Nonlin. Processes Geophys., 14, 1–2, 2007
www.nonlin-processes-geophys.net/14/1/2007/

Nonlinear Processes in Geophysics

Comment on "Spatio-temporal filling of missing points in geophysical data sets" by D. Kondrashov and M. Ghil, Nonlin. Processes Geophys., 13, 151–159, 2006

T. Schneider

California Institute of Technology, Pasadena, CA, USA

Received: 10 July 2006 – Revised: 1 October 2006 – Accepted: 22 December 2006 – Published: 15 January 2007

Kondrashov and Ghil (2006) (KG hereafter) describe a method for imputing missing values in incomplete datasets that can exploit both spatial and temporal covariability to estimate missing values from available values. Temporal covariability has not been exploited as widely as spatial covariability in imputing missing values in geophysical datasets, but, as KG show, doing so can improve estimates of missing values. However, there are several inaccuracies in KG's paper. Since similar inaccuracies have surfaced in other recent papers, for example, in the literature on paleo-climate reconstructions, I would like to point them out here.

(i) In estimating covariance matrices, KG treat an incomplete dataset with imputed values filled in as if it were a complete dataset. Possible variations of the missing values around the imputed values are ignored, leading to biased estimates of covariance matrices (Little and Rubin, 2002). The expectation-maximization (EM) algorithm takes variations of the missing values around the imputed values into account, which is essential to obtain maximum likelihood estimates of parameters such as covariance matrices with their attendant optimality properties (Dempster et al., 1977). Regularized variants of the EM algorithm (Schneider, 2001) likewise take variations of the missing values around the imputed values into account in the estimation of variances and covariances, such that they reduce, in the limit of no regularization, to the EM algorithm for Gaussian data. The same could be done in KG's method by adding estimated covariance matrices of the imputation error to the sample covariance matrix of the completed dataset. This would improve the accuracy of KG's method, for example, in estimating variability. Accurate variability estimates are particularly important for estimating higher-order statistics, such as extreme value statistics, which can be strongly biased if variations of missing values around imputed values coming from the center of a distribution of possible values are not taken into account.

(ii) Except for neglecting the variations of the missing values around the imputed values and for an unusual order of iterations – iteratively re-estimating individual principal components – KG's method is similar to the regularized EM algorithm exploiting spatial and stationary temporal covariability described in Schneider (2001). (A principal component technique similar to that of KG and Beckers and Rixen (2003) but with the more usual order of iterations – iteratively re-estimating covariance matrices and all relevant principal components – was presented by Everson and Sirovich (1995).) KG's principal component technique for imputing missing values corresponds to an orthogonal or truncated total least squares (auto-)regression (Fierro et al., 1997) and can be used in a regularized EM algorithm as discussed in Schneider (2001). As KG's method, a regularized EM algorithm with truncated total least squares regression uses leading principal components based on the entire dataset, including all records and variables with missing and available values. An innovation in KG's method is to make the time lag up to which temporal covariability is exploited an adaptive parameter.

(iii) As a result of the similarity of KG's method and a regularized EM algorithm exploiting spatio-temporal covariability with truncated total least squares regressions, several of KG's claims of how their method differs from regularized EM algorithms are incorrect. For example, KG's contrasting of their method as being "non-parametric" as opposed to the "parametric" regularized EM algorithm is incorrect. The EM algorithm for Gaussian data yields maximum likelihood estimates of mean values, covariance matrices, and missing values, with their attendant optimality properties, but it and its regularized variants can also be justified under weaker assumptions (as least squares methods or regularized variants). KG's method is just as parametric as the regularized EM algorithm.

Correspondence to: T. Schneider
(tapio@caltech.edu)

Published by Copernicus GmbH on behalf of the European Geosciences Union and the American Geophysical Union.

(iv) There are other inaccuracies, particularly where KG contrast their method with other methods. For example, it is not correct that an “EM-based method . . . [relies] on the randomness in time of the missing values.” The EM algorithm and regularized variants rely on the assumption that missing values are missing at random, which does not mean that values are missing randomly in time or in space but that the probability that a value is missing is independent of the missing value – the central necessary condition for the mechanisms responsible for missingness to be ignorable (Little and Rubin, 2002). KG’s method relies on the same assumption.

While using different terms and concepts may create the impression that methods used to estimate statistics from incomplete data differ more strongly than they do, actual methodological differences, even if small, may be important in determining the performance of the methods. A systematic exploration of the advantages and disadvantages of different methods is desirable, including methods such as that of KG that exploit spatio-temporal covariability.

Edited by: B. D. Malamud
Reviewed by: B. D. Malamud

References

Beckers, J. M. and Rixen, M.: EOF calculations and data filling from incomplete oceanographic datasets, J. Atmos. Oceanic Technol., 20, 1839–1856, 2003.

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

Everson, R. and Sirovich, L.: Karhunen-Loève procedure for gappy data, J. Opt. Soc. Am. A, 12, 1657–1664, 1995.

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

Kondrashov, D. and Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets, Nonlin. Processes Geophys., 13, 151–159, 2006, http://www.nonlin-processes-geophys.net/13/151/2006/.

Little, R. J. A. and Rubin, D. B.: Statistical Analysis with Missing Data, Series in Probability and Mathematical Statistics, Wiley, New York, 2nd edn., 2002.

Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, J. Climate, 14, 853–871, 2001.