Proceeding of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences
Florianopolis-SC, Brazil, July 10-13, 2012.

Some expectation-maximization (EM) algorithm simplifications for spatial data

Daniel A. Griffith[1]

[1] University of Texas at Dallas, School of Economic, Political, and Policy Sciences, 800 W. Campbell Rd., GR31, Richardson, TX 75080-3021, USA
dagriffith@utdallas.edu

Abstract

The EM algorithm is a generic tool that offers maximum likelihood solutions when data sets are incomplete with data values missing at random or completely at random. At least for its simplest form, the algorithm can be rewritten in terms of an ANCOVA regression specification. This formulation allows several analytical results to be derived that permit the EM algorithm solution to be expressed in terms of new observation predictions and their variances. Implementations can be made with a linear regression, with a nonlinear regression, and with a generalized linear model routine, allowing missing value imputations, even when they must satisfy constraints or involve dependent observations. This paper extends to spatially correlated data findings already reported for non-spatial data, linking the EM algorithm solution with spatial autoregression, geostatistical kriging, and eigenvector spatial filtering. One theorem is proved, and two corollaries are derived that broadly contextualize imputation findings in terms of the theory, methodology, and practice of spatial statistical science.

Keywords: ANCOVA, eigenvector spatial filter, EM algorithm, kriging, spatial autoregression.

1. Introduction

The *Ex*pectation-*M*aximization (EM) algorithm (Dempster, Laird, and Rubin 1977), an iterative procedure for computing maximum likelihood estimates (MLEs) when data sets are incomplete, with data values missing at random (MAR) or completely at random (CMAR), is a useful device for helping to solve a wide range of model-based estimation problems. Flury and Zoppè (2000, p. 209) emphasize that

> it can not be stressed enough that the E-step does not simply involve replacing missing data by their conditional expectations (although this is true for many important applications of the algorithm).

But frequently model-based estimation problems desire just this type of imputation output from the algorithm. Furthermore, in certain situations, focusing on imputation dramatically simplifies the EM solution.

Descriptions of the EM algorithm may be found in Flury and Zoppè (2000), Meng (1997), and McLachlan and Krishnan (1997), among others. The objective of this paper is to present spatial regression solutions that render conditional expectations for missing values in a data set that are equivalent to EM algorithm results.

Because the EM procedure requires imputation of the complete-data sufficient statistics, rather than just the individual missing observations, the equivalency discussed here initially derives from an assumption of normality, for which the means and covariances constitute the sufficient statistics. An assumption of normality links ordinary least squares (OLS) and MLE regression results, too; application of the Rao-Blackwell factorization theorem verifies that the means and covariances are sufficient statistics in this situation.

1.1. Classical background

Yates (1933) shows for analysis of variance (ANOVA) that if each missing observation is replaced by a parameter to be estimated (i.e., the conditional expectation for a missing value), the resulting modified analysis becomes straightforward by treating the estimated missing value as a parameter (i.e., an imputation). Rewriting the ANOVA as a standard regression problem involves introducing a binary indicator variable for each missing value—the value of 1 denoting the missing value observation in question, and 0 otherwise—with the estimated regression coefficients for these indicator variables being the negative of the missing value estimates. This approach is equivalent to subtracting each observation's missing value, in turn, from each side of a regression equation. Generalizing this regression formulation to include covariates allows missing values to be estimated with an analysis of covariance (ANCOVA) regression specification, an approach suggested by Bartlett (1937) and by Rubin (1972). Replacing the arbitrarily assigned value of 1 in each individual observation missing value indicator variable by the value –1 yields estimated regression parameters with the correct sign.

Consider a bivariate set of n observed values, each pair denoted by (y_i , x_i), i=1, 2, …, n. Suppose only the response variable, Y, contains incomplete data. First, the n_m missing values need to be replaced by 0. Second, n_m 0/–1 indicator variables, $-I_m$ (m = 1, 2, …, n_m), need to be constructed; I_m contains n–1 0s and a single 1 corresponding to the m^{th} missing value observation. The minus sign for $-I_m$ indicates that a –1 actually is entered into each of the m indicator variables. Regressing Y on a complete data predictor variable, X—which furnishes the redundant information that is exploited to compute imputations—together with the set of m indictor variables constitutes the ANCOVA.

Suppose $\mathbf{Y}_o$ denotes the n_o-by-1 ($n_o = n - n_m$) vector of observed response values, and $\mathbf{Y}_m$ denotes the n_m-by-1 vector of missing response values. Let $\mathbf{X}_o$ denote the vector of predictor values for the set of observed response values, and $\mathbf{X}_m$ denote the vector of predictor values for the set of missing response values. Further, let $\mathbf{1}$ denote an n-by-1 vector of ones that can be partitioned into $\mathbf{1}_o$, denoting the vector of ones for the set of observed response values, and $\mathbf{1}_m$, denoting the vector of ones for the set of missing response values. Then the ANCOVA specification of the regression model may be written in partitioned matrix form as

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \begin{pmatrix} \mathbf{1}_o & \mathbf{X}_o \\ \mathbf{1}_m & \mathbf{X}_m \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \mathbf{0}_{om} \\ -\mathbf{I}_{mm} \end{pmatrix} (\boldsymbol{\beta}_m) + \begin{pmatrix} \boldsymbol{\varepsilon}_o \\ \mathbf{0}_m \end{pmatrix}, \quad (1)$$

where $\mathbf{0}_j$ (j = o, m) is an n_j-by-1 vector of zeroes, $\mathbf{0}_{om}$ is an n_o-by-n_m matrix of zeroes, α and β respectively are the bivariate intercept and slope regression parameters, $\boldsymbol{\beta}_m$ is an n_m-by-1 vector of conditional expectation regression parameters, $\mathbf{I}_{mm}$ is an n_m-by-n_m identity matrix, and $\boldsymbol{\varepsilon}_o$ is an n_o-by-1 vector of random error terms. The bivariate OLS regression coefficient estimates, a and b, of α and β, respectively, for this ANCOVA specification are given by

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} n - n_m & \mathbf{1}_o^T \mathbf{X}_o \\ \mathbf{X}_o^T \mathbf{1}_o & \mathbf{X}_o^T \mathbf{X}_o \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}_o^T \mathbf{Y}_o \\ \mathbf{X}_o^T \mathbf{Y}_o \end{pmatrix}, \quad (2)$$

where T denotes matrix transpose, which is the regression results for the observed data only. In addition, the regression coefficients, $\mathbf{b}_m$, for the indicator variables are given by

$$\mathbf{b}_m = a\mathbf{1}_m + b\mathbf{X}_m = \hat{\mathbf{Y}}_m \ , \quad (3)$$

which is the vector of point estimates for additional observations (i.e., the prediction of new observations) that should have X values within the interval defined by the extreme values contained in the vector $\mathbf{X}_o$. This is a standard OLS regression result, as is the prediction error that can be attached to it (see, for example, Montgomery and Peck, 1982, pp. 31-33). In addition, the values here are positive because the $-I_m$ indicator variables contain negative ones.

1.2. Spatial statistics background

Haining, Griffith and Bennett (e.g., 1989) outline a spatial EM algorithm for estimating missing spatial data values. Consider the following n-by-n partitioned spatial covariance matrix, which captures spatial autocorrelation (see, e.g., Cliff and Ord, 1980) effects:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{om} \\ \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{mm} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_{oo} & \mathbf{V}_{om} \\ \mathbf{V}_{mo} & \mathbf{V}_{mm} \end{pmatrix}^{-1} \sigma^2 \ , \quad (4)$$

where, as before, the subscript o denotes observed data, and the subscript m denotes missing data. For a multivariate normal probability model, the MLE for missing data is given by

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m\boldsymbol{\beta} + \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}(\mathbf{Y}_o - \mathbf{X}_o\boldsymbol{\beta}) \ , \quad (5)$$

which is the kriging equation of geostatistics (see Christensen, 1991, p. 268; Griffith, 1993). Using the preceding notation, Haining, Griffith and Bennett show that for an autoregressive model specification, equation (5) becomes

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m\boldsymbol{\beta} - \mathbf{V}_{mm}^{-1}\mathbf{V}_{mo}(\mathbf{Y}_o - \mathbf{X}_o\boldsymbol{\beta}) \ , \quad (6)$$

which reduces to the following equation for the conditional autoregression (CAR) model specification based upon a binary 0/1 geographic connectivity matrix, $\mathbf{C}$ ─where $c_{ij} = 1$ if locational units i and j are neighbors, and $c_{ij} = 0$ otherwise─and spatial autocorrelation parameter ρ :

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m\boldsymbol{\beta} + \rho(\mathbf{I} - \rho\mathbf{C}_{mm})^{-1}\mathbf{C}_{mo}(\mathbf{Y}_o - \mathbf{X}_o\boldsymbol{\beta}) \ .$$

In other words, the spatial EM and geostatistical kriging solutions are identical, and are equivalent to predicting new observations, yielding the following theorem:

> THEOREM 1. The MLE for missing georeferenced values described by a spatial autoregressive model specification, given by equation (1), is equivalent to the best linear unbiased predictor kriging equation of geostatistics, given by equation (5).
>
> PF: Substituting the partitioned matrix components of matrix $\mathbf{V}^{-1}$ in equation (4) into their partitioned matrix $\boldsymbol{\Sigma}$ counterparts appearing in equation (5) yields equation (6).

This theorem highlights that spatial autoregression specifications deal with an inverse covariance matrix, whereas semivariogram specifications deal directly with the corresponding covariance matrix itself.

2. Spatial autoregressive model specification imputations

Theorems from Griffith (2010) coupled with the simultaneous autoregressive (SAR) model specification based upon the row-standardized version of matrix **C**, namely matrix **W** (i.e., $w_{ij} = c_{ij} \Big/ \sum_{j=1}^{n} c_{ij}$), yields the following spatial SAR EM algorithm solution:

COROLLARY 1. Employing an autoregressive model specification to account for spatial autocorrelation in georeferenced data renders the imputation equation

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \rho \mathbf{W} \begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} + (\mathbf{I} - \rho \mathbf{W}) \mathbf{X} \boldsymbol{\beta} + \sum_{m=1}^{M} y_m (-\mathbf{I}_m + \rho \mathbf{W}^*_{om}) + \boldsymbol{\varepsilon}, \quad (7)$$

where missing values in the vector **Y** are replaced by 0s, $\mathbf{I}_m$ is the indicator variable vector for missing value m that contains n–1 0s and a single 1 in each of its m columns (hence, the vector $-\mathbf{I}_m$ has 0s and –1s), $\mathbf{W}^*_{om}$ is the column of the geographic weights matrix **W** associated with the m^{th} missing value, and M is the number of missing values.

The particular solution equations are given by

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m \hat{\boldsymbol{\beta}}_o - \hat{\mathbf{V}}_{mm}^{-1} \hat{\mathbf{V}}_{mo} (\mathbf{Y}_o - \mathbf{X}_o \hat{\boldsymbol{\beta}}_o),$$

where $\mathbf{V} = (\mathbf{I} - \rho \mathbf{W})^T (\mathbf{I} - \rho \mathbf{W})$, and

$$\mathbf{s}^2_{y_m} = \hat{\sigma} \{ \hat{\mathbf{V}}_{mm}^{-1} + \hat{\mathbf{V}}_{mm}^{-1} [\mathbf{B} - \mathbf{A} \hat{\mathbf{V}}_{mm}^{-1} \mathbf{A}]^{-1} \hat{\mathbf{V}}_{mm}^{-1} \}_{diag},$$

where $\mathbf{A} = -(\hat{\mathbf{V}}_{mo} \mathbf{X}_o + \hat{\mathbf{V}}_{mm} \mathbf{X}_m)$ and $\mathbf{B} = \mathbf{X}_n^T \hat{\mathbf{V}} \mathbf{X}_n$. These are the predicted values and the predicted variances for new georeferenced observations.

3. Eigenvector spatial filter model specification imputations

Theorems from Griffith (2010) coupled with an eigenvector spatial filter (ESF) model specification (see Griffith, 2000, 2002, 2004) yields the following spatial filter EM algorithm solution:

COROLLARY 2. Employing a spatial filter constructed from eigenvectors selected to account for spatial autocorrelation in georeferenced data renders the imputation equation

$$\begin{pmatrix} \mathbf{Y}_o \\ \mathbf{0}_m \end{pmatrix} = \alpha \mathbf{1} + \mathbf{X} \boldsymbol{\beta}_X - \sum_{m=1}^{M} y_m \mathbf{I}_m + \sum_{k=1}^{K} \mathbf{E}_k \beta_{\mathbf{E}_k} + \boldsymbol{\varepsilon}, \quad (8)$$

where $\boldsymbol{\beta}_X$ is the vector of regression coefficients for the set of X attribute variable covariates, K eigenvectors, denoted by $\mathbf{E}_k$, are selected from the candidate set extracted from matrix

$$(\mathbf{I} - \mathbf{1}\mathbf{1}^T/n) \mathbf{C} (\mathbf{I} - \mathbf{1}\mathbf{1}^T/n),$$

an expression that appears in the numerator of the Moran Coefficient, and β_{E_k} is the regression coefficient for the k^{th} selected eigenvector.

The particular solution equations are given by

$$\hat{\mathbf{Y}}_m = \mathbf{X}_m \mathbf{b}_X + \mathbf{E}_{m,K} \mathbf{b}_{E_K}, \text{ and}$$

$$\mathbf{s}^2_{y_m} = \hat{\sigma}\{\mathbf{I}_m + \left(\mathbf{X}_m \quad \mathbf{E}_{m,K}\right)\left[\left(\mathbf{X}_m \quad \mathbf{E}_{m,K}\right)^T \left(\mathbf{X}_m \quad \mathbf{E}_{m,K}\right)\right]^{-1} \left(\mathbf{X}_m \quad \mathbf{E}_{m,K}\right)^T\}_{diag}.$$

These are the predicted values and the predicted variances for new georeferenced observations. The spatial filter can be constructed using stepwise regression procedures.

4. An empirical example

Consider the percentage, p, of farm land harvested in Puerto Rico during 2002 (see http://www.nass.usda.gov/Statistics_by_State/Puerto_Rico/index.asp). Four municipalities have missing values, whereas the number of farms, F, contained in each municipality is not considered confidential and is reported. The log-odds ratio displays weak positive spatial autocorrelation [its Moran Coefficient is 0.21 (z = 39.6)]. Estimation of the parameters of equation (7) requires a Jacobian modification (see Martin 1984), which reduces to a denominator of n_o in this particular case (i.e., the missing values are dispersed, resulting in the standard Jacobian term being the sum of the log-eigenvalues divided by n_o rather than n). The spatial autoregressive parameter estimate is $\hat{\rho}$ = 0.3426, and the imputations for three of these missing data municipalities (Vieques, an island located southeast of the main island, is not included) are

obser-	log-normal approximation		Binomial regression		ESF Multiple
vation	SAR	ESF	conventional	ESF	imputations
29	0.26	0.23	0.23	0.23	0.24 (se = 0.099)
49	0.29	0.31	0.24	0.27	0.32 (se = 0.116)
70	0.13	0.14	0.15	0.15	0.15 (se = 0.076)

These results highlight that the SAR and ESF specifications yield very similar results, which also are very similar to those obtained with a conventional and an ESF binomial random variable EM algorithm. The SAS (1999) multiple imputation procedure produces standard errors for these imputations.

5. Conclusion

Griffith (2010) presents a simplification of the EM algorithm for the linear model and non-spatial data. This paper extends that more classical work with a theorem and two corollaries to the situation of spatially autocorrelated, georeferenced data. In parallel with the non-spatial solution, the spatial autoregressive EM algorithm is found to be equivalent to kriging in geostatistics. Although autoregressive modeling furnishes a spatial EM algorithm solution for linear modeling, spatial

filtering provides the necessary tools to extend this analysis to generalized linear modeling (e.g., binomial and Poisson regression).

In conclusion, this paper presents methodology that allows a spatial scientist to handle a source of uncertainty in georeferenced datasets, namely missing data. It also outlines how to obtain standard errors for imputed spatial data values. Just as a kriging map should be accompanied by a prediction error map, imputed values should be accompanied by standard errors.

Acknowledgments

Daniel A. Griffith is an Ashbel Smith Professor of Geospatial Information Sciences at the University of Texas at Dallas.

References

Bartlett, M. 1937. Some examples of statistical methods of research in agriculture and applied biology, *Journal of the Royal Statistical Society*, Supplement 4: 137-183.

Christensen, R. 1991. *Linear Models for Multivariate, Time Series, and Spatial Data*. Berlin: Springer-Verlag.

Cliff, A., and J. Ord. 1980. *Spatial Processes*. London: Pion.

Dempster, A., Laird, N., and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*. Series B, 39: 1-38.

Flury, B., and A. Zoppè. 2000. Exercises in EM, *The American Statistician*. 54: 207-209.

Griffith, D. 1993. Advanced spatial statistics for analyzing and visualizing geo-referenced data, *International Journal of Geographical Information Systems*. 7: 107-123.

Griffith, D. 2000. A linear regression solution to the spatial autocorrelation problem, *J. of Geographical Systems*, 2: 141-156.

Griffith, D. 2002. A spatial filtering specification for the auto-Poisson model, *Statistics & Probability Letters*. 58: 245-251.

Griffith, D. 2004. A spatial filtering specification for the auto-logistic model, *Environment & Planning A*. 36: 1791-1811.

Griffith, D. 2010 Some simplifications for the Expectation-Maximization (EM) algorithm: the linear regression model case, *InterStat*, March article 2 (http://interstat.statjournals.net/YEAR/2010/abstracts/1003002.php, http://interstat.statjournals.net/YEAR/2010/articles/1003002.pdf), 23 pp.

Haining, R., D. Griffith, and R. Bennett. 1989. Maximum likelihood estimation with missing spatial data and with an application to remotely sensed data, *Communications in Statistics*. 18: 1875-1894.

Martin, R. 1984. Exact maximum likelihood for incomplete data from a correlated Gaussian process, *Communications in Statistics*. 13: 1275-1288.

McLachlan, G., and T. Krishnan. 1997. *The EM-algorithm and Extensions*. New York: Wiley.

Meng, X. 1997. The EM algorithm, in *Encyclopedia of Statistical Sciences*, Update Vol. 1, edited by S. Kotz, C. Read and D. Banks, pp. 218-227, New York: Wiley.

Montgomery, D., and E. Peck. 1982. *Introduction to Linear Regression Analysis*. New York: Wiley.

Rubin, D. 1972. A non-iterative algorithm for least squares estimation of missing values in any analysis of variance with missing data, *Applied Statistics*. 21: 136-141.

SAS. 1999. "Chapter 9. The MI Procedure," *OnlineDoc*™, Version 8. support.sas.com/rnd/app/papers/miv802.pdf (accessed 4/14/2006).

Yates, F. 1933. The analysis of replicated experiments when the field results are incomplete, *Empirical Journal of Experimental Agriculture*. 1: 129-142.