_{aa}$ when the diagonal elements are inflated by the factor $1 + h^2$. This method of regularizing the inverse of a matrix, in which a regularized inverse is formed as the inverse of the sum of the matrix and a multiple of a positive definite matrix, is called ridge regression in the statistics literature and Tikhonov regularization in the literature on numerical linear algebra

[Hoerl and Kennard (1970a,b); Tikhonov and Arsenin (1977); see Hansen (1997, chapter 5) for a review and Tarantola (1987, chapter 1) for a Bayesian justification]. In statistics, the regularization parameter h is known as the ridge parameter.

First, we will develop a representation of the regularized estimates of the regression parameters that makes some properties of ridge regression manifest and leads to a procedure for computing the regression parameters in the regularized EM algorithm. Second, we will describe a criterion for the choice of the regularization parameter h. Third, we will juxtapose two variants of ridge regression, both of which can be used in the regularized EM algorithm. A more detailed discussion of the methods presented below can be found in the referenced literature.

a. Ridge regression

In terms of the correlation matrix

$$\hat{\mathbf{\Sigma}}'_{aa} \equiv \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{\Sigma}}_{aa}\hat{\mathbf{D}}^{-1/2}$$

and the scaled cross-covariance matrix

$$\hat{\mathbf{\Sigma}}'_{am} \equiv \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{\Sigma}}_{am},$$

the regularized estimate of the regression coefficients can be written as

$$\hat{\mathbf{B}}_h = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{B}}'_h, \quad (11)$$

where

$$\hat{\mathbf{B}}'_h \equiv (\hat{\mathbf{\Sigma}}'_{aa} + h^2\mathbf{I})^{-1}\hat{\mathbf{\Sigma}}'_{am} \quad (12)$$

is termed the standard form of the estimate (cf. Hansen 1997, chapter 2.3). The fact that the correlation matrix $\hat{\mathbf{\Sigma}}'_{aa}$ and the scaled cross-covariance matrix $\hat{\mathbf{\Sigma}}'_{am}$ can be factored in similar ways can be exploited to cast the problem of estimating the regression coefficients from scaled submatrices $\hat{\mathbf{\Sigma}}'_{aa}$ and $\hat{\mathbf{\Sigma}}'_{am}$ of a given covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ into the more conventional form of estimating regression coefficients from given data matrices. This recasting of the estimation problem will lead to a representation of the regularized regression coefficients that makes some properties of ridge regression manifest and translates into a procedure for computing the regression coefficients in the regularized EM algorithm.

The correlation matrix $\hat{\mathbf{\Sigma}}'_{aa}$, the scaled cross-covariance matrix $\hat{\mathbf{\Sigma}}'_{am}$, and the submatrix $\hat{\mathbf{\Sigma}}_{mm}$ of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ can be decomposed into factors $\mathbf{X}_a \in \Re^{\tilde{n} \times p_a}$ and $\mathbf{X}_m \in \Re^{\tilde{n} \times p_m}$, such that

$$\hat{\mathbf{\Sigma}}'_{aa} = \mathbf{X}_a^{\mathrm{T}}\mathbf{X}_a/\tilde{n}, \qquad \hat{\mathbf{\Sigma}}'_{am} = \mathbf{X}_a^{\mathrm{T}}\mathbf{X}_m/\tilde{n} \quad \text{and} \quad (13a)$$

$$\hat{\mathbf{\Sigma}}_{mm} = \mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m/\tilde{n}. \quad (13b)$$

The factors $\mathbf{X}_a$ and $\mathbf{X}_m$ can be viewed as analogues of data matrices whose second-moment matrices $\mathbf{X}_a^{\mathrm{T}}\mathbf{X}_a/\tilde{n}$, $\mathbf{X}_a^{\mathrm{T}}\mathbf{X}_m/\tilde{n}$, and $\mathbf{X}_m^{\mathrm{T}}\mathbf{X}_m/\tilde{n}$ are the scaled submatrices (13) of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$.

The sampling error of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ contributes to the error of the imputed values and hence will play a role in determining the regularization parameter h (see section 3b and the appendix). Let us assume that the sampling error of the covariance matrix estimate is equal to the sampling error that would be expected if the dataset were complete and if the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ were the sample covariance matrix. The distribution of the sampling error of a sample covariance matrix is a function of the number $\tilde{n}$ of degrees of freedom available for the estimation of the covariance matrix (see, e.g., Mardia et al. 1979, chapter 3.4), and so, in order for the assumed sampling error of the scaled submatrices (13) to be equal to the sampling error that would be expected for second moment matrices of actual data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$, it is necessary that the number of rows of the factors $\mathbf{X}_a$ and $\mathbf{X}_m$ be equal to the number $\tilde{n}$ of degrees of freedom. That is, the factors $\mathbf{X}_a$ and $\mathbf{X}_m$ must have $\tilde{n} = n - 1$ rows if one mean vector μ is estimated from the dataset $\mathbf{X}$ and $\tilde{n} = n - S$ rows if S mean vectors of S groups of records are estimated (see section 4c).

The factorization (13) of the scaled submatrices can, for instance, be obtained from an eigendecomposition $\hat{\mathbf{\Sigma}}^{(t)} = \mathbf{T}\mathbf{\Phi}^2\mathbf{T}^{\mathrm{T}}$ of the covariance matrix estimate, with a matrix $\mathbf{T} \in \Re^{p \times \tilde{n}}$ containing as its columns the mutually orthogonal eigenvectors of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ and with a diagonal matrix $\mathbf{\Phi}^2 = \text{Diag}(\phi_j^2)$ of eigenvalues ϕ_j^2 ($j = 1, \ldots, \tilde{n}$). Let the submatrix $\mathbf{T}_a \in \Re^{p_a \times \tilde{n}}$ consist of those rows of the eigenvector matrix $\mathbf{T}$ that belong to the variables for which, in the record under consideration, the values are available, and let the submatrix $\mathbf{T}_m \in \Re^{p_m \times \tilde{n}}$ consist of the remaining rows of the eigenvector matrix $\mathbf{T}$ that belong to the variables for which, in the record under consideration, the values are missing. In terms of the partitioned eigendecomposition of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$, the factors $\mathbf{X}_a$ and $\mathbf{X}_m$ can be written as

$$\mathbf{X}_a = \sqrt{\tilde{n}}\mathbf{\Phi}\mathbf{T}_a^{\mathrm{T}}\hat{\mathbf{D}}^{-1/2} \quad \text{and} \quad \mathbf{X}_m = \sqrt{\tilde{n}}\mathbf{\Phi}\mathbf{T}_m^{\mathrm{T}}, \quad (14)$$

which shows that a factorization of the form (13) exists. If the number p of variables is greater than or equal to the number $\tilde{n}$ of degrees of freedom available for the estimation of the covariance matrix, the number $\tilde{n}$ of degrees of freedom is just the number of nonzero eigenvalues ϕ_j^2 of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$. If the number p of variables is less than the number $\tilde{n}$ of degrees of freedom, the number of nonzero eigenvalues ϕ_j^2 is less than the number $\tilde{n}$ of degrees of freedom. In this latter case, the factors $\mathbf{X}_a$ and $\mathbf{X}_m$ could be a product of the above form (14), provided that the matrix $\mathbf{\Phi}$ with the square roots of the eigenvalues ϕ_j^2 is completed with zeros to have $\tilde{n}$ rows. However, the form of the factors is irrelevant for the present argument. What is relevant is that a factorization (13) of the scaled submatrices of the covariance matrix estimate $\hat{\mathbf{\Sigma}}^{(t)}$ exists.

The factors $\mathbf{X}_a$ and $\mathbf{X}_m$ can be interpreted as the data matrices in the linear regression model

$$\mathbf{X}_m = \mathbf{X}_a \mathbf{B}' + \mathbf{E}, \qquad (15)$$

where $\mathbf{E} \in \mathfrak{R}^{\tilde{n} \times p_m}$ is a matrix of residuals. From the factorization (13) of the scaled submatrices $\hat{\boldsymbol{\Sigma}}'_{aa}$ and $\hat{\boldsymbol{\Sigma}}'_{am}$ of the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$ follows that estimating the regression coefficients $\mathbf{B}'$ of the regression model (15) from given data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$ is equivalent to estimating the standard form $\mathbf{B}' = \mathbf{D}^{1/2}\mathbf{B}$ of the regression coefficients of the model (1) from a given covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$. The standard form $\hat{\mathbf{B}}'_h = (\hat{\boldsymbol{\Sigma}}'_{aa} + h^2\mathbf{I})^{-1}\hat{\boldsymbol{\Sigma}}'_{am}$ of the regularized regression coefficients expressed in terms of the submatrices of the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$ is identical to the standard form $\hat{\mathbf{B}}'_h = (\mathbf{X}_a^{\mathrm{T}}\mathbf{X}_a + \tilde{n}h^2\mathbf{I})^{-1}\mathbf{X}_a^{\mathrm{T}}\mathbf{X}_m$ of the regularized regression coefficients expressed in terms of the data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$. Moreover, for any estimate $\hat{\mathbf{B}}'$ of the standard form regression coefficients $\mathbf{B}'$, the second-moment matrix $\hat{\mathbf{E}}^{\mathrm{T}}\hat{\mathbf{E}}/\tilde{n}$ of the estimated residuals

$$\hat{\mathbf{E}} = \mathbf{X}_m - \mathbf{X}_a\hat{\mathbf{B}}',$$

is identical to the generic estimate (3) of the residual covariance matrix $\mathbf{C}$ of the regression model (1). Hence, estimating the regression coefficients and the residual second-moment matrix of the regression model (15) from given data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$ is equivalent to estimating the regression coefficients and the residual covariance matrix of the regression model (1) from a given covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$. Since, under the above assumptions on the sampling error of the covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$, the expected sampling errors of the estimated parameters also coincide, estimating the parameters of the regression model (1) from a given estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$ of the covariance matrix is equivalent to estimating the parameters of the regression model (15) from given data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$. This equivalence makes it possible to apply standard methods for the regularization of conventional regression models (15) to the regression model (1) figuring in the EM algorithm.

A revealing representation of the ridge regression coefficients results from a singular value decomposition of the matrix $\mathbf{X}_a$ (cf. Hansen 1997, chapter 5). Let us rescale the factors $\tilde{\mathbf{X}}_a = \mathbf{X}_a/(\tilde{n})^{1/2}$ and $\tilde{\mathbf{X}}_m = \mathbf{X}_m/\tilde{n}^{1/2}$ such that in the factorization of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa} = \tilde{\mathbf{X}}_a^{\mathrm{T}}\tilde{\mathbf{X}}_a$ and of the scaled cross-covariance matrix $\hat{\boldsymbol{\Sigma}}'_{am} = \tilde{\mathbf{X}}_a^{\mathrm{T}}\tilde{\mathbf{X}}_m$ the number $\tilde{n}$ of degrees of freedom no longer appears explicitly. Whatever form is ascribed to the rescaled factor $\tilde{\mathbf{X}}_a$, it has a singular value decomposition $\tilde{\mathbf{X}}_a = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_j)$ is the diagonal matrix of singular values λ_j. In the basis of the singular value decomposition, the correlation matrix becomes $\hat{\boldsymbol{\Sigma}}'_{aa} = \mathbf{V}\boldsymbol{\Lambda}^2\mathbf{V}^{\mathrm{T}}$, which implies that the squared singular values λ_j^2 are the eigenvalues of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ and that the right singular vectors $\mathbf{V}_{\cdot j}$, the columns of the matrix $\mathbf{V}$, are the corresponding eigenvectors (see, e.g., Golub and van Loan 1993, chapter 2.5). Substituting the factorization (13) and the singular value decomposition of the rescaled

factor $\tilde{\mathbf{X}}_a$ into the standard form estimate (12) yields the representation

$$\hat{\mathbf{B}}'_h = \mathbf{V}\,\mathrm{Diag}\!\left(\frac{\lambda_j}{\lambda_j^2 + h^2}\right)\mathbf{F}, \qquad (16)$$

of the regression coefficients. The elements of the matrix $\mathbf{F} \equiv \mathbf{U}^{\mathrm{T}}\tilde{\mathbf{X}}_m$ are called Fourier coefficients, in analogy to inverse problems in which the counterpart of the matrix $\tilde{\mathbf{X}}_a$ is a convolution operator whose singular value decomposition is equivalent to a Fourier expansion (cf. Wahba 1977; Anderssen and Prenter 1981).

The representation (16) of the regression coefficients shows that, in the standard form, the columns of the regression coefficient matrix $\hat{\mathbf{B}}'_h$ are linear combinations of the eigenvectors $\mathbf{V}_{\cdot j}$ of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$. Only the eigenvectors $\mathbf{V}_{\cdot j}$ belonging to nonzero eigenvalues λ_j^2 contribute to the regression coefficients. The weights of the eigenvectors $\mathbf{V}_{\cdot j}$ are given by the products of the scalars $\lambda_j/(\lambda_j^2 + h^2)$ and the Fourier coefficients $\mathbf{F}_{j\cdot}$, which implies that only those rows $\mathbf{F}_{j\cdot}$ of the Fourier coefficient matrix that belong to nonzero eigenvalues λ_j^2 contribute to the regression coefficients.

The Fourier coefficients can be expressed in terms of the scaled cross-covariance matrix $\hat{\boldsymbol{\Sigma}}'_{am}$ and of the nonzero eigenvalues and corresponding eigenvectors of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$. Since, in terms of the singular value decomposition of the rescaled factor $\tilde{\mathbf{X}}_a$, the scaled cross-covariance matrix $\hat{\boldsymbol{\Sigma}}'_{am} = \tilde{\mathbf{X}}_a^{\mathrm{T}}\tilde{\mathbf{X}}_m$ can be written as $\hat{\boldsymbol{\Sigma}}'_{am} = (\mathbf{V}\boldsymbol{\Lambda}\mathbf{U}^{\mathrm{T}})\tilde{\mathbf{X}}_m = \mathbf{V}\boldsymbol{\Lambda}\mathbf{F}$, we can take

$$\mathbf{F} = \boldsymbol{\Lambda}^+\mathbf{V}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}'_{am} \qquad (17)$$

as the matrix of Fourier coefficients, the diagonal matrix $\boldsymbol{\Lambda}^+ = \mathrm{Diag}(\lambda_j^+)$ being the pseudoinverse of the singular value matrix $\boldsymbol{\Lambda}$; that is, the diagonal elements of the pseudoinverse $\boldsymbol{\Lambda}^+$ are $\lambda_j^+ = 1/\lambda_j$ if $\lambda_j > 0$ and $\lambda_j^+ = 0$ if $\lambda_j = 0$. [In actual computations, an element λ_j^+ of the pseudoinverse should be set to zero if the singular value λ_j is smaller than a threshold value ε that depends on the machine precision; see, e.g., Golub and van Loan (1993, chapter 5).] If the jth eigenvalue λ_j^2 of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ is zero, the jth row $\mathbf{F}_{j\cdot}$ of the Fourier coefficient matrix (17) consists of zeros and might thus differ from the jth row of the matrix $\mathbf{U}^{\mathrm{T}}\tilde{\mathbf{X}}_m$ that was originally defined to be the matrix of Fourier coefficients. But since all other rows of these matrices—the rows belonging to nonzero eigenvalues λ_j^2—agree, the differences in the rows belonging to zero eigenvalues do not affect the estimate (16) of the regression coefficients.

Thus, we can compute the regression coefficients $\hat{\mathbf{B}}'_h$ from the partitioned covariance matrix estimate $\hat{\boldsymbol{\Sigma}}^{(t)}$ as a product (16) that involves the nonzero eigenvalues and corresponding eigenvectors of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ and the Fourier coefficients (17). If there are $\tilde{n}$ degrees of freedom for the estimation of the covariance matrix $\boldsymbol{\Sigma}$ and p_a available values in the record for which the regression parameters are estimated, the number r of nonzero eigenvalues of the correlation matrix

is at most $\tilde{n}$ or p_a, whichever is smaller. Henceforth, we let the eigenvalue matrix $\Lambda^2 \in \Re^{r \times r}$ and the eigenvector matrix $\mathbf{V} \in \Re^{p_a \times r}$ contain only the r nonzero eigenvalues and corresponding eigenvectors, and we similarly restrict the Fourier coefficient matrix $\mathbf{F} \in \Re^{r \times p_m}$ to the r relevant rows. The expression (16) for the standard form estimate of the regression coefficients remains valid with these restricted matrices.

The covariance matrix of the residuals, which, in updating the covariance matrix estimate at the end of each EM iteration, is added to the cross-products (9) of the completed data matrix, can also be represented in a factored form. Substituting the estimate $\hat{\mathbf{B}}_h = \hat{\mathbf{D}}^{-1/2}\hat{\mathbf{B}}'_h$ of the regression coefficients into the generic expression (3) for the residual covariance matrix yields the estimate

$$\hat{\mathbf{C}}_h = \hat{\mathbf{C}}_0 + \mathbf{F}^{\mathrm{T}}\,\mathrm{Diag}\left[\frac{h^4}{(\lambda_j^2 + h^2)^2}\right]\mathbf{F}. \qquad (18)$$

The term

$$\hat{\mathbf{C}}_0 \equiv \hat{\mathbf{\Sigma}}_{mm} - \mathbf{F}^{\mathrm{T}}\mathbf{F},$$

which is independent of the regularization parameter h, vanishes if the regression coefficients are not overdetermined, which is the case if the number $\tilde{n}$ of degrees of freedom for the estimation of the covariance matrix $\mathbf{\Sigma}$ is less than or equal to the number p_a of variables with available values. Since the residual covariance matrix depends on the regularization method and on the regularization parameter, both of which cannot usually be chosen a priori, without reference to the dataset under consideration, the residual covariance matrix is not, as in the well-posed case, the conditional covariance matrix of the imputation error. The uncertainties about the adequacy of the regularization method and the regularization parameter contribute to the conditional imputation error given the estimates of the mean and of the covariance matrix, but the residual covariance matrix does not account for these uncertainties. Nevertheless, substituting the residual covariance matrix (18) for the conditional covariance matrix of the imputation error in updating the covariance matrix estimate at the end of each EM iteration seems a plausible heuristic.

The representation (16) makes manifest the way in which ridge regression regularizes the regression coefficients, that is, the way in which the noise, the high-frequency or small-scale components of the data, is filtered out. Both the regression coefficients regularized by a truncated principal component analysis of the correlation matrix $\hat{\mathbf{\Sigma}}'_{aa}$ and the regression coefficients regularized by ridge regression can be written as

$$\hat{\mathbf{B}}'_h = \mathbf{V}\,\mathrm{Diag}(f_j)\Lambda^+\mathbf{F},$$

where what are called the filter factors f_j depend on the regularization method (Hansen 1997, chapter 4.2). For principal component regression, the filter factors f_j of the retained principal component vectors (EOFs) $\mathbf{V}_{.j}$ are unity, and the filter factors of the discarded principal component vectors are zero. Regularization by a truncated principal component analysis of the correlation matrix $\hat{\mathbf{\Sigma}}'_{aa}$, which is what applied mathematicians call regularization by truncated singular value decomposition (cf. Schneider and Griffies 1999; Hansen 1997, chapter 3), corresponds to filtering with a stepfunction filter. For ridge regression, the filter factors are

$$f_j = \frac{\lambda_j^2}{\lambda_j^2 + h^2}. \qquad (19)$$

This filter function is structurally identical to the Wiener filter. The eigenvalues λ_j^2 are the correlate of the spectral density of what is called the signal in Wiener filtering, and the squared regularization parameter h^2 is the correlate of the spectral density of what is called the noise in Wiener filtering (Papoulis 1991, chapter 14.1; Anderssen and Prenter 1981). The filter function of ridge regression decays more slowly with decreasing eigenvalues λ_j^2 than the step function filter of principal component regression. Principal component vectors with eigenvalues λ_j^2 much greater than the squared regularization parameter h^2 are unaffected by the filtering. Principal component vectors with eigenvalues λ_j^2 much smaller than the squared regularization parameter h^2 are effectively filtered out (Hansen 1997, chapter 4.2).

For typical climate data, which do not have an evident gap in the eigenvalue spectrum and whose samples are so small that only a few principal components can be retained in a truncated principal component analysis, leaving only a small choice of possible truncation parameters, the smoother filtering afforded by ridge regression and the greater flexibility of a continuous regularization parameter could offer advantages over principal component regression. The structural parallels between the ridge regression filter and the optimal Wiener filter moreover suggest that ridge regression might suppress noise in the data in a more robust way and with less loss of relevant information than principal component regression (cf. Anderssen and Prenter 1981). Indeed, ridge regression also arises as a regularization method when the observational error in the available data, which is ignored in the regression model (1), is explicitly taken into account (Golub et al. 2000). In the regression model (1), the available values $\mathbf{x}_a$ in a record are taken as known and observational errors are neglected, but ridge regression in the form presented here is still an adequate regularization method if the available values are affected by a non-negligible observational error whose relative variance—the variance of the observational error relative to the variance of the observed variable—is homogeneous throughout the dataset. By choosing a regularized inverse (10) with a different matrix $\hat{\mathbf{D}}$, one that, in contrast to the matrix $\hat{\mathbf{D}}$ above, does not consist of the diagonal elements of the covariance matrix $\hat{\mathbf{\Sigma}}_{aa}$, other variance structures of the observational error can be accommodated in ridge regression (cf. Golub et al. 2000). To be sure, observational errors

are also taken into account in a regularization method known as truncated total least squares, in which regression coefficients are computed in a truncated basis of principal components of the overall covariance matrix $\hat{\boldsymbol{\Sigma}}^{(i)}$ instead of the scaled submatrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ (Fierro et al. 1997). But the continuous regularization parameter of ridge regression might still offer advantages over a truncated principal component analysis when there is only a small choice of possible truncation parameters.

b. Generalized cross-validation

In the regularized EM algorithm, the estimated regression coefficients are not of interest in their own right but only as intermediaries in the imputation of missing values. As a criterion for the choice of the regularization parameter h, it is therefore suitable to require that the error of the imputed values be as small as possible. As the regularization parameter tends to zero, the imputed values are increasingly affected by noise, implying an increasing sampling error. Conversely, as the regularization parameter tends to infinity, the ridge regression coefficients tend to zero and the imputed values (4) tend to the estimated mean values, implying an increasing regularization error (cf. Hansen 1997, chapter 7). A good choice of regularization parameter, in between the limiting cases of zero and infinity, should minimize the total imputation error, the sum of the sampling error and the regularization error.

Golub et al. (1979) argued that the regularization parameter h that minimizes the expected mean-squared error of predictions with an estimated linear regression model (15) is approximately equal to the minimizer of the generalized cross-validation (GCV) function

$$\mathcal{G}(h) \equiv \tilde{n} \frac{\|\mathbf{X}_a \hat{\mathbf{B}}'_h - \mathbf{X}_m\|_F^2}{\text{tr}(\mathbf{I} - \mathbf{X}_a \mathbf{X}_a^\dagger)^2}, \qquad (20)$$

an object function that resembles the object function of ordinary cross-validation but is, in contrast to the latter, invariant under orthogonal transformations of the data [see Wahba (1990, chapter 4) and Hansen (1997, chapter 7.4) for reviews]. The notations $\|\mathbf{A}\|_F$ and tr $\mathbf{A}$ indicate the Frobenius norm and the trace of a matrix $\mathbf{A}$,[4] and the matrix

$$\mathbf{X}_a^\dagger \equiv (\mathbf{X}_a^T \mathbf{X}_a + \tilde{n} h^2 \mathbf{I})^{-1} \mathbf{X}_a^T$$

in the denominator of the GCV function is the regularized pseudoinverse of the data matrix $\mathbf{X}_a$. With the regularized pseudoinverse $\mathbf{X}_a^\dagger$ of the data matrix $\mathbf{X}_a$, the regularized regression coefficients of the model (15) can be written

as $\hat{\mathbf{B}}'_h = \mathbf{X}_a^\dagger \mathbf{X}_m$, which, if the data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$ are again regarded as the factors in the decomposition (13) of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ and of the scaled cross-covariance matrix $\hat{\boldsymbol{\Sigma}}'_{am}$, is identical to the standard form (12) of the regularized regression coefficients. Under the assumptions of section 3a on the sampling error of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ and of the scaled cross-covariance matrix $\hat{\boldsymbol{\Sigma}}'_{am}$, the sampling error of the regularized regression coefficients is equal to the sampling error that would be expected if the regularized regression coefficients were estimated from actual data matrices $\mathbf{X}_a$ and $\mathbf{X}_m$ with $\tilde{n}$ records, so that the regularization parameter h that minimizes the expected mean-squared error of the imputed values is likewise approximately equal to the minimizer of the GCV function (20). Therefore, in each iteration of the regularized EM algorithm, the regularization parameter h for each record with missing values is chosen as the minimizer of the GCV function (20).

An alternative form of the GCV function follows from the eigendecomposition of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ and the derived representations of the regression coefficients and of the residual covariance matrix. Since the squared Frobenius norm of a matrix is equal to the trace of the product of the matrix and its transpose, $\|\mathbf{A}\|_F^2 = \text{tr}(\mathbf{A}^T \mathbf{A})$, the squared Frobenius norm $\|\mathbf{X}_a \hat{\mathbf{B}}'_h - \mathbf{X}_m\|_F^2 = \|\hat{\mathbf{E}}\|_F^2$ in the numerator of the GCV function is proportional to the trace of the residual covariance matrix $\hat{\mathbf{C}}_h = \hat{\mathbf{E}}^T \hat{\mathbf{E}} / \tilde{n}$. Hence, the GCV function can be written as

$$\mathcal{G}(h) = \frac{\tilde{n}^2}{\mathcal{T}^2(h)} \text{tr} \, \hat{\mathbf{C}}_h$$

where

$$\mathcal{T}(h) = \text{tr}(\mathbf{I} - \mathbf{X}_a \mathbf{X}_a^\dagger),$$

an effective number of degrees of freedom for the estimation of the residual covariance matrix $\hat{\mathbf{C}}_h$ (Wahba 1990, chapter 4), can be expressed in terms of the filter factors (19) as

$$\mathcal{T}(h) = \tilde{n} - \sum_{j=1}^{r} f_j \qquad (21)$$

(Hansen 1997, chapter 7.2). For a given regularization parameter h, evaluating both the trace tr $\hat{\mathbf{C}}_h$ of the residual covariance matrix from the factored representation (18) and the effective number of degrees of freedom $\mathcal{T}(h)$ from the filter factors (19) requires $O(r)$ operations, where r is the number of nonzero eigenvalues of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$ (cf. Hansen 1994, 1997, chapter 4.6). That is, if the ridge regression is computed via an eigendecomposition of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$, only a small additional effort is required to find, with one of the common scalar optimization methods, the regularization parameter h that minimizes the GCV function.

With the regularization parameter determined by generalized cross-validation, the regularized estimates of the imputed values are usually reliable, even when the noise in the data, which might be a result of observa-

---

[4] The GCV function is usually defined for a single dependent variable (i.e., $p_m = 1$), so that the data matrix $\mathbf{X}_m$ and the regression coefficients $\hat{\mathbf{B}}'_h$ are vectors, for which the Frobenius norm in the numerator of the GCV function (20) becomes the Euclidean norm. Replacing the Euclidean norm for vectors with the Frobenius norm for matrices leads to the extension of standard results presented here.

tional errors, is not Gaussian and has an inhomogeneous variance structure (see, e.g., Wahba 1990, chapter 4.9). Since with small but nonzero probability the GCV function has a minimum near zero, generalized cross-validation occasionally leads to a regularization parameter near zero when, in fact, a greater regularization parameter would be more appropriate (Wahba and Wang 1995). Choosing too small a regularization parameter in such cases can be avoided by constructing a lower bound for the regularization parameter from a priori guesses of the magnitude of the imputation error (see, e.g., Hansen 1997, chapters 7.7 and 7.2).

c. Multiple and individual ridge regressions

If ridge regression with generalized cross-validation is used in the regularized EM algorithm as described above, the regularization of the regression coefficients is controlled by one regularization parameter per record with missing values. For each record, the regression coefficients of all variables with missing values are estimated jointly by multiple ridge regression. With generalized cross-validation, the regularization parameter is chosen such as to minimize, approximately, the expected mean-squared error of the imputed values.

However, with the above methods, it is also possible to estimate individually regularized regression coefficients for each missing value. The matrix of regression coefficients (11) can be computed columnwise with an individual regularization parameter for each column. Instead of only one regularization parameter per record in multiple ridge regressions, in individual ridge regressions we can, for a record with p_m missing values, adjust p_m regularization parameters. Choosing the regularization parameter for each column of the regression coefficient matrix by generalized cross-validation approximately minimizes not only the expected average error of the imputed values in the record, but also the expected error of each individual imputed value.

The computation of individual ridge regressions is similar to the computation of a multiple ridge regression. If the ridge regression is computed via an eigendecomposition of the correlation matrix $\hat{\boldsymbol{\Sigma}}'_{aa}$, one obtains the standard form estimate (16) of the regression coefficients columnwise from the columns of the Fourier coefficient matrix (17), with an individual regularization parameter h_j ($j = 1, \ldots, p_m$) for each column. The regularization parameters h_j are determined as the minimizers of the GCV function (20), where the numerator of the GCV function reduces from the squared Frobenius norm of a residual matrix to the squared Euclidean norm of a residual vector. Generalizing the factored representation (18) of the residual covariance matrix from the case of a multiple ridge regression to that of individual ridge regressions, one finds that the residual covariance matrix of individual ridge regressions consists of the elements

$$(\hat{\mathbf{C}}_h)_{kl} = (\hat{\mathbf{C}}_0)_{kl} + (\mathbf{F}_{:k})^{\mathrm{T}}\boldsymbol{\Gamma}^{(k)}\boldsymbol{\Gamma}^{(l)}\mathbf{F}_{:l}, \qquad (22)$$

where $\boldsymbol{\Gamma}^{(j)} \equiv h_j^2(\boldsymbol{\Lambda}^2 + h_j^2\mathbf{I})^{-1}$ is a diagonal matrix and h_j is the regularization parameter for the jth column of the matrix of regression coefficients. In a regularized EM algorithm with individual ridge regressions, this residual covariance matrix is added to the cross-products $\hat{\mathbf{x}}_m^{\mathrm{T}}\hat{\mathbf{x}}_m$ of the imputed values when a new estimate of the covariance matrix (6) is assembled.

Thus, the additional computational expense of individual ridge regressions in place of a multiple ridge regression is merely that which is required to minimize the GCV function p_m times for p_m residual vectors, compared with minimizing it once for one residual matrix with p_m column vectors. As long as the greater number of regularization parameters to be estimated does not lead to the estimated regularization parameters becoming unreliable, the greater accuracy of the imputed values that can be expected with individual ridge regressions suggests the use of individual ridge regressions in the regularized EM algorithm whenever computationally feasible.

4. Exploiting spatial and temporal covariability

By arranging a dataset into data matrices **X** in different ways, one can, with the regularized EM algorithm, exploit spatial and temporal covariability for the imputation of missing values. Suppose the dataset under consideration consists of time series of m geophysical variables. The time series of each of the m variables would, if the dataset were complete, span N instants (N years, for example), but the values of some of the variables are missing for some instants. The m variables could, for example, represent yearly mean surface temperatures at m stations or at m grid points. Or some of the m variables could represent local surface temperature measurements and others proxies of surface temperatures, such as dendroclimatic data or isotope fractions in ice cores, so that the regularized EM algorithm exploits the covariability of the measurements and the proxy data to impute missing measurement values given available proxy data. Let the m-dimensional row vector $\mathbf{y}_\nu$ contain the values of the m variables at instant ν ($\nu = 1, \ldots, N$). Whether the regularized EM algorithm estimates, and exploits for the imputation of missing values, the spatial covariance matrix or a mixed spatiotemporal covariance matrix depends on how the N data vectors $\mathbf{y}_\nu \in \mathfrak{R}^{1\times m}$ are arranged into the data matrix $\mathbf{X} \in \mathfrak{R}^{n\times p}$.

a. Spatial covariability

The fundamental kind of covariability that the regularized EM algorithm exploits for the imputation of missing values is spatial covariability, a synchronic covariability of the variables at a single instant ν. If the data vectors $\mathbf{y}_\nu$ form the rows of the data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_N \end{pmatrix}, \tag{23}$$

with $n = N$ and $p = m,$ the regularized EM algorithm calculates an estimate of the temporal mean μ of the data vectors $\mathbf{y}_\nu$ and an estimate of the spatial covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{y}_\nu, \mathbf{y}_\nu)$. The missing values for an instant ν are filled in with imputed values that are conditioned on the available values for the same instant ν and on the estimates of the mean and of the spatial covariance matrix.

b. Stationary temporal covariability

In addition to the synchronic spatial covariability at a single instant ν, the variables of the dataset may also possess diachronic covariability across different instants ν. This diachronic covariability can be exploited to improve the accuracy of the imputed values and of the estimates of the temporal mean and of the spatial covariance matrix of the data. If the instants for which the data are available are equally spaced in time and if the temporal variability of the data is stationary, the diachronic covariability across three neighboring instants can be taken into account in the regularized EM algorithm by arranging the data vectors $\mathbf{y}_\nu$ for three successive instants ν into the rows of the data matrix

$$\mathbf{X} = \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 \\ \vdots & \vdots & \vdots \\ \mathbf{y}_{\nu-1} & \mathbf{y}_\nu & \mathbf{y}_{\nu+1} \\ \mathbf{y}_\nu & \mathbf{y}_{\nu+1} & \mathbf{y}_{\nu+2} \\ \vdots & \vdots & \vdots \\ \mathbf{y}_{N-2} & \mathbf{y}_{N-1} & \mathbf{y}_N \end{pmatrix}, \tag{24}$$

with $n = N - 2$ and $p = 3m$. With such a data matrix, the regularized EM algorithm calculates an estimate of the spatiotemporal covariance matrix

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_0 & \mathbf{\Sigma}_1 & \mathbf{\Sigma}_2 \\ \mathbf{\Sigma}_{-1} & \mathbf{\Sigma}_0 & \mathbf{\Sigma}_1 \\ \mathbf{\Sigma}_{-2} & \mathbf{\Sigma}_{-1} & \mathbf{\Sigma}_0 \end{pmatrix}, \tag{25}$$

a block Toeplitz matrix[5] composed of the spatial covariance matrix $\mathbf{\Sigma}_0 = \mathrm{Cov}(\mathbf{y}_\nu, \mathbf{y}_\nu)$, of the lag-1 covariance matrices $\mathbf{\Sigma}_{-1} = \mathrm{Cov}(\mathbf{y}_\nu, \mathbf{y}_{\nu-1})$ and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_{-1}^{\mathrm{T}}$, and

of the lag-2 covariance matrices $\mathbf{\Sigma}_{-2} = \mathrm{Cov}(\mathbf{y}_\nu, \mathbf{y}_{\nu-2})$ and $\mathbf{\Sigma}_2 = \mathbf{\Sigma}_{-2}^{\mathrm{T}}$.

In each iteration of the regularized EM algorithm, the imputed values and the residual covariance matrix need to be computed only once for each data vector $\mathbf{y}_\nu$ ($\nu = 1, \ldots, N$) and can then be used to update both the copis of the vector $\mathbf{y}_\nu$ in the data matrix (24) and the corresponding elements of the covariance matrix estimate $\hat{\mathbf{\Sigma}}$. The imputed values in a data vector $\mathbf{y}_\nu$ are conditioned on the available values in the vector $\mathbf{y}_\nu$ itself and on the available values in two other data vectors for neighboring instants. Whether the imputed values for an instant ν are conditioned on the available values for the two preceding instants, the two succeeding instants, or one preceding and one succeeding instant depends on the position, within the data matrix, of the vector $\mathbf{y}_\nu$ for which the imputed values are computed. For example, the imputed values of a vector $\mathbf{y}_\nu$ in the first m-column block of the data matrix (24) are conditioned on the available values in the vector $\mathbf{y}_\nu$ itself and on the available values in the two succeeding vectors $\mathbf{y}_{\nu+1}$ and $\mathbf{y}_{\nu+2}$. Usually, one wants the imputed values for an instant to be conditioned on the available values for the instant itself and on the available values for the two instants nearest to it. Thus, in addition to computing the imputed values for the data vectors $\mathbf{y}_1$ and $\mathbf{y}_N$ at the ends of the dataset, which occur only once in the data matrix (24), one should compute the imputed values for the data vectors $\mathbf{y}_2, \ldots, \mathbf{y}_{N-1}$ in the central m-column block of the data matrix. The remaining elements of the data matrix and the estimate of the covariance matrix (25) should then be updated by copying the imputed values and the residual covariance matrices from the data vectors $\mathbf{y}_2, \ldots, \mathbf{y}_{N-1}$ in the central m-column block of the data matrix.

The resulting regularized EM algorithm calculates an estimate of the temporal mean of the data vectors $\mathbf{y}_\nu$ and an estimate of the spatiotemporal covariance matrix (25). The ridge regression filter (19) acts not on the EOFs, the eigenvectors of the estimated spatial correlation matrix, but on the eigenvectors of the estimated spatiotemporal correlation matrix that belongs to the spatiotemporal covariance matrix (25). The missing values for an instant ν are filled in with imputed values that are conditioned on the available values for the same instant ν, on the available values for the two instants nearest to ν, and on the estimates of the mean and of the spatiotemporal covariance matrix.

It is possible to modify the above way of exploiting spatiotemporal covariability in the regularized EM algorithm. For example, higher-order temporal covariability can be taken into account by increasing the data matrix (24) further, that is, by arranging more data vectors $\mathbf{y}_\nu$ into a row of the data matrix. The algorithm presented here is the simplest algorithm that exploits temporal covariability in a symmetric manner: in the imputation of missing values for a given instant in the interior of the dataset, the available data immediately

---

[5] That the spatiotemporal covariance matrix (25) has a block Toeplitz structure means that along its diagonals are identical copies of the spatial covariance matrix $\mathbf{\Sigma}_0$ or of the lag-1 covariance matrices $\mathbf{\Sigma}_{-1}$ and $\mathbf{\Sigma}_1$. However, from the structure of the data matrix (24) one can infer that the estimate $\hat{\mathbf{\Sigma}}$ of the covariance matrix (25) will not have a block Toeplitz structure. The estimates of the submatrices $\mathbf{\Sigma}_0$, $\mathbf{\Sigma}_{-1}$, and $\mathbf{\Sigma}_1$ along the diagonals of $\hat{\mathbf{\Sigma}}$ will differ in that they include or exclude contributions of data vectors for the first instants ($\nu = 1$ or 2) and for the final instants ($\nu = N$ or $N - 1$) of the dataset.

preceding and immediately succeeding that instant are taken into account in a manner that is invariant under time reversal.

c. Cyclostationary temporal covariability

If the dataset under consideration contains, for example, monthly mean temperatures for all months of several years, the temporal variability of the data is not stationary but cyclostationary. To take cyclostationary temporal covariability into account, it often suffices to allow for a periodically varying mean of the data. Suppose each data vector $\mathbf{y}_\nu$ ($\nu = 1, \ldots, N$) belongs to one of S regimes s ($s = 1, \ldots, S$). The regimes could be, for example, the months of the year or the seasons. The regularized EM algorithm allows for a regime-dependent mean if, in each iteration and for each record with missing values, the parts $\hat{\mu}_a$ and $\hat{\mu}_m$ of the mean vector in the estimated regression (4) are assembled not from an overall mean (5) of the records, but from regime-dependent mean values

$$\hat{\mu}_s = \frac{1}{N_s} \sum_{\nu \in s} \mathbf{y}_\nu, \qquad N_s = \sum_{\nu \in s} 1,$$

where it is understood that, for the computation of the regime-dependent mean values for a given EM iteration, the missing values in the data vectors $\mathbf{y}_\nu$ are filled in with the imputed values of the preceding EM iteration. The loss of degrees of freedom resulting from estimating S regime-dependent mean vectors, instead of only one mean vector, must be taken into account in the covariance matrix estimate (6) and in the GCV function (20) by taking $\tilde{n} = n - S$, in place of $\tilde{n} = n - 1$. Allowing for a regime-dependent mean, which is revised in each EM iteration, amounts to performing the regression (1) with the anomaly data of the deviations from the regime-dependent mean. If the spatial covariance matrix of the anomaly data is regime independent, the spatial covariability of the anomaly data can be exploited with an algorithm that is, up to the distinction of regime-dependent mean values, identical to that for spatial data (section 4a). If the spatiotemporal covariance matrix of the anomaly data is likewise regime independent, the temporal covariability can be exploited with an algorithm that is, up to the distinction of regime-dependent mean values, identical to that for stationary data (section 4b).

If a regime-dependent mean does not suffice to account for the cyclostationary variability of the data, one can also allow for a periodically varying spatiotemporal covariance matrix. It is then necessary to define a separate data matrix $\mathbf{X}_s$ for each regime s, such that the regularized EM algorithm calculates separate estimates of the mean and of the spatiotemporal covariance matrix for each regime. For each data vector $\mathbf{y}_\nu$ ($\nu = 2, \ldots, N - 1$) in the interior of the dataset, the data matrix $\mathbf{X}_s$ of the regime s to which the vector $\mathbf{y}_\nu$ belongs must

contain, like the data matrix (24) for stationary data, a row vector

$$(\mathbf{y}_{\nu-1} \quad \mathbf{y}_\nu \quad \mathbf{y}_{\nu+1})$$

comprising the data vector itself and the data vectors for the preceding and the succeeding instant. The missing values of the data matrices $\mathbf{X}_s$ and the corresponding estimates of the spatiotemporal covariance matrices must be updated in each EM iteration. As in the regularized EM algorithm for stationary data, the imputed values and the residual covariance matrices should be computed for the data vectors $\mathbf{y}_1$ and $\mathbf{y}_N$ at the ends of the dataset and for the data vectors $\mathbf{y}_2, \ldots, \mathbf{y}_{N-1}$ in the central m-column blocks of the data matrices $\mathbf{X}_s$. The remaining elements of the data matrices and the regime-dependent estimates of the spatiotemporal covariance matrix should then be updated by copying, possibly across different data matrices $\mathbf{X}_s$, the imputed values and the residual covariance matrices from the data vectors $\mathbf{y}_2, \ldots, \mathbf{y}_{N-1}$ in the central m-column blocks of the data matrices. The resulting regularized EM algorithm calculates regime-dependent estimates of the mean and of the spatiotemporal covariance matrix and, using these estimates, fills in the missing values for an instant ν with imputed values that are conditioned on the available values for the same instant ν and on the available values for the two instants nearest to ν.

The values imputed with a regularized EM algorithm that exploits one of the above forms of temporal covariability are potentially more accurate than the values imputed with a regularized EM algorithm that exploits only spatial covariability. For a given set of m-dimensional data, the number p of variables in the data matrix and the number S of regimes distinguished can be viewed as discrete regularization parameters that control the level of detail with which temporal covariability is taken into account. As the size of the data matrix increases and more regimes are distinguished, the estimated temporal covariance information becomes more detailed, but, per covariance matrix element to be estimated, the number of degrees of freedom decreases. Moreover, the computational complexity of the regularized EM algorithm increases with the level of detail of the estimated covariance information. Whether employing a more complex rather than a simpler regularized EM algorithm is worth the added computational effort can be checked by comparing the values of the GCV function, and thus comparing rough estimates of the errors of the imputed values (see the appendix), for various compositions of the data matrix.

5. Other imputation techniques for climate data

Conventional techniques for the imputation of missing values in climate data, such as the techniques with which Smith et al. (1996), Kaplan et al. (1997, 1998), and Mann et al. (1998, 1999) impute missing values in

surface temperature data, differ from the regularized EM algorithm in two respects.

First, the conventional techniques neglect the interdependence of the imputed values and the estimated statistics of the data. In all above-cited studies, a covariance matrix, be it a spatial or a spatiotemporal covariance matrix, is estimated from the complete or almost complete more recent records of the surface temperature datasets under consideration, whereupon missing values in the records for the years further past are filled in with imputed values that are conditioned on the estimated covariance matrix. If the statistics of the data are estimated from only a small portion of the records in the dataset, it is possible that in filling in missing values with imputed values, long-term climate variability is underestimated. The regularized EM algorithm reduces such an underestimation of long-term climate variability by allowing for the dependence of the estimated statistics of the data on all available values in the dataset, including the available values in incomplete records.

Second, the conventional techniques regularize the ill-posed or ill-conditioned problems that arise in the imputation of missing values by one or the other form of truncated principal component analysis, in which typically a single truncation parameter, or a single discrete regularization parameter, is chosen for an entire dataset. For example, Smith et al. (1996) estimate imputed temperature values for a given region by analyzing the regression of grid points with missing values on grid points with available values in a truncated basis of principal components of a covariance matrix estimate. Their regularized regression method, although it is not presented in terms of regression analysis, resembles what is known to statisticians as latent root regression and to applied mathematicians as truncated total least squares—a regularization method that arises, like ridge regression, when observational errors in the available data are explicitly taken into account (Webster and Gunst 1974; Fierro et al. 1997; van Huffel and Vandewalle 1991, chapter 3). Smith et al., as well as the other above-cited authors, choose a single regularization parameter per dataset, which means that the regularization does not adapt to the density and the pattern of the available values in the incomplete records. In the regularized EM algorithm, ridge regression regularizes the ill-posed or ill-conditioned estimation of regression coefficients, which, as argued in section 3a, might offer advantages over regularization methods that are based on truncated principal component analyses. Moreover, the continuous regularization parameter of ridge regression is chosen adaptively by generalized cross-validation, such that the regularization adapts to the density and the pattern of the available values in each incomplete record.

In that the above-cited conventional techniques estimate the statistics of the data under consideration only from a subset of the available data and regularize the ill-posed regression problems nonadaptively, the conventional techniques can be viewed as approximations to the regularized EM algorithm. Thus, on theoretical grounds one would expect that the regularized EM algorithm yields imputed values that are at least as accurate as the values imputed with one of the conventional techniques. In the limit of sufficiently large sample sizes in which regularization becomes unnecessary and the regularization parameter h can be set to zero, the regularized EM algorithm reduces to the EM algorithm, for which Dempster et al. (1977) proved some general optimality properties. However, that the EM algorithm for typical climate data involves the solution of ill-posed problems is more than a mere inconvenience that complicates the use of an otherwise optimal method: there are no general, problem-independent criteria according to which the optimality of a method for ill-posed problems can be established (Linz 1984). Hence, any claim that the regularized EM algorithm or any other technique for the imputation of missing values in climate data is ''optimal'' in some general sense would be unjustified. The performance of the regularized EM algorithm must be assessed in practice.

6. Test with simulated surface temperature data

The regularized EM algorithm was tested with sets of simulated monthly mean surface temperature data in which values were deleted in a manner characteristic for observational data. Because the missing values in the resulting sets of test data were known, the accuracy with which the regularized EM algorithm imputes missing values could be assessed. To be able to compare the performance of the regularized EM algorithm with that of a conventional noniterative imputation technique, the missing values in the test datasets were also imputed with the technique of Smith et al. (1996).

a. The imputation technique of Smith et al. (1996)

Smith et al. impute missing values in a given record of an incomplete dataset by first estimating a regression model of the same form as the regression model (1) of the EM algorithm and then filling in missing values with the values that the estimated regression model predicts given the available values in the record. The way in which Smith et al. estimate the regression models entails certain limitations on the applicability of their technique, and these limitations also restrict the extent to which their technique can be compared with the regularized EM algorithm. To understand these limitations, it is necessary to consider the technique of Smith et al. in more detail.[6]

---

[6] Smith et al. do not describe their technique in terms of regression analysis, but the equivalence of their description and the one given here can be seen by expressing the formulas in the paper by Smith et al. in terms of matrix operations. The imputation technique of Mann et al. (1998) can similarly be interpreted as a regression analysis.

Smith et al. estimate the coefficients of the regression model for a given record from a spatial covariance matrix that is the sample covariance matrix of a complete set of training data. The training dataset contains complete records of monthly mean surface temperature data for the years 1982–93, with the data given as anomaly data, that is, as the deviations of the monthly mean temperatures for a given month from the estimated climatological mean temperatures for that month. The spatial covariance matrix is estimated as the sample covariance matrix of the 144 monthly records for all months of the 12 years from 1982 through 1993. Smith et al. assume that a periodically varying mean suffices to account for the cyclostationary variability of the data and that the spatial covariance matrix of the anomaly data can be approximated by a stationary covariance matrix (cf. section 4c).

The matrix of regression coefficients **B** of the regression model (1) for a given record is computed from the estimated spatial covariance matrix by a regularization method known as truncated total least squares (Fierro et al. 1997), which leads to the estimate

$$\hat{\mathbf{B}}_q \equiv (\mathbf{T}_{aq}^{+})^{\mathrm{T}}(\mathbf{T}_{mq}^{\mathrm{T}}),$$

where $\mathbf{T}_{aq} \in \Re^{p_a \times q}$ and $\mathbf{T}_{mq} \in \Re^{p_m \times q}$ are submatrices of the matrix **T** that contains as its columns the eigenvectors of the estimated spatial covariance matrix, and $\mathbf{T}_{aq}^{+} = (\mathbf{T}_{aq}^{\mathrm{T}}\mathbf{T}_{aq})^{-1}\mathbf{T}_{aq}^{\mathrm{T}}$ is a pseudoinverse of the matrix $\mathbf{T}_{aq}$. The submatrix $\mathbf{T}_{aq}$ consists of those rows of the first q eigenvectors $\mathbf{T}_{:,1:q}$ that belong to the variables for which, in the given record, the values are available, and the submatrix $\mathbf{T}_{mq}$ consists of those rows of the first q eigenvectors $\mathbf{T}_{:,1:q}$ that belong to the variables for which, in the given record, the values are missing (cf. the partitioning of the eigenvector matrix **T** in section 3a). The truncation parameter q is a discrete regularization parameter, which Smith et al. choose with an ad hoc technique. In truncated total least squares, the regression coefficients are linear combinations of the subvectors $\mathbf{T}_{aq}$ of the q leading eigenvectors of the estimated overall covariance matrix. Fierro et al. (1997) investigate properties of regularization by truncated total least squares, and Golub et al. (2000) show how considerations of data errors can give rise both to truncated total least squares and to ridge regression.

The spatial covariance matrix and the regression coefficients are estimated from a training dataset that spans only 12 years. To account for interdecadal variability of the monthly mean temperature, Smith et al. estimate the mean vector μ of the regression model (1) for a given record as an 11-yr running mean. For the estimation of the mean values of all variables in a given record, it is therefore necessary that sufficient data are available in the 11 years surrounding the year to which the given record belongs. The running mean is smoothed spatially, in this test with a radial Gaussian smoothing filter of standard deviation 500 km, in order to stabilize the estimation of the mean vector and to fill in missing values

in the mean vector with nearby available mean values when, for any grid point, there are not sufficient data in an 11-yr window.

Thus, the technique of Smith et al. is based on different estimation and regularization procedures for the same regression model (1) that is used in the regularized EM algorithm. That the technique of Smith et al. requires the estimation of a running mean limits its applicability to datasets that do not contain variables for which values are missing in many consecutive records.

b. Test data

To obtain incomplete datasets with which the regularized EM algorithm and the imputation technique of Smith et al. could be tested, some values were deleted in nine sets of simulated surface temperature data. The simulated datasets were obtained from nine integrations of a coupled climate model of low (R15) resolution [see Dixon and Lanzante (1999) for a description of the ensemble of integrations]. The climate model was developed at the Geophysical Fluid Dynamics Laboratory and, in the past decade, variants of it have been used in several studies to simulate anthropogenic climate change (e.g., Manabe et al. 1991; Manabe and Stouffer 1994). Monthly means of the simulated surface temperature data were interpolated to the 5° × 5° latitude–longitude grid of the set of merged land surface and sea surface temperature data of Jones (1994) and Parker et al. (1994, 1995). Corresponding to the locations and times for which values are missing in this set of observational temperature data, values were deleted in the nine sets of simulated temperature data.

Requirements on the spatial and temporal continuity of data coverage in a study of interdecadal climate change—the study that motivated the development of the regularized EM algorithm—and the limitations of the technique of Smith et al. led to some restrictions on the test datasets: only the simulated monthly mean temperatures for July were considered, and the datasets were restricted temporally to the 53 simulated years from 1946 through 1998 and spatially to the region between the 25°S and the 60°N latitude circles. Within this period and this region, grid points for which more than 30% of the temperature values had been deleted were excluded from the datasets. With these restrictions on the datasets, the spatially smoothed 11-yr running mean, required by the technique of Smith et al., was available for all variables and for all records. What resulted were nine test datasets, each consisting of $N = 53$ records with simulated values of the mean July surface temperature for $m = 1156$ grid points. In each of the nine datasets, 3.3% of the values were missing.

c. Test results

With a regularized EM algorithm with individual ridge regressions, the temporal mean, the spatial co-
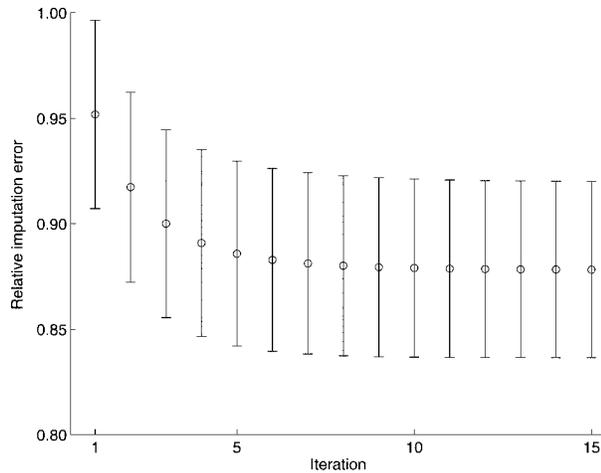
FIG. 1. Mean (circles) and standard deviation (bars) of the rms relative imputation errors for the nine datasets after each of the first 15 iterations of the regularized EM algorithm.

variance matrix, and the missing values of the simulated July temperatures were estimated from the incomplete test datasets. Only the spatial covariance matrix was estimated and exploited for the imputation of missing values (cf. section 4a). For each of the nine test datasets, the regularized EM algorithm was initialized with the mean vector estimated as the sample mean of all values available in the test dataset and with a covariance matrix estimated as the sample covariance matrix of the test dataset with estimated mean values substituted for missing values. The regularized EM algorithm was iterated until it reached the stopping criterion

$$\left[\frac{\sum\limits_{\text{miss}} (\mathbf{X}_{ij}^t - \mathbf{X}_{ij}^{t-1})^2}{\sum\limits_{\text{miss}} (\mathbf{X}_{ij}^{t-1})^2} \right]^{1/2} < 5 \times 10^{-3},$$

where $\mathbf{X}^t$ is the data matrix with the imputed values of the tth iteration filled in for missing values, and the sums extend over all missing values. The stopping criterion was reached after 14–17 iterations.[7]

Figure 1 shows the mean and the standard deviation of the root-mean-square (rms) relative imputation errors for the nine datasets after each of the first 15 iterations of the regularized EM algorithm. For a given set of test data, the rms relative imputation error after the tth EM iteration is defined as

$$\delta \mathbf{X}^t \equiv \left[\frac{1}{M} \sum_{\text{miss}} \frac{(\mathbf{X}_{ij}^t - \mathbf{X}_{ij}^c)^2}{(\sigma_j^c)^2} \right]^{1/2}, \tag{26}$$

where the matrix $\mathbf{X}^c$ contains the complete set of simulated surface temperature data from which values had

been deleted to obtain the incomplete set of test data. The normalization constant M is the total number of missing values in a dataset, and σ_j^c is the standard deviation of the jth variable of the complete dataset $\mathbf{X}^c$. The squared relative imputation error [the fraction inside the sum (26)] averaged over the nine datasets was spatially and temporally fairly homogeneous. Figure 1 shows that the rms relative imputation error averaged over the nine datasets decreased monotonically from iteration to iteration. Indeed, in the 14–17 iterations that the regularized EM algorithm needed to reach the stopping criterion, the rms relative imputation error was found to decrease monotonically for each individual dataset. With the test datasets, then, the regularized EM algorithm converged reliably, albeit slowly.

Smith et al. estimate the spatial covariance matrix needed by their imputation technique from a training dataset consisting of in situ temperature measurements that were completed with satellite data. For the comparison of their imputation technique with the regularized EM algorithm, the spatial covariance matrices needed by the technique of Smith et al. were estimated from complete datasets for all months of the simulated years 1982–93, after the climatological monthly mean of the simulated years 1961–90 had been subtracted from each month of the simulated data. To account for the expected decrease of the temperature variance with the area of the grid box that a variable of a dataset represents, the simulated data for each grid point were scaled by the square root of the cosine of the latitude of the grid point. (No such scaling is necessary in the regularized EM algorithm because the regularized EM algorithm is based on eigendecompositions of correlation matrices, not of covariance matrices.) The imputation technique of Smith et al. was tried with various truncation parameters q for the total least squares regularization. The results reported below belong to the truncation parameter $q = 20$ that, averaged over the nine datasets, yielded the smallest rms relative imputation error. Thus, in choosing the truncation parameter with the smallest rms relative imputation error, information from the complete datasets was used that would not ordinarily be available. The spatial covariance matrices, moreover, were estimated from training datasets that contained some of the values that were missing in the test datasets, whereas, in practice, the training dataset and the incomplete dataset in which missing values are to be imputed are distinct. Therefore, the conditions for the technique of Smith et al. were, in this test, more favorable than they would be in practice. The regularized EM algorithm, on the other hand, used only the kind of data that would have been available in practice, namely, only the incomplete test datasets with simulated mean temperatures for July.

Table 1 displays the relative errors that occurred with the regularized EM algorithm and with the technique of Smith et al. For each of the nine datasets, the rms relative imputation error $\delta \mathbf{X}$ of the regularized EM algorithm

---

[7] Details of the implementation of the regularized EM algorithm can be taken from the program code, which is available online (Schneider 1999).

TABLE 1. Relative errors of the regularized EM algorithm and of the technique of Smith et al. (1996) for the nine datasets: rms relative imputation error $\delta\mathbf{X}$, estimated rms relative imputation error $\widehat{\delta\mathbf{X}}$, and relative error $\delta(\mathrm{tr}\,\boldsymbol{\Sigma})$ of the trace of the estimated covariance matrix.

Data	Regularized EM			Smith et al.
	$\delta\mathbf{X}$	$\widehat{\delta\mathbf{X}}$	$\delta(\mathrm{tr}\,\boldsymbol{\Sigma})$	$\delta\mathbf{X}$
1	0.966	0.782	−0.022	0.974
2	0.844	0.765	−0.020	0.913
3	0.827	0.779	−0.017	0.882
4	0.893	0.796	−0.018	0.950
5	0.874	0.760	−0.018	0.912
6	0.868	0.787	−0.019	0.915
7	0.914	0.787	−0.019	0.922
8	0.857	0.772	−0.017	0.904
9	0.860	0.784	−0.016	0.916

was smaller than the rms relative imputation error of the technique of Smith et al. That is, although the regularized EM algorithm used only the incomplete datasets of simulated July temperatures, it led to more accurate imputed values than the technique of Smith et al., which additionally required complete sets of training data with simulated temperatures for every month of a 12-yr period and for which the truncation parameter was chosen from the complete datasets such as to minimize the rms relative imputation error averaged over the nine datasets.

The standard errors of the values imputed with the regularized EM algorithm were estimated as described in the appendix. From the estimated standard errors, the estimated rms relative imputation error $\widehat{\delta\mathbf{X}}$ was calculated in analogy to the actual rms relative imputation error (26), with the estimated squared errors replacing the actual squared errors in the numerator of the relative imputation error. The comparison of the rms relative imputation error $\delta\mathbf{X}$ and of the estimated rms relative imputation error $\widehat{\delta\mathbf{X}}$ shows that the estimated imputation error tends to underestimate the actual imputation error. The estimated rms relative imputation error was, on the average, 11% smaller than the actual rms relative imputation error.

The underestimation of the imputation error points to a general difficulty in estimating errors in ill-posed problems. Error estimates in ill-posed problems depend on the regularization method employed and on the regularization parameter, but one rarely has a priori reasons, independent of the particular dataset under consideration, for the choice of a regularization method and a regularization parameter. In addition to the uncertainty about the adequacy of the regression model (1), the uncertainties about the adequacy of the regularization method and of the regularization parameter contribute to the imputation error. Since in the estimated imputation error, these uncertainties are neglected, the estimated imputation error underestimates the actual imputation error.

Since only 3.3% of the values in the test datasets were missing, estimating the mean of the records from the

incomplete test datasets was unproblematic. The sample means of the complete datasets and the sample means of the completed test datasets with imputed values filled in for missing values did not differ significantly. In the test datasets, 500 of the $p = 1156$ variables had at least one missing value. The individual sample means of these 500 variables, estimated from the test datasets completed with the regularized EM algorithm and with the technique of Smith et al., fell within the 95% confidence intervals of the sample mean of the complete datasets. Thus, the mean values estimated with the regularized EM algorithm and with the technique of Smith et al. were statistically indistinguishable from the sample mean of the complete datasets.

Because of the small fraction of missing values in the test datasets, the differences between the variances estimated with the regularized EM algorithm from the incomplete test datasets and the sample variances of the complete datasets were also of a similar magnitude as the expected sampling errors of the sample variances. However, the mean of the differences between the variances estimated from the incomplete test datasets and the sample variances of the complete datasets was not zero. As a measure of the bias of the estimated variances, Table 1 displays the error

$$\delta(\mathrm{tr}\,\boldsymbol{\Sigma}) \equiv \frac{\mathrm{tr}\,\hat{\boldsymbol{\Sigma}} - \mathrm{tr}\,\boldsymbol{\Sigma}^c}{\mathrm{tr}\,\boldsymbol{\Sigma}^c}$$

of the trace of the spatial covariance matrix $\hat{\boldsymbol{\Sigma}}$ estimated with the regularized EM algorithm relative to the trace of the sample covariance matrix $\boldsymbol{\Sigma}^c$ of the complete set of simulated data. The spatial covariance matrix $\hat{\boldsymbol{\Sigma}}$ is the spatial covariance matrix of simulated July temperatures that the regularized EM algorithm estimates and exploits for the imputation of missing values. That the trace of the estimated covariance matrix was, for all nine sets of test data, smaller than the trace of the sample covariance matrix of the complete datasets indicates that the regularized EM algorithm underestimates the variances. This underestimation of the variances is a consequence of using the residual covariance matrix of the regularized regression model in place of the unknown conditional covariance matrix of the imputation error (cf. section 3a). The residual covariance matrix of the regularized regression model underestimates the conditional covariance matrix of the imputation error for the same reason that the estimate of the imputation error in the appendix underestimates the actual imputation error: the error estimates neglect the uncertainties about the regularization method and the regularization parameter. To be sure, the traces of the estimated covariance matrices, on the average, have a relative error of only about 1.8%, but for datasets in which a greater fraction

of the values is missing, the underestimation of the variances will be greater.[8]

The test of the regularized EM algorithm demonstrates that the algorithm is applicable to typical sets of incomplete climate data and that it leads to more accurate estimates of the missing values than the technique of Smith et al., even though for the technique of Smith et al., the conditions of the test were more favorable than they would be in practice. The limitations of the technique of Smith et al. make it difficult to compare that technique with the regularized EM algorithm in a more complex test problem with a greater fraction of missing values. However, that the regularized EM algorithm already in this relatively simple test led to more accurate imputed values than the technique of Smith et al. suggests that in more complex tests the regularized EM algorithm would also perform better than conventional noniterative imputation techniques that resemble the one of Smith et al.

7. Summary and discussion

Two characteristics complicate the multivariate analysis of typical sets of climate data: most sets of climate data are incomplete, and they contain more variables than records from which the statistics of the data can be estimated. If an incomplete dataset has more records than variables so that it is of full rank, the EM algorithm can be used both to compute the maximum likelihood estimates of the statistics of the data and to fill in missing values with their conditional expectation values given the available values and the estimated statistics. But if an incomplete dataset, like most sets of climate data, has more variables than records so that it is rank-deficient, the conditional expectation values of the missing values are underdetermined and the EM algorithm cannot be used. The EM algorithm for Gaussian data, which is based on iterated linear regression analyses, was taken as the point of departure for the development of a regularized EM algorithm, in which ridge regression with generalized cross-validation replaces the ill-posed conditional maximum likelihood estimation of the regression parameters in the conventional EM algorithm. With the regularized EM algorithm, the mean and the covariance matrix can be estimated from incomplete datasets with more variables than records and missing values in such datasets can be filled in with imputed values.

Since replacing the conditional maximum likelihood estimation of the regression parameters with the estimation of regularized regression parameters by ridge regression is a mere heuristic, the regularized EM algorithm is no longer justified on grounds of general principles such as the maximum likelihood principle. The proofs that the conventional EM algorithm leads to consistent and unbiased estimators and converges monotonically (Little and Rubin 1987, chapter 7) are not transferable to the regularized EM algorithm. Nevertheless, in a test with sets of simulated surface temperature data, the regularized EM algorithm did converge reliably. The values imputed with the regularized EM algorithm were more accurate than the values imputed with a conventional noniterative imputation technique that exploits statistics estimated from a complete training dataset to fill in missing values in an incomplete dataset. But the test with the simulated data also showed that the variances estimated with the regularized EM algorithm are too small, a consequence of an underestimation of the imputation error that the ridge regression entails. Covariance matrices estimated with the regularized EM algorithm and statistics derived from them must therefore be interpreted cautiously, particularly when the fraction of missing values in an incomplete dataset is large.

The regularized EM algorithm differs from conventional imputation techniques for climate data in two respects. It rests on a construal of the problem of estimating statistics from incomplete datasets and imputing missing values as nonlinear, and ill-posed problems arising in the imputation of missing values are regularized adaptively. In conventional imputation techniques for climate data, the nonlinear problem of estimating statistics from incomplete data and imputing missing values is linearized by estimating the statistics of the data from a training dataset, a complete subset of the available data, and exploiting the estimated statistics for the imputation of missing values in the subset of the data in which there are values missing. In the regularized EM algorithm, the statistics are estimated from all available data, including the available values in records with missing values, which makes it unnecessary to distinguish between training datasets and incomplete datasets in which values are to be imputed. Moreover, in conventional imputation techniques, usually a single regularization parameter is chosen for the regularization of the ill-posed problems arising in the imputation of missing values in different records of an incomplete dataset. In the regularized EM algorithm, the ill-posed problems are regularized adaptively, with regularization parameters chosen by generalized cross-validation such as to adapt to the density and pattern of available values in the different records of an incomplete dataset.

The regularized EM algorithm can not only be applied to incomplete datasets containing values of a single geophysical field, such as values of the surface temperature at various grid points or at various stations, but it can also be used to construct historic surface temperature datasets from proxy data. The regularized EM algorithm estimates the correlations between the variables of a dataset and exploits the estimated correlations to fill in

---

[8] The underestimation of the variances can be reduced by multiplying the residual covariance matrices $\hat{\mathbf{C}}_h$ with a scalar inflation factor $\alpha > 1$ before adding them to the cross products (9) of the imputed values. The inflation factor α might be determined from simulation results like the ones above.

missing values with imputed values. Some of the variables in a dataset might represent surface temperature values at the nodes of a spatial grid and other variables might represent proxies of the surface temperature (cf. Mann et al. 1998). The regularized EM algorithm then would estimate the correlations between the temperature proxies and the temperature values at the various grid points and exploit these correlations to impute missing temperature values from available proxy data. Time series of spatial averages of the surface temperature field, regional or global averages, for example, could be computed as one-dimensional projections of the completed set of spatial temperature data. If the relative observational errors of the available temperature measurements and of the proxy variables are of different magnitudes, modifying the regularization method of the regularized EM algorithm so as to take the variance structure of the observational errors explicitly into account might lead to more accurate imputed temperature values than the regularization method presented above (cf. section 3a and Golub et al. 2000).

An extensive literature exists both on the EM algorithm and on ridge regression as a regularized regression method and on generalized cross-validation as a method of determining a regularization parameter. Ridge regression and generalized cross-validation as they are commonly presented had only to be adapted to make them fit into the framework of the EM algorithm. Given the ubiquity of ill-posed problems in the atmospheric and oceanic sciences, be it in the initialization of weather forecasts, in the detection of climate change, or in any of the numerous other problems that involve the solution of ill-conditioned or singular linear systems, it seems that regularization methods other than the widely used truncated principal component analysis and criteria such as generalized cross-validation for determining a regularization parameter deserve more attention than they currently receive. To facilitate experiments with the regularized EM algorithm and with the regularization methods it is composed of, the program code with which the test problems of this paper were computed has been made available online (Schneider 1999).

APPENDIX

Estimation of the Imputation Error

The derivation of the GCV function by Golub et al. (1979) suggests a heuristic for estimating the standard error of the imputed values. Golub et al. showed that the GCV function (20) is a cross-validatory estimate of the mean-squared error of predictions with an estimated ridge regression model, provided that the cross-validation is carried out after transforming the data to a basis in which the individual records of the dataset are strongly coupled (see also Wahba 1990, chapter 4). That the GCV function $\mathcal{G}(h)$ is an estimate of the predictive mean-squared error suggests taking, in the notation of section 3,

$$(\Delta \hat{\mathbf{x}}_m)_j^2 = \frac{\tilde{n}^2}{\mathcal{T}^2(h_j)} (\hat{\mathbf{C}}_h)_{jj} \qquad \text{(A1)}$$

as a rough estimate of the squared standard error $(\Delta \hat{\mathbf{x}}_m)_j^2$ of the imputed value $(\hat{\mathbf{x}}_m)_j$.

However, using the GCV function value as an estimate of the squared imputation error is a heuristic without theoretical foundation. The theoretical foundation of generalized cross-validation is furnished by the fact that the regularization parameter that minimizes the GCV function is approximately equal to the regularization parameter that minimizes the expected mean-squared imputation error (cf. Wahba 1977; Golub et al. 1979). That does not, however, imply that the GCV function value itself is approximately equal to the expected mean-squared imputation error. The uncertainties about the adequacy of the regression model (1), of the regularization method, and of the regularization parameter all contribute to the imputation error, but the error estimate (A1) does not account for these uncertainties. Therefore, the estimated imputation error is merely a lower bound on the actual imputation error.

Nevertheless, the error estimate (A1) has a structure that makes it plausible as a heuristic lower bound on the imputation error. One of the factors $\tilde{n}/\mathcal{T}(h_j)$ that is multiplied with the residual variance $(\hat{\mathbf{C}}_h)_{jj}$ can be interpreted as a bias-correction factor that corrects the assumed number of degrees of freedom in the estimate of the residual variance from $\tilde{n}$ to $\mathcal{T}(h_j)$, thus taking into account the loss of degrees of freedom resulting from the estimation of the ridge regression coefficients. The other factor $\tilde{n}/\mathcal{T}(h_j)$ that is multiplied with the residual variance $(\hat{\mathbf{C}}_h)_{jj}$ can be interpreted as a variance inflation factor that accounts for the sampling error of the ridge regression coefficients and thus of the imputed values.

REFERENCES

Anderssen, R. S., and P. M. Prenter, 1981: A formal comparison of methods proposed for the numerical solution of first kind integral equations. *J. Austral. Math. Soc. B,* **22,** 488–500.

Beale, E. M. L., and R. J. A. Little, 1975: Missing values in multivariate analysis. *J. Roy. Stat. Soc. B,* **37,** 129–146.

Buck, S. F., 1960: A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. Roy. Stat. Soc. B,* **22,** 302–306.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Stat. Soc. B,* **39,** 1–38.

Dixon, K. W., and J. R. Lanzante, 1999: Global mean surface air temperature and North Atlantic overturning in a suite of coupled

GCM climate change experiments. *Geophys. Res. Lett.,* **26,** 1885–1888.

Fierro, R. D., G. H. Golub, P. C. Hansen, and D. P. O'Leary, 1997: Regularization by truncated total least squares. *SIAM J. Sci. Comput.,* **18,** 1223–1241.

Golub, G. H., and C. F. van Loan, 1993: *Matrix Computations.* 2d ed. Johns Hopkins University Press, 642 pp.

——, M. T. Heath, and G. Wahba, 1979: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21,** 215–223.

——, P. C. Hansen, and D. P. O'Leary, 2000: Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.,* **21,** 185–194.

Hansen, P. C., 1994: Regularization tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Numer. Algorithms,* **6,** 1–35.

——, 1997. *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion.* SIAM Monogr. on Mathematical Modeling and Computation, Society for Industrial and Applied Mathematics, 247 pp.

Hoerl, A. E., and R. W. Kennard, 1970a: Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics,* **12,** 55–67.

——, and ——, 1970b: Ridge regression: Applications to non-orthogonal problems. *Technometrics,* **12,** 69–82. Correction, **12,** 723.

Horn, R. A., and C. R. Johnson, 1985: *Matrix Analysis.* Cambridge University Press, 561 pp.

Jones, P. D., 1994: Hemispheric surface air temperature variations: A reanalysis and an update up to 1993. *J. Climate,* **7,** 1794–1802.

Kaplan, A., Y. Kushnir, M. A. Cane, and M. B. Blumenthal, 1997: Reduced space optimal analysis for historical data sets: 136 years of Atlantic sea surface temperatures. *J. Geophys. Res.,* **102C,** 27 835–27 860.

——, M. A. Cane, Y. Kushnir, A. C. Clement, M. B. Blumenthal, and B. Rajagopalan, 1998: Analyses of global sea surface temperature 1856–1991. *J. Geophys. Res.,* **103C,** 18 567–18 589.

Linz, P., 1984: Uncertainty in the solution of linear operator equations. *BIT,* **24,** 92–101.

Little, R. J. A., and D. B. Rubin, 1987: *Statistical Analysis with Missing Data.* Series in Probability and Mathematical Statistics, Wiley, 278 pp.

Manabe, S., and R. J. Stouffer, 1994: Multiple-century response of a coupled ocean–atmosphere model to an increase of atmospheric carbon dioxide. *J. Climate,* **7,** 5–23.

——, ——, M. J. Spelman, and K. Bryan, 1991: Transient response of a coupled ocean-atmosphere model to gradual changes of

atmospheric CO_2. Part I: Annual mean response. *J. Climate,* **4,** 785–818.

Mann, M. E., R. S. Bradley, and M. K. Hughes, 1998: Global-scale temperature patterns and climate forcing over the past centuries. *Nature,* **392,** 779–787.

——, ——, and ——, 1999: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. *Geophys. Res. Lett.,* **26,** 759–762.

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: *Multivariate Analysis.* Series in Probability and Mathematical Statistics, Academic Press, 518 pp.

Papoulis, A., 1991: *Probability, Random Variables, and Stochastic Processes.* 3d ed. McGraw Hill, 666 pp.

Parker, D. E., P. D. Jones, C. K. Folland, and A. Bevan, 1994: Interdecadal changes of surface temperature since the late nineteenth century. *J. Geophys. Res.,* **99D,** 14 373–14 399.

——, C. K. Folland, and M. Jackson, 1995: Marine surface temperature: Observed variations and data requirements. *Climatic Change,* **31,** 559–600.

Rubin, D. B., 1976: Inference and missing data. *Biometrika,* **63,** 581–592.

Schneider, T., cited 1999: Analysis of incomplete climate data: Matlab code. [Available online at http://www.math.nyu.edu/~tapio/imputation/.]

——, and S. M. Griffies, 1999: A conceptual framework for predictability studies. *J. Climate,* **12,** 3133–3155.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate,* **9,** 1403–1420.

Tarantola, A., 1987: *Inverse Problem Theory: Methods for Data Fitting and Model Parameter Estimation.* Elsevier, 613 pp.

Tikhonov, A. N., and V. Y. Arsenin, 1977: *Solution of Ill-Posed Problems.* Scripta Series in Mathematics, V. H. Winston and Sons, 258 pp.

van Huffel, S., and J. Vandewalle, 1991: *The Total Least Squares Problem: Computational Aspects and Analysis.* Frontiers in Applied Mathematics, Vol. 9, Society for Industrial and Applied Mathematics, 300 pp.

Wahba, G., 1977: Practical approximate solutions to linear operator equations when the data are noisy. *SIAM J. Numer. Anal.,* **14,** 651–667.

——, 1990: *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Society for Industrial and Applied Mathematics, 169 pp.

——, and Y. Wang, 1995: Behavior near zero of the distribution of GCV smoothing parameter estimates. *Stat. Probabil. Lett.,* **25,** 105–111.

Webster, J. T., and R. F. Gunst, 1974: Latent root regression analysis. *Technometrics,* **16,** 513–522.