Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

BAYESIAN IMPLEMENTATION OF A COMPLEX HIERARCHICAL MODEL

BY A. P. DEMPSTER AND J. S. HWANG
Harvard University

Estimates are published each month for rates of employment and unemployment in each of the 50 United States plus the District of Columbia. The basic data source for these estimates is the Current Population Survey (CPS), a national survey that samples approximately 60,000 households each month according to a complex multistage design with rotating panel structure. Estimation procedures based on successively more detailed models can lead to corresponding successive improvements in accuracy of estimation. A three phase project-in-progress (i) models the multivariate time series structure of sampling errors from 8 parallel subsamples called streams, (ii) introduces time series models for the true series that are a basis for mean square error reduction through signal extraction methods, and (iii) jointly models true and covariate time series, from nonCPS sources, to achieve still more error reduction. Models and associated Bayesian posterior-sampler computing techniques are sketched in detail for the first phase that studies sampling error only.

1. Introduction. Any segment of a national population, such as U. S. residents aged 16 to 65, may be classified using appropriate definitions into "not in the labor force", "in the labor force and unemployed", or "in the labor force and employed", leading in turn to rates of employment (EMP) and unemployment (UNEMP) on any particular date. U. S. official statistics use the definitions embodied in the Current Population Survey (CPS) which is also the basic data source for the estimates released each month.

For practical reasons related to cost, and to a lesser extent related to accuracy of estimation, the CPS does not select an independent random sample from the population each month. Before introducing the actual design, and the consequent autocovariances of sampling errors implicit in the design, it may illuminate our analysis strategy, and forestall confusions, if we separate

---

AMS 1980 Subject Classifications: Primary 62D05.
Key words and phrases: Complex survey designs, sampling error autocovariances, small area estimates.

two distinct roles played by these autocovariances. The first role is presented in detail in this paper, namely, the use of the panel structure of the design to motivate a general form of a variance-reduction technique known in the sample survey literature as compositing. The second role arises in the second and third phases of the project, not discussed in detail here, where the autocovariance structure of the sampling error is balanced against the autocovariance structure of the time series of underlying true population values of EMP and UNEMP, in search of optimum signal extraction.

The latter use of autocovariances, in combination with the joint time series behavior of other covarying time series, is capable of producing dramatic gains in efficiency, as much as tenfold. This we plan to show in later reports. The bases of such gains are (i) that smoothness of the true phenomena allow the use of projections from previous months' sample observations in combination with the current month's data, (ii) that common patterns across states allow borrowing strength from other states' correlated time series, and (iii) that nonCPS measures of the phenomena also track the survey measures and hence can be used to improve accuracy of estimation. Note that these uses of autocorrelation structure would be important even if independent samples had been drawn each month, whence no autocorrelation of sampling errors and no possibility of compositing had existed.

It is important to study compositing in advance of general time series modelling for two reasons. First, the theory of composite estimation in effect defines the efficient use of the CPS panel data. Second, the theory specifies the autocovariance structure of these efficient estimates, to be used as inputs to the later phases of analysis where optimum combination with time series properties of the underlying target phenomena are considered. Reducing these inputs by a factor of 8 yields important savings of computing effort.

The empirical input to our study is a pair of $51 \times 48 \times 8$ data arrays, one giving estimates of EMP and the other estimates of UNEMP. The two arrays come from the same surveys, and could be analyzed jointly. However, for ease of modelling and analysis, we treat them separately in two parallel analyses. The "51" dimension refers to the 50 U.S. states plus Washington DC, referred to in the sequel as "51 states". The "48" dimension refers to the 48 calendar months from Jan. 1986 through Dec. 1989, and the "8" refers to the stream dimension that provides replication internal to the survey design, and is the basis for compositing. The survey design giving rise to the stream structure is sketched in Section 2.

The variation in each of the EMP and UNEMP data sets was studied, and each was modelled using a set of 4 variance components for each of the 51 states, as detailed in Section 3. Since the variance of a rate estimator depends functionally on the mean of the estimator, we may not have constant variance components across time within each state. In the models, a variance stabilizing transform is applied to each raw estimate Z in the data arrays

and expressed as $Y = \arcsin(\sqrt{Z})$, which is known as the angular transform. The analysis has hierarchical structure in that the 51 states are treated as a random sample, or in Bayesian terms as exchangeable, so that variation both among and within states is analyzed formally. At the higher level, the 4-vectors of logs of variances for each state are treated as independent normal 4-vectors, and approximations to the likelihood functions of these 51 4-vectors are devised that allow efficient sampling of their posterior distributions. At the lower level of within state variation, the conditional distribution of the sample values given the sampled 4-vector is Gaussian, and posteriors and likelihoods are computed by standard numerical linear algebra methods.

Composite estimation for a given state in a given month may be described in terms of covariance adjustment of the raw estimates formed by averaging across the 8 streams, using as covariates the 7 contrasts among the 8 streams that in effect estimate zero. In its simplest form, compositing assumes that the autocovariances within streams are known. In practice, we sample the posterior distribution of the 4 variances, and in effect sample the posterior distribution of the within stream autocovariances that are determined by the 4 variances, thus arriving at a generalized form of compositing that averages over the posterior distribution of the composite estimates given our $51 \times 48 \times 8$ data set. We stress that averaging here includes averaging over the posteriors of the 51×4 array of variances as well as averaging over the Gaussian linear model posteriors that define composite estimation given the variances.

2. CPS Design Features. A full description of the complex multistage design and of the many adjustments made during analysis is beyond our present scope. Further details may be found in U.S. Bureau of the Census (1978). The major sources of estimation error are addressed in our analyses, but not all can be quantified from the data. For example, one important source of bias relates to the 4-8-4 rotation pattern whereby a household is interviewed in 4 successive months, dropped for 8 months, then interviewed again for 4 months. The first and fifth interviews are conducted in person, while the rest are conducted by telephone. Systematic differences are visible in the data between different techniques, whence fixed effects can be fitted, but from sample data alone it is not possible to make an adjustment which identifies the bias associated with each interview technique. In effect, it is assumed that the average bias over the 8 interviews is zero. Similarly, a selection is made among a few primary sampling units (PSUs) in rural areas that are "non-self-representing" and that are typically unchanged over years if not decades. Sampling variances from this source are by and large not assessed in our analyses, but are believed by experts to contribute very little to overall sampling error.

The ultimate sampling units (USUs) are mostly clusters of 4 neighboring households selected from a list by systematic sampling with a random start.

Each individual in a sampled household receives a weight such that aggregate sample weights produce a population of the correct size and distribution by age, sex, and race within each stream for each state and month.

Almost exactly one-eighth of the households interviewed in a given month are in each of the month-in-sample categories. Sets of households phased in over 8 month periods beginning in Dec. 1984, August 1985, $\cdots$, are called a "sample" by the survey managers, and appear roughly in ordered sets of 8, equally spaced along the list that defines the sampling frame, with separation chosen to achieve the desired sampling rate. When the households in each "sample" complete their 16 month cycle, they are replaced not by new random draws, but rather by the next USUs in the list. Consequently the data presents itself in 8 staggered replicates that we call streams, where stream 1 consists of data from households observed initially in Dec. 1984, Jan. 1985, Feb. 1985, March 1985, then households from a previous "sample" observed in their second year in April 1985 through July 1985, and so forth. Streams 2, 3, $\cdots$, 8 are similar, except that the patterns of dates are shifted ahead by one month in successive streams. Weights are computed for each individual sampled each month, separately by stream, whence the 8×48 data sets from each state can be regarded as 8 independent subsamples operating on 8 staggered timetables. Because the replacement of households within a stream is typically by neighboring households, it is evident that 8 random stream effects can be anticipated to persist across the 48 month observation period selected for study.

3. Sampling Error Model. For each state the angular transformed estimates, Y_{tj}, for EMP and UNEMP in the j^{th} stream at time t is represented as the sum of month level μ_t, month-in-sample bias effect ν, plus three random components, S_j for stream, V_{jg} for sample g within stream j,W_{jg} for annual change within sample g within stream j, and residual e_{tj},

$$Y_{tj} = \mu_t + \nu + S_j + V_{jg} + W_{jg} + e_{tj},$$
$$t = 1,2,\cdots,n,\ j = 1,2,\cdots,8 \quad \text{and} \quad g = 1,2,\cdots,g_j,$$

where $n = 48$ is the number of months and g_j, a function of t within each j, is the number of different samples in stream j. The four random components are assumed normally distributed as follows:

$$\begin{aligned} S_j &\sim \mathrm{N}\,(0,\sigma^2_{\mathrm{str}}) \\ V_{jg} &\sim \mathrm{N}\,(0,\sigma^2_{\mathrm{sam}}) \\ W_{jg} &\sim \mathrm{N}\,(0,\sigma^2_{\mathrm{lag}}) \\ e_{tj} &\sim \mathrm{N}\,(0,\sigma^2), \end{aligned}$$

and all these random quantities are assumed independent.

The standard estimates of fixed effects that may be obtained either by maximum likelihood or by generalized least squares are also the Bayesian posterior means obtained in the limiting case where the prior variance of the effects tends to infinity. Because both Bayesians and non-Bayesians estimate random effects by using conditional means given the data and the variance parameters, it follows that the Bayesian approach unifies the treatment of the fixed and random effects.

In order to apply general computing algorithms, we rewrite the specific model in the general form

$$Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3,$$

where β_i is normal with mean vector zero and covariance matrix Σ_i.

The connections between the general and specific notations are:

$$\begin{aligned}
Y^T &= (Y_{11}, \cdots, Y_{1n}, \cdots, Y_{81}, \cdots, Y_{8n}),\\
\beta_1^T &= (\mu_1, \mu_2, \cdots, \mu_n, \nu),\\
\beta_2^T &= (S_1, \cdots, S_8, V_{11}, \cdots, V_{1g_1}, \cdots, V_{81}, \cdots, V_{8g_8},\\
&\qquad W_{11}, \cdots, W_{1g_1}, \cdots, W_{81}, \cdots, W_{8g_8})\\
\beta_3^T &= (e_{11}, \cdots, e_{1n}, \cdots, e_{81}, \cdots, e_{8n}),
\end{aligned}$$

where the three covariance matrices are

$$\begin{aligned}
&\Sigma_1 \to \infty,\\
&\Sigma_2 = \operatorname{diag}(\sigma^2_{\text{str}} I_{k_1},\ \sigma^2_{\text{sam}} I_{k_2},\ \sigma^2_{\text{lag}} I_{k_3}),\\
&\Sigma_3 = \sigma^2 I_{k_4},
\end{aligned}$$

and the dimensions are $k_1 = 8$, $k_2 = k_3 = g_1 + \cdots + g_8$, $k_4 = 8n$. Finally, the three design matrices are $X_3 = I_{k_4}$ and

$$X_1 = \begin{bmatrix} I_n & \vdots \\ \vdots & c_{kj} \\ I_n & \vdots \end{bmatrix}, \qquad c_{kj} = \begin{cases} 1 & \text{if the } k^{th} \text{ month sample is the} \\ & \text{first or fifth one in stream } j, \\ -1 & \text{otherwise,} \end{cases}$$

$$X_2 = \left[\begin{array}{cccc|cccc|cccc} J & & & O & A_1 & & & O & B_1 & & & O \\ & J & & & & A_2 & & & & B_2 & & \\ & & \ddots & & & & \ddots & & & & \ddots & \\ O & & & J & O & & & A_8 & O & & & B_8 \end{array}\right]$$

where

$$J^T = (1, 1, \cdots, 1), \qquad A_j = (a_{jkl}), \qquad B_j = (b_{jkl}),$$

$$a_{jkl} = \begin{cases} 1 & \text{if the } k^{th} \text{ month sample is} \\ & \text{the } l^{th} \text{ one in stream } j, \\ 0 & \text{otherwise,} \end{cases}$$

$$b_{jkl} = \begin{cases} 1 & \text{if the } k^{th} \text{ month sample is the } l^{th} \text{ one} \\ & \text{in stream } j \text{ and in the first 4 months,} \\ -1 & \text{if the } k^{th} \text{ month sample is the } l^{th} \text{ one} \\ & \text{in stream } j \text{ and in the last 4 months,} \\ 0 & \text{otherwise.} \end{cases}$$

4. Estimation of Variance Components. Estimation methods for the linear covariance components model were developed and illustrated in Dempster, Rubin and Tsutakawa (1981). These techniques include Bayesian estimation of fixed and random effects when the variances and covariances are known and point estimation of unknown variances and covariances using an EM algorithm.

Spelling out the details, we begin with the joint normal distribution

$$\begin{bmatrix} Y \\ \beta \end{bmatrix} \sim \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} Q & X\Sigma \\ \Sigma X^T & \Sigma \end{bmatrix} \right), \tag{1}$$

where

$$\begin{aligned} X &= (X_1,\ X_2,\ X_3), \\ \beta^T &= (\beta_1^T,\ \beta_2^T,\ \beta_3^T), \\ \Sigma &= \text{diag}\,(\Sigma_1,\ \Sigma_2,\ \Sigma_3), \\ Q &= X\Sigma X^T \\ &= X_1\Sigma_1 X_1^T + X_2\Sigma_2 X_2^T + \sigma^2 I_{k_4}. \end{aligned}$$

The conditional distribution of β given Y is

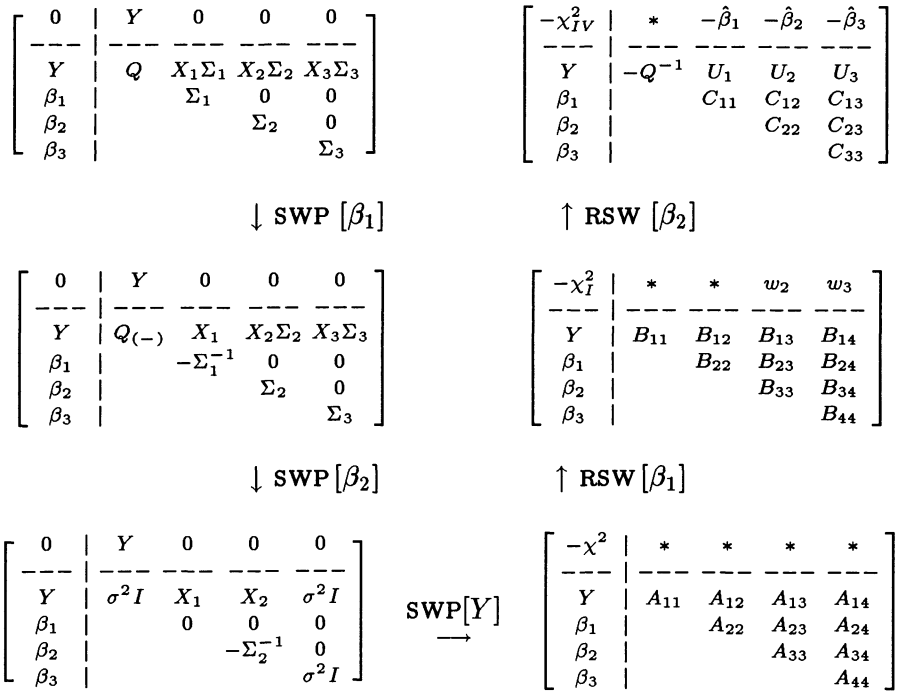
$$\beta \mid Y \sim \text{N}(\hat{\beta}, C), \tag{2}$$

where

$$\begin{aligned} &\hat{\beta} = U^T Y, \quad U = Q^{-1} X\Sigma, \\ &C = \Sigma - \Sigma X^T Q^{-1} X\Sigma. \end{aligned}$$

The operation of finding the above conditional distribution (2) from the marginal distribution (1) can be expressed in SWP terms as shown from the upper left array to the upper right array in Figure 1. The computing operations SWP (for sweep) and RSW (for reverse sweep) provide compact derivations of formulas for Bayesian linear model calculations. The definitions of

the these operators and their properties are illustrated briefly in appendix. See also, e.g., Dempster (1969, 1982) and Carlin (1990) for more detailed discussions and applications.



$$\chi_I^2 = Y(I-H)Y^T\sigma^{-2} \qquad \chi_{IV}^2 = YQ^{-1}Y^T$$
$$H = X_1(X_1^TX_1)^{-1}X_1^T \qquad \chi^2 = YY^T\sigma^{-2}$$
$$B_{11} = -\sigma^{-2}(I-H) \qquad A_{11} = -\sigma^{-2}I$$
$$B_{12} = X_1(X_1^TX_1)^{-1} \qquad A_{12} = \sigma^{-2}X_1$$
$$B_{13} = \sigma^{-2}(I-H)X_2 \qquad A_{13} = \sigma^{-2}X_2$$
$$B_{14} = I-H \qquad A_{14} = I$$
$$B_{22} = \sigma^2(X_1^TX_1)^{-1} \qquad A_{22} = -\sigma^{-2}X_1^TX_1$$
$$B_{23} = -(X_1^TX_1)^{-1}X_1^TX_2 \qquad A_{23} = -\sigma^{-2}X_1^TX_2$$
$$B_{24} = -\sigma^2(X_1^TX_1)^{-1}X_1^T \qquad A_{24} = -X_1^T$$
$$B_{33} = -\Sigma_2^{-1} - \sigma^{-2}X_2^T(I-H)X_2 \qquad A_{33} = -\Sigma_2^{-1} - \sigma^{-2}X_2^TX_2$$
$$B_{34} = -X_2^T(I-H) \qquad A_{34} = -X_2^T$$
$$B_{44} = \sigma^2 H \qquad A_{44} = 0$$
$$w_2 = -Y^TB_{13} \qquad w_3 = -Y^TB_{14}$$

Figure 1. Schematic Picture of Sweeps

Clearly, taking the limit $\Sigma_1 \to \infty$ in the prior distribution results in an improperly defined probability measure. But under mild conditions, see Dempster (1982), we can deal with the infinite variances case using the algorithm embodied in the SWP/RSW manipulation. We first carry out a SWP on indices corresponding to β_1, and later undo this operation with a RSW, meanwhile allowing $\Sigma_1^{-1} \to 0$. This first step SWP$[\beta_1]$ may be represented in the usual framework as shown in the first two arrays of the schematic picture of sweeps in Figure 1.

When the $\dim(Q)$ is much bigger than $\dim(\Sigma_2)$, we may run another SWP on β_2 at the second step SWP$[\beta_2]$, and later we will undo this operation with RSW on β_2 again. This is because the block corresponding to Y reduces to $\sigma^2 I_{k_4}$ after sweeping on β_2. Processing SWP$[Y]$ is then trivial.

The real computations are the fourth step RSW$[\beta_1]$ and the final step RSW$[\beta_2]$. The resulting array equals that which would be obtained by simple application of SWP$[Y]$ alone to the original array. For this identity we require only that $Q_{(-)} = Q - X_1 \Sigma_1 X_1^T$ be of full rank, a trivially satisfied condition in the present case. But, the suggested 5-step route to the final array would reduce computation burden and the possibility of numerical roundoff errors.

The basis of the EM algorithm in the standard variance components model is that the observable vector Y is written as a linear combination of unobservable random effects β_2, β_3, each of whose distributions is normal with mean 0 and variance known up to a simple scale factor (" the component of variance"). If the random effects were in fact observable quantities, maximum likelihood estimates of the scale factors would be available very simply in the usual way, by equating "observed" and "expected" variances. The EM algorithm proceeds by alternatively (E-Step) filling in values for the random effects (more precisely, for the corresponding sufficient statistics), using conditional expectations given the observed Y and current estimated variances, and (M-step) obtaining new estimates from the filled-in data. This iterative procedure can be shown in general (Dempster, Laird and Rubin 1977) to increase the likelihood at every step. In our specific model the processing of the above sweep operations will complete an E-step. The M-step of the EM algorithm is then trivial.

Conditions for convergence to a global maximum of the likelihood are complicated, but practical experience with variance component models has shown generally good results, although convergence may be very slow. Fortunately there are several possible ways of speeding up iterations, for example, a simple algorithm in Hwang (1992).

5. Computing Approximate Likelihood. To adjust for different sample sizes, the four variance estimates in each state are multiplied by the state's household number to a per household basis. Since the survey designs are similar across states, variances per household should be roughly stable. We

will explore the data for evidence of variation of the underlying true variance components over the 51 states. Most especially in the case of the stream variance, since there are only 7 degrees of freedom among the 8 streams in each state, there is far too much sampling variation to think of substituting maximum likelihood estimates directly into models. The maximum likelihood estimates, certainly for stream variance, cannot possibly be accurate and so are not very informative about the phenomenon. One sensible method is to devise a sampling scheme to draw sets of four variances from a posterior distribution.

The posterior is proportional to the likelihood function times prior. The direct computation of the exact posterior requires more likelihood computations than we can afford. We therefore develop simple approximations that both are accurate enough and greatly simplify the computations of estimating the four variances.

There are $N = 48\times8$ observations for each state in the 48 months data. There are 49 degrees of freedom for the fixed row effects and a bias effect, n_1=7 degrees of freedom for streams, $n_2 = 47$ degrees of freedom for sample, $n_3 = 47$ degrees of freedom for sample-lag, leaving $n_4 = N - 150$ for residual. The five subspaces that span the N-space are not exactly orthogonal, and even worse the degree of nonorthogonality changes a bit as the variances themselves change. In fact, this is what makes the exact computation of likelihood difficult, i.e., we must reinvert big covariance matrices every time we change the values of the variances in EM steps. But suppose the five subspaces are very nearly orthogonal whatever the values of the variances. Then there are four approximately sufficient statistics, and estimating the four variances is just a matter of equating these sufficient sums of squares with their expectations.

Recall the schematic picture of sweeps in Figure 1, the first element of the extra row of the right bottom matrix, $-YY^T\sigma^{-2}$, after SWP[β_1], SWP[β_2], and SWP[Y] implemented, is just the familiar sum of squares of ANOVA treated with a minus sign and division by σ^2. If we look at what the RSW[β_1] step does to this sum of squares, we see that it is just doing a similarly doctored version of ordinary multiple regression on the fixed effects, and in effect reduces the sum of squares to the residual sum of squares after removing the 49 degrees of freedom for fixed effects in the usual way.

Next we consider what happens to this residual sum of squares in the RSW[β_2] step. This step includes the three sets of random effects for stream, sample within stream, and sample-lag within stream. During these three successive sweeping operations, the sum of squares continues to shrink just as it did in the preceding RSW[β_1] step.

Apart from the negative sign, these two values are all chi-squares with the same degrees of freedom $N - 49$, specifically, the model with just the fixed effects and error terms, that is χ^2_I, and the model that adds stream random

effects, sample random effects, and sample-lag random effects χ^2_{IV}, where

$$\chi^2_I = Y(I - H)Y^T\sigma^{-2} > \chi^2_{IV} = YQ^{-1}Y^T.$$

Another way to describe what these chi-squares are is to say that after multiplying by $-\frac{1}{2}$ they are the exponents in the normal likelihood function of the data under the random effects model.

The result here is exact, i.e., not dependent on approximate orthogonality, so it can be used as part of exact likelihood calculation of likelihood of fitted models as we go through the sweeps. The other factor in the likelihood is the determinant of the covariance matrix of the normal. This determinant may be carried forward during the sweeps using the formula (3.17) in Dempster (1982).

The corrected sum of squares after removing fixed effects, $SS = \sigma^2\chi^2_I$, can be decomposed in some approximate sense into

$$SS = SS_1 + SS_2 + SS_3 + SS_4$$

with n_1, n_2, n_3, and n_4 degrees of freedom, where

$$\begin{aligned}\chi^2_I &= \frac{SS_1}{\sigma^2} + \frac{SS_2}{\sigma^2} + \frac{SS_3}{\sigma^2} + \frac{SS_4}{\sigma^2} \\ \chi^2_{IV} &= \frac{SS_1}{\sigma^2 + n\sigma^2_{str} + 8\sigma^2_{sam} + 8\sigma^2_{lag}} \\ &\quad + \frac{SS_2}{\sigma^2 + 8\sigma^2_{sam}} + \frac{SS_3}{\sigma^2 + 8\sigma^2_{lag}} + \frac{SS_4}{\sigma^2}.\end{aligned}$$

The above formulas are obvious from Gaussian linear model theory. The theory says that $Y_{(-)}$, the data vector minus fixed effects, has a multivariate normal distribution with mean vector zero and covariance matrix of rank $N - 49$ whose components in four subspaces of dimensions n_1, n_2, n_3, and n_4 are independent under any choices of the four variances. The two different chi-squares are just the chi-squares in the exponent of the multivariate normal under two choices of the variances, namely $(\sigma^2, 0, 0, 0)$ and $(\sigma^2, \sigma^2_{str}, \sigma^2_{sam}, \sigma^2_{lag})$. The various denominators are just the expected mean squares which represent the variances of the normal in the associated directions.

Under orthogonality the likelihood would be proportional to

$$\tau_1^{\frac{n_1}{2}}\ \tau_2^{\frac{n_2}{2}}\ \tau_3^{\frac{n_3}{2}}\ \tau_4^{\frac{n_4}{2}} \times \exp\left\{-\frac{1}{2}(SS_1\tau_1 + SS_2\tau_2 + SS_3\tau_3 + SS_4\tau_4)\right\}, \qquad (3)$$

where

$$\begin{aligned}\tau_1 &= (\sigma^2 + n\sigma^2_{\text{str}} + 8\sigma^2_{\text{sam}} + 8\sigma^2_{\text{lag}})^{-1} \\ \tau_2 &= (\sigma^2 + 8\sigma^2_{\text{sam}})^{-1} \\ \tau_3 &= (\sigma^2 + 8\sigma^2_{\text{lag}})^{-1} \\ \tau_4 &= \sigma^{-2}.\end{aligned}$$

The four sums of squares in (3) are replaced with

$$\begin{aligned} SS_1 &= n_1\ (\hat{\sigma}^2 + n\hat{\sigma}^2_{\text{str}} + 8\hat{\sigma}^2_{\text{sam}} + 8\hat{\sigma}^2_{\text{lag}}) \\ SS_2 &= n_2\ (\hat{\sigma}^2 + 8\hat{\sigma}^2_{\text{sam}}) \\ SS_3 &= n_3\ (\hat{\sigma}^2 + 8\hat{\sigma}^2_{\text{lag}}) \\ SS_4 &= n_4\ \hat{\sigma}^2, \end{aligned} \tag{4}$$

where $\hat{\sigma}^2, \hat{\sigma}^2_{\text{str}}, \hat{\sigma}^2_{\text{sam}}$, and $\hat{\sigma}^2_{\text{lag}}$ are the maximum likelihood estimates.

The 4 components $SS_i\tau_i$ are chi-squares with degrees of freedom n_i. To draw a sample of 4 variances from (3) is equivalent to the following: draw 4 chi-square values from each $\chi^2_{n_i}$; set the 4 chi-square values equal to the 4 $SS_i\tau_i$ and solve the 4 equations.

However, for some of the variances, solving could produce a negative value, and any among the samples for which this happened, would be rejected as not being in the parameter space, so having zero prior and posterior probability density. For the states with extremely small stream variances, the rejection fraction in the original chi-square sampling would be large. The marginal distributions of the simulated 4 log transformed variances are bell-shaped, and three of them are reasonable symmetric (not stream component), so normal fits them more or less well. We then assume the simulated 4 variances in log scale are approximate normal, and let the sample mean and sample variance of the i^{th} variance component of the simulated sample for the k^{th} state be θ_{ki} and η_{ki}.

6. Posteriors of Variance Components. After obtaining the approximate likelihood function we choose proper prior distributions for the four variances for each state to get a posterior distribution of the four variances. The prior distribution of the four variances in principle could be different for each state, e.g., if we had special information about the sample design for each state. Instead, however, we use variation among the states to arrive at a plausible prior for all states. This is the idea behind "borrowing strength" across the states, which could also be called empirical Bayes, the idea being that we treat all states as *exchangeable* for the purposes of formalizing prior uncertainty about the four *true* variances of any particular state.

We start by looking directly at original maximum likelihood estimates of the eight variances, i.e., four each for unemployment rates and employment rates. The maximum likelihood estimates are all multiplied by the number of households in each state to standardize for sample size. The histograms of the log transforms of the variances are fairly normal except stream components which have a cluster of extremely small values. The pairwise scatter plots of the log variance estimates also show little correlation. Therefore it is natural to start with four independent log normal priors for the four variances. The means and standard deviations of the log normal distributions may be obtained

from the sample means and sample standard deviations of the 51×4 maximum likelihood estimates (in log scale).

Let δ_i and ζ_i be the log normal prior mean and variance of i^{th} variance component. Then, the posterior of the i^{th} variance component for the k^{th} state is log normal with mean and variance

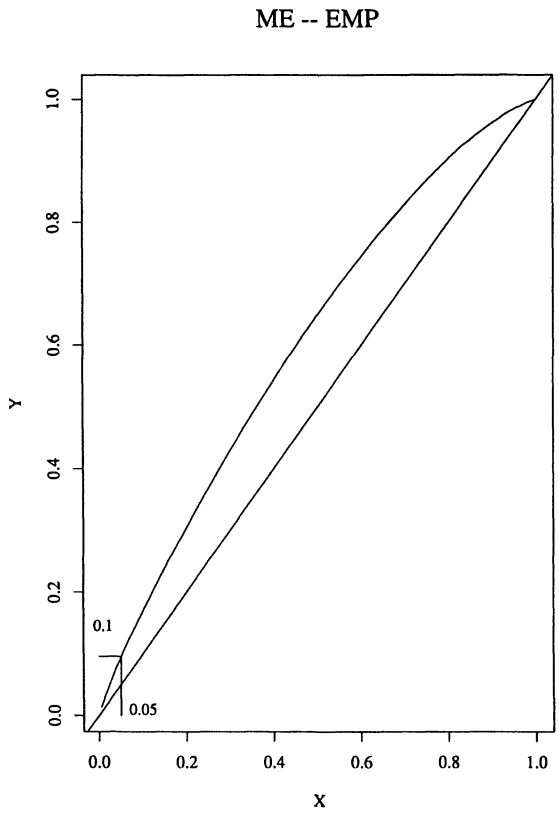


Figure 2. Frequencies of frequencies plot

$$\frac{\theta_{ki}/\eta_{ki} + \delta_i/\zeta_i}{\eta_{ki}^{-1} + \zeta_i^{-1}} \qquad \text{and} \qquad (\eta_{ki}^{-1} + \zeta_i^{-1})^{-1}.$$

The next step is to assess the approximate log normal posterior from which we attempt to draw. One way is to look at the distribution of the correct importance sampling weights, the ratio of exact likelihood times the prior and the approximate log normal posterior. The basic question is whether the distribution of weights is acceptable. In other words, we cannot use this approximate posterior when the total weights is provided by a few largest weights. A descriptive graphic for detecting too many large values in a sample is the *frequencies of frequencies* plot. For example, suppose we have computed a large sample of weights, ordered them from largest to smallest, and calculated the fraction y of the total weight provided by each fraction x of the largest weights. The degree of nonlinearity of the plot of y vs x indicates nonconstancy

of the weights, as illustrated in Figure 2 for EMP in the state of Maine (ME), where the largest 5% of the weights provide 10% of the total weight, not a severe deviation from the straight line. Figure 2 is typical of our Monte Carlo runs.

Table 1. The posterior variance ratios of raw estimators to composite estimators in the transform scale

State	EMP	UNEMP	State	EMP	UNEMP	State	EMP	UNEMP
AL	1.602	1.119	KY	1.622	1.101	ND	1.710	1.109
AK	1.408	1.081	LA	1.696	1.125	OH	1.581	1.123
AZ	1.436	1.088	ME	1.525	1.065	OK	1.542	1.043
AR	1.681	1.081	MD	1.470	1.033	OR	1.591	1.070
CA	1.681	1.061	MA	1.546	1.104	PA	1.551	1.088
CO	1.473	1.064	MI	1.636	1.119	RI	1.553	1.050
CT	1.516	1.049	MN	1.662	1.083	SC	1.686	1.107
DE	1.577	1.049	MS	1.561	1.097	SD	1.548	1.092
DC	1.471	1.056	MO	1.650	1.102	TN	1.609	1.057
FL	1.431	1.037	MT	1.541	1.110	TX	1.550	1.054
GA	1.651	1.085	NE	1.669	1.074	UT	1.617	1.101
HI	1.553	1.061	NV	1.604	1.089	VT	1.667	1.049
ID	1.454	1.098	NH	1.543	1.059	VA	1.673	1.082
IL	1.604	1.101	NJ	1.685	1.109	WA	1.496	1.078
IN	1.588	1.054	NM	1.608	1.083	WV	1.537	1.063
IA	1.693	1.080	NY	1.521	1.088	WI	1.553	1.084
KS	1.679	1.097	NC	1.607	1.067	WY	1.507	1.141

7. Optimal Composite Estimation. We propose a k-lag composite estimator of the t^{th} month rate, which is the sum of the t^{th} month weighted average and the previous k months' adjustments. Let $Y_{t,j}$ be the angular transformed estimate at month t from the j^{th} stream, and the unbiased composite estimator in the transform scale is defined as

$$Y_t^{[k]} = \sum_{j=1}^{8} c_j Y_{t,j} + \sum_{l=1}^{k} \sum_{j=1}^{8} b_{lj} Y_{t-l,j},$$

where

$$\sum_{j=1}^{8} c_j = 1, \qquad \sum_{j=1}^{8} b_{lj} = 0 \qquad \text{for} \quad l = 1, \cdots, k.$$

The special estimator defined by $k = 0$ and $c_j = 1/8$, $j = 1, 2, \cdots, 8$ is the uncomposited or raw estimator that simply averages the original stream estimates in the transform scale. In the following, we describe how to obtain an optimal composite estimate under the sampling error model.

Given a sample of the four variances from the approximate log normal posterior, the conditional estimate of the fixed effect of the n^{th} month is the optimal $(n-1)$-lag composite estimator with coefficients c_j and b_{lj} found in U_1 in Figure 1. The conditional posterior variance of the raw estimator is the sum of the simulated four variances divided by 8. The conditional posterior means and variances of the composite estimators were discussed in Section 3. These are shown as in $\hat{\beta}_1$ and C_{11} in Figure 1.

To obtain approximate posterior means and variances of the composite and raw estimators, we repeat drawing four variances from the approximate log normal posterior and dividing the four variances by each state's household number back to a per sample basis. The posterior variance of the raw estimator is the average of these conditional posterior variances. The posterior mean of the composite estimator is the average of the simulated conditional posterior means. The approximate posterior variance is the sample variance of the simulated means plus the average of the simulated conditional posterior variances.

Table 1 shows the posterior variance ratios of raw estimators to composite estimators in the transform scale, which are based on 200 draws of the four variances from the approximate log normal posteriors. The optimal UNEMP composite estimates improve about 5%. There are substantial gains for the optimal EMP composite estimates.

Acknowledgements. This work has been supported in part by National Science Foundation grant DMS-90-03216, and also by the Bureau of Labor Statistics through Task Orders 7 and 14.

Appendix: The SWP Operator — Definition and Properties. Suppose that M is an $r \times r$ symmetric matrix with (i,j) element m_{ij}. For any k such that $1 \leq k \leq r$, *sweeping*, or *pivoting*, on k produces a new $r \times r$ matrix which we denote by SWP[k]: M and has (i,j) elements m^*_{ij}, where

$$\begin{aligned} m^*_{kk} &= -1/m_{kk} \\ m^*_{ik} &= m^*_{ki} = m_{ik}/m_{kk} \\ m^*_{ij} &= m_{ij} - m_{ik}m_{kj}/m_{kk}, \qquad \text{for} \quad i \neq j,\ j \neq k. \end{aligned} \tag{5}$$

The value m_{kk} found at the k^{th} diagonal element of M before applying SWP[k] is called the *pivot* element associated with SWP[k]. It is straightforward to check that the elementary sweep operations SWP[1], SWP[2], ..., SWP[r] are commutative. Hence one can define SWP[$\boldsymbol{u}$] for any subset $\boldsymbol{u}$ of

the integers $\{1, \cdots, r\}$ to the result of applying SWP[k] for each $k \in \boldsymbol{u}$, in any order.

The result of the block form of SWP is easily displayed if we assume that $\boldsymbol{u} = \{1, 2, \cdots, s\}$ and $\boldsymbol{v} = \{s+1, \cdots, r\}$, and make a corresponding partition of M. For any s on $1 \leq s \leq r$, it may readily be shown (e.g., Dempster, 1969) that

$$\text{SWP}[\boldsymbol{u}] \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} = \begin{bmatrix} -M_{11}^{-1} & M_{11}^{-1} M_{12} \\ M_{12}^T M_{11}^{-1} & M_{22} - M_{12}^T M_{11}^{-1} M_{12} \end{bmatrix} \tag{6}$$

Note that the block form (6) reduces to the single-index form (5) in the case $r = 2$, $s = 1$, as it obviously must.

The effect of SWP is readily reversed by an analogous operator, RSW, which plays an important role in the methods described in this paper. It is defined that RSW[k] is identical to SWP[k] except for a change of sign in the second line of the definition. The block version, RSW[$\boldsymbol{u}$], is defined analogously, and gives the same result as (6) except for a change of sign in the off-diagonal block.

It is clear that successive application of SWP[k] in any order to M yields $-M^{-1}$. An important by-product of performing SWP[1], $\cdots$, SWP[r] is that the determinant $\det M$ may be calculated by multiplying the pivot elements associated with each of the r steps.

REFERENCES

CARLIN, J. B. (1990). An algorithmic approach to Bayesian linear model calculations. *Austral. J. Statist.*, **32**(1), 29–43.

DEMPSTER, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley, Reading, Mass.

DEMPSTER, A. P. (1982). Some formulas useful for covariance estimation with Gaussian linear component models. In Statistics and Probability: Essays in Honor of C.R. Rao, eds. G. Kallianpur *et al.*, Amsterdam: North-Holland, 213–229.

DEMPSTER, A. P. LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist.* Soc. B, **39**, 1–38.

DEMPSTER, A. P., RUBIN, D. B. and TSUTAKAWA, R. K. (1981). Estimation in covariance components models. *J. Amer. Statist. Assoc.* **76**, 341–353.

HWANG, J. S. (1992). Prototype Bayesian estimation of US state employment and unemployment rates. Ph. D. Thesis, Department of Statistics, Harvard University.

U.S. Bureau of the Census (1978). The Current Population Survey: Design and methodology, Technical Paper 40, Washington, D.C.

DEPARTMENT OF STATISTICS
HARVARD UNIVERSITY
ONE OXFORD ST.
CAMBRIDGE, MA 02138
U.S.A.