MODEL BASED ESTIMATION OF COVARIANCE MATRICES WITH APPLICATIONS TO THE EM–ALGORITHM

Stephen M. Woodruff, Bureau of Labor Statistics, 441 G St. N.W., Washington D.C. 20212

Key Words: Imputation, Regression Superpopulation Models.

1. INTRODUCTION

When the minimization of mean square error (or variance) is a primary criterion for chosing an estimator of means or totals, then second moment estimates are often necessary too. Some examples of this are composite estimators where the component weights are functions of the component variances, generalized least squares estimators where an estimate of a covariance matrix is required, and the normal EM–algorithm where the sufficient statistics are functions of first and second moments. In summary, estimation of first moments is often intertwined with estimation of second moments. In cases where variances are not required in the first moment estimators, it may be necessary to estimate the variance of these first moment estimators, and this will require second moment estimates.

The variances of second moment estimators are generally functions of population fourth moments and thus these second moment estimators can be very unstable. In some cases data relationships expressed by superpopulation models can impose restrictions on variance/covariance structure and suggest variance/covariance estimators, which themselves have relatively small variance. We consider a case where this information takes the form of a Markov superpopulation model which specifies that the expected current value of an item for a population unit is a function of the realized value of that item in the immediate past. For example, a manufacturer's expected output this year may be roughly proportional to his actual output last year. Such relationships can be expressed in terms of regression superpopulation models, and these models imply restrictions on the covariance matrix of the random variables that describe this longitudinal data. These restrictions can reduce the number of parameters and second moment terms that need to be estimated.

This paper expands on two other papers by Woodruff (1989) and Johnson and Woodruff (1990). Both these papers apply one of the covariance matrix estimators analyzed here to generalized least squares estimation of finite population means and totals in the Bureau of Labor Statistics' (BLS) Current Employment Statistics (CES) survey.

The CES survey is the Bureau's largest employment survey. It measures total national employment each month in about 1500 industry cells. Every month, the Bureau publishes estimates of total cell employment for past reference months based on all the CES survey data that is available for survey reference periods one, two, and three months in the past. Due to delayed reporting, this short time between reference date and initial publication date (one month) means that initial CES employment estimates may be based on relatively few sample units (often only about half) and as time passes and more data arrives, substantial revisions to these initial estimates are sometimes necessary. An estimator developed for the CES survey which depends on the model based covariance matrix estimator studied here can substantially reduce these revisions.

To summarize the data flow, we can say that within one month of the reference date about half the units have responded, within two months this proportion is about three quarters, and within three months it is about nine tenths.

All the sample data that is available for the m most current months for the n sample units in an estimation cell can be summarized in an nxm matrix, M, with missing entries, where an entry is missing if a given sample unit (row of M) did not have data for a given reference month (column of M).

In the next section, a regression superpopulation model for M is described. This model is similar to the superpopulation models considered by Royall and Cumberland (1981a). An improved estimator of employment level in the CES survey, Johnson and Woodruff (1990), requires the variance/covariance matrix, Σ, defined in this model. This paper describes an estimator of Σ based on the regression superpopulation model (REM estimate) and compares it empirically with the estimator of Σ from the Normal EM–algorithm (NEM estimate). Little and Rubin (1987) give a clear and complete description of the Normal EM–algorithm (NEM). For more detail on the EM–algorithm see Beale and Little (1975) or Dempster and Laird (1977). In the simulation study, absolute error of the two estimators is compared.

The REM estimate of Σ uses additional stochastic structure beyond the multivariate normality from which the NEM is derived. Thus, it is not surprising that the REM has smaller absolute error than the NEM for estimating Σ. However, the size of this reduction in absolute error is surprising.

Although this application of a regression superpopulation model to derive the REM covariance matrix estimate may be of marginal interest in itself, this REM estimate is an important component of the GLS estimator (Link90) described in Johnson and Woodruff (1990). In addition, it is computationally far cheaper than the NEM estimate since it does not involve iterative recomputations, possibly several hundred, to convergence.

2. THE NORMAL EM ALGORITHM AND THE MODEL ENHANCED ALTERNATIVE

In this section we describe the Normal EM–algorithm (NEM) and a model enhanced alternative for estimating a covariance matrix when some addition stochastic structure is used (REM). Let the available data be summarized in the matrix, M. Suppose further that M is the result of two processes. Under the first process, an nxm matrix is generated where the rows of this matrix are the outcomes of iid random vectors. Under the second process, certain components are deleted from the matrix that was generated by the first process (nonresponse). We further assume that these two processes are stochastically independent (i.e. nonresponse is independent of the sample measurements).

This missing data mechanism is called ignorable nonresponse and it is an essential assumption underlying the EM–algorithm. Although it does not hold for the CES data, some conditional properties of the CES data suffice to make ignorability a good enough approximation.

Then M can be modelled as the element by element product of two matrices, $M = (Z_1,Z_2,Z_3,\ldots\ldots ,Z_n)'xR$ where the $\{Z_i\}$ are iid m–dimension column vectors, each $Z_i \sim N(\mu,\Sigma)$, and R is an nxm matrix of zeros and ones. The $(i,j)^{th}$ component of R is zero if the j^{th} component of Z_i was deleted and one otherwise. Thus a zero entry in M denotes a missing datum (item nonresponse). The unknowns, μ and Σ, are to be estimated via the EM–algorithm.

If there were no missing data (R is all ones), the sufficient statistics for (μ,Σ) when the $\{Z_i\}$ iid normal are the realizations of:

$T_1 = \sum_{i=1}^{n} Z_i$ and $T_2 = \sum_{i=1}^{n} Z_i Z_i'$ where the $\{Z_i\}_{i=1}^{n}$ are from a simple random sample of size n from a population of N units.

Then $E(T_1) = n\mu$ and $E(T_2) = n(\Sigma + \mu\mu')$. $\hat{\mu} = (1/n)T_1$ and $\hat{\Sigma} = (1/n)T_2 - \hat{\mu}\hat{\mu}'$ are the maximum likelihood estimates of (μ,Σ) for the complete data case.

When data are missing, the EM–algorithm proceeds as follows. For each vector Z_i with missing values let Z_{i1} denote the missing components and Z_{i2} denote the observed components. Then, without loss of generality, $Z_i = (Z_{i1}',Z_{i2}')'$. Note that for arbitrary patterns of missing values this block partition of Z_i does not hold, and the expressions to follow would then be preceded and followed by indicator matrices of zeros and ones. This is a notational complexity that contributes nothing to the understanding of the EM–algorithm, so this notational complexity is omitted. Let $\hat{\mu}(1)$ be the vector of column means computed from the available data (nonzero entries) in each column of M. If the i^{th} column mean derived this way is imputed for each missing entry in the i^{th} column of M then let $\hat{\Sigma}(1)$ be the MLE estimator of Σ from this imputation for M (i.e. $\hat{\Sigma}(1)=(1/n)M_o'M_o - \hat{\mu}(1)\hat{\mu}(1)'$ where M_o denotes M with these column means imputed for the zeros). Given these initial estimates, $\hat{\mu}(1)$ and $\hat{\Sigma}(1)$, impute values for Z_{i1} as $\hat{Z}_{i1} = E(Z_{i1}|Z_{i2},\hat{\mu}(1),\hat{\Sigma}(1)) = \hat{\mu}_1(1) + \hat{\Sigma}_{12}(1)\hat{\Sigma}_{22}^{-1}(1)(Z_{i2} - \mu_2(1))$, where $\hat{\Sigma}(q) = \begin{bmatrix} \hat{\Sigma}_{11}(q) & \hat{\Sigma}_{12}(q) \\ \hat{\Sigma}_{21}(q) & \hat{\Sigma}_{22}(q) \end{bmatrix}$ and $\hat{\mu}(q) = (\hat{\mu}_1'(q),\hat{\mu}_2'(q))'$, for q=1,2,3, etc. Thus $\hat{\Sigma}_{11}(q)$ is an estimated covariance matrix of Z_{i1}, and $\hat{\mu}_1(q)$ is an estimate for the mean of Z_{i1}, etc. $\hat{Z}_{i1}$ estimates the contribution of Z_{i1} to T_1.

The predicted contribution of Z_{i1} to T_2 is $\widehat{Z_{i1}Z_{i1}'} = E(Z_{i1}Z_{i1}'|Z_{i2},\hat{\mu}(1),\hat{\Sigma}(1)) = \Sigma_{11}(1) - \Sigma_{12}(1)\Sigma_{22}^{-1}(1)\Sigma_{21}(1) + \hat{Z}_{i1}\hat{Z}_{i1}'$ and $\widehat{Z_{i1}Z_{i2}'} = E(Z_{i1}Z_{i2}'|Z_{i2},\hat{\mu}(1),\hat{\Sigma}(1)) = \hat{Z}_{i1}Z_{i2}'$. These contributions to the i^{th} term in T_1 and T_2 are then inserted for the missing parts of the i^{th} term for each $1 \le i \le n$. Let $\hat{T}_1$ and $\hat{T}_2$ denote these imputations for T_1 and T_2. The next iterate MLEs for (μ,Σ) are $\hat{\mu}(2) = \hat{T}_1/n$ and $\hat{\Sigma}(2) = \hat{T}_2/n - \hat{\mu}(2)\hat{\mu}'(2)$. With these new values, $(\hat{\mu}(2),\hat{\Sigma}(2))$, we re–impute the above conditional expected values to get new $\hat{T}_1$ and $\hat{T}_2$ and the next iterate MLEs, $(\hat{\mu}(3),\hat{\Sigma}(3))$. This process continues until convergence of the $\hat{\mu}(q)$ and this occurs when

$$\max_{1 \le j \le m} |\hat{\mu}(q)_j - \hat{\mu}(q-1)_j| < \varepsilon \text{ for some } \varepsilon \text{ where } \hat{\mu}(q)_j \text{ is}$$

the j^{th} component of $\hat{\mu}(q)$. This describes the NEM.

If this process converges after q iterations then $\hat{\Sigma}(q)$ is the NEM estimate of Σ.

The estimator for Σ which uses some additional structure is described next and refered to as the REM estimate of Σ. To derive this estimator for Σ, we first describe the additional structure, which models the relationships between the components of a Z_i. Now let Z_{ij} be the j^{th} component of Z_i (as opposed to the above usage where the second subscript is used to denote a partition of Z_i). For each i, let the $\{\varepsilon_{ij}\}_{j=1}^{m}$ be pairwise uncorrelated with expected value zero. Let $\{\beta_j\}_{j=1}^{m}$ be unknown constants. Then suppose for each i, $Z_{i1} = \beta_1 + \varepsilon_{i1}$ and $Z_{ij} = \beta_j Z_{ij-1} + \varepsilon_{ij}$ for j = 2,3,.. m. This is the regression superpopulation model and this model implies that the off–diagonal entries of Σ can be written as:

$$b_{kj} = \sigma_k^2 \prod_{l=k+1}^{j} \beta_l \text{ for } k<j\le m \text{ where the } \{\sigma_k^2\} \text{ are}$$

diagonal entries of this covariance matrix (i.e. $\sigma_k^2 = V(Z_{ik})$, the variance of Z_{ik}, for all $1 \le i \le n$ and $1 \le k \le m$).

Note that this Σ can be written as the element by element product of Σ_1 and Σ_2 where:

$$\Sigma_1 = \begin{bmatrix} 1 & \beta_{22} & \cdots & \beta_{2m} \\ \beta_{22} & 1 & \cdots & \beta_{3m} \\ \vdots & & \ddots & \vdots \\ \beta_{2m} & \beta_{3m} & \cdots & 1 \end{bmatrix} \text{ with } \beta_{ij} = \prod_{k=i}^{j} \beta_k$$

$$\text{and } \Sigma_2 = \begin{bmatrix} \sigma_1^2 & \sigma_1^2 & \cdots & \sigma_1^2 \\ \sigma_1^2 & \sigma_2^2 & \cdots & \sigma_2^2 \\ \vdots & & \ddots & \vdots \\ \sigma_1^2 & \sigma_2^2 & \cdots & \sigma_m^2 \end{bmatrix}$$

If $\beta_i \doteq 1$ for all i then $\Sigma = \Sigma_2$ (i.e. m parameters to estimate).

If $\sigma_i^2 \doteq \sigma_1^2$ for all i then $\Sigma = \sigma_1^2 \Sigma_1$ (& m+1 parameters to estimate).

When neither of these simplifications is appropriate, note that Σ^{-1} is the tri–diagonal matrix with diagonal entries $(1/\tau_j) - (\beta_{j+1}^2/\tau_{j+1})$ for $1 \le j < m$ and for j=m, (the last diagonal entry) $1/\tau_m$, where $\tau_1 = \sigma_1^2$ and $\tau_j = \sigma_j^2 - \beta_j^2\sigma_{j-1}^2$ for $2 \le j \le m$. The $(j,j+1)^{th}$ and $(j+1,j)^{th}$ off–diagonal entries of Σ^{-1} for $1 \le j < m$ are $-(\beta_{j+1}/\tau_{j+1})$. All other entries of Σ^{-1} are zero.

Σ will be estimated by first estimating Σ^{-1} and then inverting to get the model based $\hat{\Sigma}$. By the results in the paragraph above, it suffices to estimate the pairs (τ_j,β_j) for j=1,2,3,.. ,m. For j>1 τ_j is the expected value of the conditional variance of Z_{ij} given the realization of Z_{ij-1} [$E(V(Z_{ij}|Z_{ij-1}))$], and τ_1 is the variance (unconditional) of Z_{i1}. For j>1 β_j is the regression coefficient of Z_{ij} on Z_{ij-1}

and β_1 is the expected value of Z_{i1}. A robust estimator for β_j is Z_{sj}/Z_{sj-1} for j>1 where Z_{sj} is the sum of the Z_{ij} over the rows of M which have data in both columns j and j–1. β_1 is estimated with the sample mean of Z_{i1} over the rows of M which have data in column one. Denote these estimators as $\{\hat{\beta}_j\}$.

Estimate τ_j with the error sum of squares of the regression of Z_{ij} on Z_{ij-1} for j>1. Then $\hat{\tau}_j$ $=(1/[n_{s_j}-1]) \sum_{i \varepsilon s_j} (Z_{ij} - \hat{\beta}_j Z_{ij-1})^2$ for j>1

and $\hat{\tau}_1 = (1/[n_{s_1}-1]) \sum_{i \varepsilon s_1} (Z_{i1}-\bar{Z}_{.1})^2$.

n_{s_j} is the number of rows in M with data (nonzero entries) in both column j and column j–1. n_{s_1} is the number of rows with data in column one. s_j is the set of rows in M with data in columns j and j–1. s_1 is the set of rows in M with data in column one. $\bar{Z}_{.1}$ in the mean of Z_{i1} over the rows in M with data in column one.

This completes the description of the REM estimate of Σ. The REM estimate of μ is derived by applying the NEM with $\hat{\Sigma}$ REM used in place of $\hat{\Sigma}(q)$ in the $(q+1)^{st}$ iteration for each q.

3. SIMULATION STUDIES

Two simulation studies were done, one with data from a Normal random number generator and a second with data that was slightly skewed.

a) Under Normality (Table 1)

The first data matrix, M, is generated exactly according to the superpopulation model given in section two. m=3, n=30, β_1= 500, β_j= 1+(j/100) for j=2, and 3; for all i, $\varepsilon_{ij} \sim N(0,\sigma_j^2)$ where σ_1^2 = 10,000 . The variance of ε_{ij} for each i and j is 225 and thus σ_2^2 = 10,629 and σ_3^2 = 11,501. From M, the entries were deleted independently at random with probability of deletion for an entry equal to .4. With a fixed M, 80 replicates of this deletion of the entries of M were used to compare the NEM and the REM. For each replicate, absolute error for both covariance matrix estimators was computed. The number of iterations till convergence were computed for the NEM. These absolute errors and iteration counts were averaged over the 80 replicates of the random deletion process.

These estimates of absolute error for the NEM and the REM estimates of Σ are given in table one together with the estimates of simulation variance, Sim Var, which measures statistical significance for the tabled entries. Simulation variance is the estimated variance of these average absolute errors, where the 80 individual replicate absolute error estimates constitute the sample (n=80) used to estimate simulation variance.

In summary, these results show that the REM has roughly one third the absolute error of the NEM covariance matrix estimator.

Iterations for the NEM were halted when the sum of the absolute differences between the components of $\hat{\mu}(q)$ and $\hat{\mu}(q-1)$ is less than .02. The average over the 80 replicates of iterations until convergence is given at the bottom of each table.

b) Under approximate χ^2. (Table 2)

Table two summarizes a simulation which parellels the table one simulation except that the rows of M are no longer Normal (but still iid). The first column of each row is $Z_{i1} = \sqrt{5000}\ \chi^2(1) + 500 - \sqrt{5000}$, where $\chi^2(1)$ is a central Chi–square random variate with one degree of freedom. The noise terms $\{\varepsilon_{ij}\}$ are independent and distributed as 50(u(0,1) – .5), where u(0,1) is a uniform random variate on the unit interval. The $\{\beta_j\}$ are the same as in a) (table one).

Table 1. Average Absolute Error (in thousands) of the Covariance Matrix Estimators over the 80 Replicates.

	Abs Err			Sim Var		
	10.8	10.9	11.2	.65	.66	.69
NEM		11.0	11.3		.67	.70
			11.5			.73
	3.0	3.1	3.2	.043	.044	.047
REM		3.2	3.3		.046	.049
			3.4			.053

AVERAGE ITERATIONS FOR NEM 26.1

Table 2. Average Absolute Error (in thousands) of the Covariance Matrix Estimators over the 80 Replicates.

	Abs Err			Sim Var		
	12.3	12.2	12.6	.89	.88	.93
NEM		12.2	12.6		.87	.92
			13.0			.97
	3.3	3.4	3.5	.036	.039	.041
REM		3.4	3.5		.042	.044
			3.6			.048

AVERAGE ITERATIONS FOR NEM 25.0

In both a) and b), the REM provides a much better estimate of the covariance matrix. In both cases, the absolute error in estimating Σ using the REM is less than one third the absolute error of the NEM estimate of Σ.

4. CONCLUSIONS

Two criteria of comparison for the NEM and the REM were considered, the error in estimating the covariance matrix and the speed of convergence. When the regression superpopulation model described in part two holds, then the REM makes good use of this additional information to both speed convergence (instantaneous for REM $\hat{\Sigma}$) and greatly reduce the error in estimating Σ.

Results for estimating μ were not tabluated here but the REM was no help in reducing the error in estimating μ. Recall that the REM estimate of μ is derived by applying the NEM with $\hat{\Sigma}$ REM in place of $\hat{\Sigma}(q)$. This is another example of a comforting property of composite estimators: in general, when their MSE is considered as a function of the component weights, this MSE is very flat in a fairly large region around the optimal weights (functions of Σ).

Although the variance/covariance estimator suggested here may be of limited value for estimating the vector of means (except for speeding up convergence), it can be useful for improving the estimates of variance for these estimators of the means.

In the introduction, it was noted that this REM covariance matrix estimate finds an important application in the Bureau's CES survey. The REM estimate of Σ is used to derive the conditional covariance matrix estimate that is used to estimate the optimal weights in an estimate for total employment at the BLS.

Although there are some small differences between the $\hat{\Sigma}$ REM and the estimator of Σ examined in Johnson and Woodruff (1990), this paper on the EM–algorithm may be considered an appendix to the J/W paper.

REFERENCES

Beale, E.M.L., and Little, R.J.A. (1975). "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*. B37, 129–145.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). "Maximum Likelihood Estimation from Incomplete Data via the EM–Algorithm (with discussion)," *Journal of the Royal Statistical Society*. B37, 1–38.

Little, R J.A. and Rubin, D.B. (1987), *Statistical Analysis With Missing Data*, Wiley.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, Wiley.

Royall, R.M. and Cumberland, W.G. (1981a), "An Empirical Study of the Ratio Estimator and Estimators of Its Variance", *Journal of the American Statistical Association*, 76, 66–77.

Woodruff, S.M., (1988), "Estimation in the Presence of Nonignorable Missing Data and a Markov Superpopulation Model," *Proceedings of the American Statistical Association* (*Survey Research Methods*).

Woodruff, S.M. (1989), "A Multivariate Approach to Estimation in Finite Population Sampling When Nonignorable Missing Data is Present," *Proceedings of the American Statistical Association* (*Survey Research Methods*).

Johnson, C.H. and Woodruff, S.M., (1990), "An Application of Regression Superpopulation Models in the Current Employment Statistics Survey," *Proceedings* of the *American Statistical Association* (*Survey Research Methods*).