Notes on a Multivariate Fay-Herriot Model with AR(1) Model Errors

William Bell

April 19, 2012

Multivariate Fay-Herriot (FH) model:

$$\begin{aligned} \mathbf{y}_i &= \mathbf{Y}_i + \mathbf{e}_i \\ &= (\mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i) + \mathbf{e}_i \end{aligned}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{in})'$ are the $n \times 1$ vectors of observations (direct survey estimates) for areas $i = 1, \ldots, m$, and $\mathbf{Y}_i$ are the corresponding $n \times 1$ vectors of true population quantities estimated by $\mathbf{y}_i$ with sampling error vectors $\mathbf{e}_i$. The t^{th} equation for $t = 1, \ldots, n$ can be written as

$$y_{it} = x_{it}'\beta_t + u_{it} + e_{it},$$

which allows for different regression parameters β_t in each equation. Note that

$$\mathbf{X}_i\boldsymbol{\beta} = \begin{bmatrix} x_{i1}' & 0 & \cdots & 0 \\ 0 & x_{i2}' & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & x_{in}' \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}.$$

One could restrict the model to have common regression parameters β in each equation, in which case $\mathbf{X}_i\boldsymbol{\beta}$ would collapse to

$$\begin{bmatrix} x_{i1}' \\ x_{i2}' \\ \vdots \\ x_{in}' \end{bmatrix} \beta.$$

One could also use intermediate possibilities such as different intercept parameters for each equation but with the remaining $\beta_j s$ common across the n equations.

The general multivariate model would have $\text{Var}(\mathbf{u}_i) = \Sigma_u$ where Σ_u is a general, symmetric positive definite $n \times n$ matrix ($n(n+1)/2$ distinct elements). If n is not too large, this general model is feasible. Also, we treat $\text{Var}(\mathbf{e}_i) = \Sigma_{e_i}$ as known (estimated from survey micro-data, though possibly also smoothed over areas in some way). Assuming $\mathbf{u}_i$ and $\mathbf{e}_i$ are both independent over areas i, our model for the observed data is

$$\mathbf{y}_i \sim ind.\ N(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_{y_i}) \qquad i = 1, \ldots, m$$

where

$$\boldsymbol{\Sigma}_{y_i} = \boldsymbol{\Sigma}_u + \boldsymbol{\Sigma}_{e_i}.$$

Autoregressive (AR) models for $\mathbf{u}_i$:

If the n observations in $\mathbf{y}_i$ are for different time points, then we may specify $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_{e_i}$ to correspond to time series models reflecting the dependence in u_{it} and e_{it} where t indexes the time points. For ACS data, we assume the e_{it} are independent over time, so that

$$\mathbf{\Sigma}_{e_i} = \begin{bmatrix} \sigma^2_{e_{i1}} & & \\ & \ddots & \\ & & \sigma^2_{e_{in}} \end{bmatrix}.$$

Some alternative first-order AR models can be specified for u_{it}. These determine the form of $\mathbf{\Sigma}_u \equiv \mathbf{\Sigma}_u(\psi)$, where ψ denotes the unknown parameters to be estimated in the model for u_{it}.

Stationary AR(1): $u_{it} = \rho u_{i,t-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim$ *i.i.d.* $N(0, \sigma^2_\varepsilon)$, $\rho \in (-1, 1)$, $\sigma^2_u = \sigma^2_\varepsilon/(1-\rho^2)$, $\psi = (\sigma^2_u, \rho)'$,

$$\mathbf{\Sigma}_u = \sigma^2_u \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \cdots & \rho^2 & \rho & 1 \end{bmatrix}$$

Heteroskedastic AR(1): $\psi = (\sigma^2_{u_1}, \ldots, \sigma^2_{u_n}, \rho)'$,

$$\mathbf{\Sigma}_u = \begin{bmatrix} \sigma^2_{u_1} & & & \\ & \sigma^2_{u_2} & & \\ & & \ddots & \\ & & & \sigma^2_{u_n} \end{bmatrix} \begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \cdots & \rho & 1 \end{bmatrix} \begin{bmatrix} \sigma^2_{u_1} & & & \\ & \sigma^2_{u_2} & & \\ & & \ddots & \\ & & & \sigma^2_{u_n} \end{bmatrix}.$$

For the case of $n = 2$, this model is equivalent to a general bivariate model.

Random walk (AR(1) with $\rho = 1$): $u_{it} = u_{i,t-1} + \varepsilon_{it}$, $\varepsilon_{it} \sim$ *i.i.d.* $N(0, \sigma^2_\varepsilon)$, $\psi = (\sigma^2_{u_1}, \sigma^2_\varepsilon)'$,

$$\mathbf{\Sigma}_u = \sigma^2_{u_1} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \sigma^2_\varepsilon \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2 & \cdots & n-1 \end{bmatrix}.$$

Notice that this $\mathbf{\Sigma}_u$becomes singular if either $\sigma^2_{u_1} \to 0$ or $\sigma^2_\varepsilon \to 0$, so we must watch if parameter estimation heads towards either of these boundary values. Also, this covariance matrix is ill-conditioned if n is large, but this should not be a problem if n is small.

Maximum likelihood (ML) estimation:

The Gaussian log-likelihood function to be maximized over ψ and $\boldsymbol{\beta}$ is

$$\ell(\psi, \boldsymbol{\beta}) = -\frac{mn}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left\{\log|\boldsymbol{\Sigma}_{y_i}| + (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Sigma}_{y_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right\}.$$

Maximizing $\ell(\psi, \boldsymbol{\beta})$ can be accomplished by minimizing the summation term above, which can be done by iterative generalized least squares. At iteration r:

1. Given estimates $\hat{\boldsymbol{\beta}}^{(r)}$, maximize $\ell^{(r)}(\psi) \equiv \ell(\psi, \hat{\boldsymbol{\beta}}^{(r)})$ over ψ to get updated estimates $\hat{\psi}^{(r)}$.

2. Given $\hat{\psi}^{(r)}$, estimate $\boldsymbol{\beta}$ by generalized least squares:

$$\hat{\boldsymbol{\beta}}^{(r)} = \left(\sum_{i=1}^{m}\mathbf{X}_i'[\boldsymbol{\Sigma}_{y_i}^{(r)}]^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{m}\mathbf{X}_i'[\boldsymbol{\Sigma}_{y_i}^{(r)}]^{-1}\mathbf{y}_i\right) \tag{1}$$

   where $\boldsymbol{\Sigma}_{y_i}^{(r)} = \boldsymbol{\Sigma}_u(\hat{\psi}^{(r)}) + \boldsymbol{\Sigma}_{e_i}$.

We could start this process by estimating $\boldsymbol{\beta}$ by OLS, or by setting $\boldsymbol{\Sigma}_{y_i}^{(r)} = \Sigma_{e_i}$ and computing $\boldsymbol{\beta}$ from (1). Convergence of the above iterations should be fairly rapid. At the end, we have the MLEs $\hat{\psi}$ and $\hat{\boldsymbol{\beta}}$. The large sample covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{m}\mathbf{X}_i'\hat{\boldsymbol{\Sigma}}_{y_i}^{-1}\mathbf{X}_i\right)^{-1} \tag{2}$$

where $\hat{\boldsymbol{\Sigma}}_{y_i} = \boldsymbol{\Sigma}_u(\hat{\psi}) + \boldsymbol{\Sigma}_{e_i}$. The large sample covariance matrix of $\hat{\psi}$ is given by

$$\mathrm{Var}(\hat{\psi}) \approx -\left[\frac{\partial^2}{\partial\psi_i\partial\psi_j}\ell(\psi, \hat{\boldsymbol{\beta}})\right]^{-1}$$

evaluated at $\hat{\psi}$. This can be computed using numerical derivatives, though many optimization routines will automatically provide it. The large sample covariance between $\hat{\psi}$ and $\hat{\boldsymbol{\beta}}$ is zero.

One can also use a Bayesian approach to inference, which will require some form of simulation. A standard prior for these models would be a flat prior just restricting the variances to be non-negative and $\rho \in (-1, 1)$. With this prior the posterior, $p(\psi, \boldsymbol{\beta}|\mathbf{y})$, where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)'$ is all the data, is proportional to the likelihood, and a large sample approximation to the posterior treats it as normal with mean vector given by the MLEs $\hat{\psi}$ and $\hat{\boldsymbol{\beta}}$, and covariance matrix given by the large sample covariance matrix of $\hat{\psi}$ and $\hat{\boldsymbol{\beta}}$. One can easily simulate from this normal approximation to the posterior.

For exact Bayesian calculations it is useful to note that

$$p(\psi, \boldsymbol{\beta}|\mathbf{y}) = p(\psi|\mathbf{y})p(\boldsymbol{\beta}|\mathbf{y},\psi)$$

where $p(\boldsymbol{\beta}|\mathbf{y},\psi)$ is a normal density with mean and covariance matrix given by the GLS results (1) and (2), and $p(\psi|\mathbf{y})$ can be readily calculated from $p(\psi, \boldsymbol{\beta}|\mathbf{y})/p(\boldsymbol{\beta}|\mathbf{y},\psi)$ $= p(\psi, \boldsymbol{\beta})\ell(\psi, \boldsymbol{\beta})/p(\boldsymbol{\beta}|\mathbf{y},\psi)$ where $p(\psi, \boldsymbol{\beta})$ is the prior (possibly just a constant). If one can simulate from $p(\psi|\mathbf{y})$, appending to each simulated ψ a simulation of $\boldsymbol{\beta}$ from its conditional normal distribution, $p(\boldsymbol{\beta}|\mathbf{y},\psi)$, yields a simulation from $p(\psi, \boldsymbol{\beta}|\mathbf{y})$. For the first and third models discussed above, ψ has only two elements, so independent simulations from $p(\psi|\mathbf{y})$ could be obtained by rejection sampling. Alternatively, the Metropolis-Hastings (MH) MCMC algorithm could be used. The MH approach could also be used when ψ has a large enough dimension to make rejection sampling inefficient, which can be the case for the heteroskedastic model when n is not very small.

Small area prediction:

If we knew the parameters ψ and $\boldsymbol{\beta}$, the minimum mean squared error predictor of $\mathbf{Y}_i$ would be, from standard multivariate normal results

$$E(\mathbf{Y}_i|\mathbf{y}, \psi, \boldsymbol{\beta}) = E(\mathbf{Y}_i|\mathbf{y}_i, \psi, \boldsymbol{\beta}) = \mathbf{X}_i\boldsymbol{\beta} + \Sigma_u(\psi)\Sigma_{y_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}). \tag{3}$$

In practice, we can substitute the MLEs, $\hat{\psi}$ and $\hat{\boldsymbol{\beta}}$, into (3) yielding

$$\hat{\mathbf{Y}}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} + \Sigma_u(\hat{\psi})\Sigma_{y_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}). \tag{4}$$

Note that Σ_{y_i} also depends on $\hat{\psi}$. For the Bayesian approach, (4) with ψ instead of $\hat{\psi}$ gives $E(\mathbf{Y}_i|\mathbf{y}, \psi)$. Given simulations of ψ from its posterior, $p(\psi|\mathbf{y})$, we can calculate (4) for each simulated ψ and then average the results over the simulations to approximate the posterior mean, $E(\mathbf{Y}_i|\mathbf{y})$, which is the Bayesian MMSE predictor.

The prediction error of $\hat{\mathbf{Y}}_i$ can be written

$$\mathbf{Y}_i - \hat{\mathbf{Y}}_i = (\mathbf{Y}_i - \tilde{\mathbf{Y}}_i) + (\tilde{\mathbf{Y}}_i - \hat{\mathbf{Y}}_i) \tag{5}$$

where we use the notation $\tilde{\mathbf{Y}}_i$ for $E(\mathbf{Y}_i|\mathbf{y}_i, \psi, \boldsymbol{\beta})$ given by (3). The two terms on the right-hand-side of (5) are orthogonal, so the prediction error variance of $\hat{\mathbf{Y}}_i$ is $\mathrm{Var}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i) = \mathrm{Var}(\mathbf{Y}_i - \tilde{\mathbf{Y}}_i)+\mathrm{Var}(\tilde{\mathbf{Y}}_i - \hat{\mathbf{Y}}_i)$. To derive this we rewrite (3) as

$$\begin{aligned} E(\mathbf{Y}_i|\mathbf{y}_i, \psi, \boldsymbol{\beta}) &= \mathbf{y}_i - E(\mathbf{e}_i|\mathbf{y}_i, \psi, \boldsymbol{\beta}) \\ &= \mathbf{y}_i - \Sigma_{e_i}\Sigma_{y_i}^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}) \end{aligned}$$

and similarly for (4), from which one can easily see that

$$\tilde{\mathbf{Y}}_i - \hat{\mathbf{Y}}_i = \Sigma_{e_i}\Sigma_{y_i}^{-1}\mathbf{X}_i(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

Also, from standard results, given ψ, $\text{Var}(\mathbf{Y}_i - \tilde{\mathbf{Y}}_i) = \Sigma_u(\psi) - \Sigma_u(\psi)\Sigma_{y_i}^{-1}\Sigma_u(\psi) = \Sigma_{e_i} - \Sigma_{e_i}\Sigma_{y_i}^{-1}\Sigma_{e_i}$. Therefore,

$$\text{Var}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i) = \left\{\Sigma_{e_i} - \Sigma_{e_i}\Sigma_{y_i}^{-1}\Sigma_{e_i}\right\} + \Sigma_{e_i}\Sigma_{y_i}^{-1}\mathbf{X}_i\text{Var}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\mathbf{X}_i'\Sigma_{y_i}^{-1}\Sigma_{e_i}. \quad (6)$$

Notice that this depends on ψ through Σ_{y_i} and $\text{Var}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$. This result accounts for error due to estimating $\boldsymbol{\beta}$, but not for error due to estimating ψ. If the number of observations is large compared to the number of elements of ψ, the error due to estimating ψ should be unimportant.

For the Bayesian approach, (6) gives the variance of $\mathbf{Y}_i$ conditional on both $\mathbf{y}$ and ψ, $\text{Var}(\mathbf{Y}_i|\mathbf{y}, \psi)$. Given simulations of ψ from its posterior, $p(\psi|\mathbf{y})$, we can approximate the posterior variance, $\text{Var}(\mathbf{Y}_i|\mathbf{y})$, using

$$\text{Var}(\mathbf{Y}_i|\mathbf{y}) = E_{\psi|\mathbf{y}}[\text{Var}(\mathbf{Y}_i|\mathbf{y}, \psi)] + \text{Var}_{\psi|\mathbf{y}}[E(\mathbf{Y}_i|\mathbf{y}, \psi)]$$

where $\text{Var}(\mathbf{Y}_i|\mathbf{y}, \psi)$ is given by (6), $E(\mathbf{Y}_i|\mathbf{y}, \psi)$ is given by (4) with ψ replacing $\hat{\psi}$, and $E_{\psi|\mathbf{y}}$ and $\text{Var}_{\psi|\mathbf{y}}$ denote the expectation and variance under the posterior $p(\psi|\mathbf{y})$. We approximate $E_{\psi|\mathbf{y}}[\text{Var}(\mathbf{Y}_i|\mathbf{y}, \psi)]$ by averaging the values of $\text{Var}(\mathbf{Y}_i|\mathbf{y}, \psi)$ computed for each simulated value of ψ, and we approximate $\text{Var}_{\psi|\mathbf{y}}[E(\mathbf{Y}_i|\mathbf{y}, \psi)]$ by taking the sample variance of $E(\mathbf{Y}_i|\mathbf{y}, \psi)$ across the simulations of ψ. If we use a normal approximation to $p(\psi|\mathbf{y})$ to obtain the simulations of ψ, these results could also be taken as an extension to the ML prediction results to obtain an improved $\text{Var}(\mathbf{Y}_i - \hat{\mathbf{Y}}_i)$ that also accounts for the error in estimating ψ.