Hierarchical Model Example: California Temperature Data

February 14, 2013

Recall the general form of a hierarchical model:

$$\begin{aligned} p(\eta, \theta|y) \quad \propto \quad & p(y|\eta, \theta) \times \quad \text{“Data model”} \\ & p(\eta|\theta) \times \quad \text{“Process model”} \\ & p(\theta) \quad \text{“Parameter model”} \end{aligned}$$

Define the vectors

$$\begin{aligned} y &= (y(s_1), \ldots, y(s_n))^T \quad \text{Observations} \\ \eta_{obs} &= (\eta(s_1), \ldots, \eta(s_n))^T \quad \text{Process at observation locations} \\ \eta_{pred} &= (\eta(s_1^*), \ldots, \eta(s_m^*))^T \quad \text{Process at prediction locations} \end{aligned}$$

1 Prediction

Our main goal is to approximate $p(\eta_{pred}|y)$. Note that

$$p(\eta_{pred}|y) = \int p(\eta_{pred}|\eta_{obs}, \theta, y) p(\eta_{obs}, \theta|y) d\eta_{obs} d\theta$$

What this means is that we can first generate samples from $p(\eta_{obs}, \theta|y)$, then to generate samples from $p(\eta_{pred}|y)$ we can sample from

$$p(\eta_{pred}|\eta_{obs}^{(i)}, \theta^{(i)}, y)$$

for each of the $i = 1, \ldots, B$ samples from $p(\eta_{obs}, \theta|y)$. So I'm going to save the prediction part for the end and focus for now only on $p(\eta_{obs}, \theta|y)$.

2 Model specification

Note that we can factor

$$p(\eta_{obs}, \theta, y) \quad = \quad p(y|\eta_{obs}, \theta)p(\eta_{obs}|\theta)p(\theta)$$

To do this I haven't made any independence assumptions; I've just used the definition of conditional probability. For the model we'll specify, $\theta = (\beta, \sigma^2, \rho, \tau^2)^T$. β, σ^2, and ρ are used in the process model, and τ^2 is used in the data model. So now we have

$$p(\eta_{obs}, \theta, y) = p(y|\eta_{obs}, \tau^2)p(\eta_{obs}|\beta, \sigma^2, \rho)p(\beta, \sigma^2, \rho)$$

The two assumptions here are that Y is independent of β, σ^2, and ρ given η_{obs} and τ^2, and that η_{obs} is independent of τ^2 given β, σ^2, and ρ. Now we'll specify each of the components in this equation.

2.1 Data Model

We take the observations to be conditionally independent of each other given η_{obs} and unknown variance τ^2, with

$$Y|\eta_{obs}, \tau^2 \sim MVN(\eta_{obs}, \tau^2 I_n)$$

2.2 Process Model

We take the process to have a Gaussian process distribution. This implies multivariate normal distributions for the process at any finite set of locations. In particular, we take

$$\eta|\beta, \sigma^2, \rho \sim GP(X(\cdot)^T\beta, \sigma^2 K(\cdot, \cdot; \rho))$$

where $K(s_i, s_j; \rho) = \exp\{-||s_i - s_j||/\rho\}$.

Now we can write down the finite-dimensional distributions. We'll write η_{obs} and η_{pred} jointly for later reference.

$$\begin{pmatrix} \eta_{obs} \\ \eta_{pred} \end{pmatrix} |\beta, \sigma^2, \rho \sim MVN\left(\begin{pmatrix} X \\ X_{pred} \end{pmatrix}\beta, \sigma^2 \begin{pmatrix} \Gamma(\rho) & \gamma(\rho) \\ \gamma(\rho)^T & \Gamma_{pred}(\rho) \end{pmatrix}\right)$$

where X is an $n \times p$ matrix whose i^{th} row is $X(s_i)^T$, X_{pred} is an mxp matrix whose i^{th} row is $X(s_i^*)^T$, and the correlation matrices $\Gamma(\rho)$, $\Gamma_{pred}(\rho)$, and $\gamma(\rho)$ are $n \times n$,

$m \times m$, and $n \times m$, respectively, with

$$\begin{aligned}
\Gamma(\rho)_{ij} &= \exp\{-||s_i - s_j||/\rho\} \\
\Gamma_{pred}(\rho)_{ij} &= \exp\{-||s_i^* - s_j^*||/\rho\} \\
\gamma(\rho)_{ij} &= \exp\{-||s_i - s_j^*||/\rho\}
\end{aligned}$$

For our MCMC to sample $\eta_{obs}, \theta|y$, we need only the marginal distribution

$$\eta_{obs}|\beta, \sigma^2, \rho \sim MVN(X\beta, \sigma^2\Gamma(\rho))$$

2.3 Parameter Model

Before seeing the data, we know very little any possible dependence of the parameters. So we take them to be independent with prior distributions

$$\begin{aligned}
\beta &\sim MVN(m_\beta, V_\beta) \\
\sigma^2 &\sim InverseGamma(a_{\sigma^2}, b_{\sigma^2}) \\
\tau^2 &\sim InverseGamma(a_{\tau^2}, b_{\tau^2}) \\
\rho &\sim Gamma(a_\rho, b_\rho)
\end{aligned}$$

Note that once we condition on the data, they will generally *not* be independent, but we may still be interested in looking at their marginal posterior distributions.

3 Full Conditional Distributions

Now we can derive the full conditional distributions. For each one, we'll use the full joint distribution

$$p(\eta_{obs}, \beta, \sigma^2, \rho, \tau^2, y) = p(y|\eta_{obs}, \tau^2)p(\eta_{obs}|\beta, \sigma^2, \rho)p(\beta)p(\sigma^2)p(\rho)p(\tau^2)$$

3.1 $\beta|Rest$

$$\begin{aligned}
p(\beta|Rest) &\propto p(\eta_{obs}|\beta, \sigma^2, \rho)p(\beta) \\
&\propto \exp\left\{-\frac{1}{2}(\eta_{obs} - X\beta)^T[\sigma^2\Gamma(\rho)]^{-1}(\eta_{obs} - X\beta)\right\} \times
\end{aligned}$$

$$
\begin{aligned}
& \exp\left\{-\frac{1}{2}(\beta - m_\beta)^T V_\beta^{-1}(\beta - m_\beta)\right\} \\
\propto \quad & \exp\left\{-\frac{1}{2}\left[\beta^T(X^T[\sigma^2\Gamma(\rho)]^{-1}X + V_\beta^{-1})\beta - 2\beta^T(X^T[\sigma^2\Gamma(\rho)]^{-1}\eta_{obs} + V_\beta^{-1}m_\beta)\right\}\right\}
\end{aligned}
$$

The last expression has the form $\exp\{-\frac{1}{2}\beta^T A\beta - 2\beta^T b\}$. We recognize this as the kernel of a $MVN(\tilde{m}_\beta, \tilde{V}_\beta)$ distribution, with

$$
\begin{aligned}
\tilde{V}_\beta &= A^{-1} = [X^T\Gamma(\rho)^{-1}X/\sigma^2 + V_\beta^{-1}]^{-1} \\
\tilde{m}_\beta &= \tilde{V}_\beta b = [X^T\Gamma(\rho)^{-1}X/\sigma^2 + V_\beta^{-1}]^{-1}[X^T\Gamma(\rho)^{-1}\eta_{obs}/\sigma^2 + V_\beta^{-1}m_\beta]
\end{aligned}
$$

3.2 $\sigma^2|Rest$

$$
\begin{aligned}
p(\sigma^2|Rest) &\propto p(\eta_{obs}|\beta, \sigma^2, \rho)p(\sigma^2) \\
&\propto (\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\eta_{obs} - X\beta)^T\Gamma(\rho)^{-1}(\eta_{obs} - X\beta)\right\} \times \\
&\quad\; (\sigma^2)^{-a_{\sigma^2}-1}\exp\{-b_{\sigma^2}/\sigma^2\} \\
&\propto (\sigma^2)^{-(a_{\sigma^2}+n/2)-1}\exp\left\{-\frac{1}{\sigma^2}[b_{\sigma^2} + (\eta_{obs} - X\beta)^T\Gamma(\rho)^{-1}(\eta_{obs} - X\beta)/2]\right\}
\end{aligned}
$$

which we recognize as the kernel of an $InverseGamma(\tilde{a}_{\sigma^2}, \tilde{b}_{\sigma^2})$ distribution, where $\tilde{a}_{\sigma^2} = a_{\sigma^2} + n/2$ and $\tilde{b}_{\sigma^2} = b_{\sigma^2} + (\eta_{obs} - X\beta)^T\Gamma(\rho)^{-1}(\eta_{obs} - X\beta)/2$.

3.3 $\tau^2|Rest$

$$
\begin{aligned}
p(\tau^2|Rest) &\propto p(y|\eta_{obs}, \tau^2)p(\tau^2) \\
&\propto (\tau^2)^{-n/2}\exp\left\{-\frac{1}{2\tau^2}(y - \eta_{obs})^T(y - \eta_{obs})\right\} \times \\
&\quad\; (\tau^2)^{-a_{\tau^2}-1}\exp\{-b_{\tau^2}/\tau^2\} \\
&\propto (\tau^2)^{-(a_{\tau^2}+n/2)-1}\exp\left\{-\frac{1}{\tau^2}[b_{\tau^2} + (y - \eta_{obs})^T(y - \eta_{obs})/2]\right\}
\end{aligned}
$$

which we recognize as the kernel of an $InverseGamma(\tilde{a}_{\tau^2}, \tilde{b}_{\tau^2})$ distribution, where $\tilde{a}_{\tau^2} = a_{\tau^2} + n/2$ and $\tilde{b}_{\tau^2} = b_{\tau^2} + (y - \eta_{obs})^T(y - \eta_{obs})/2$.

3.4 $\eta_{obs}|Rest$

$$\begin{aligned}
p(\eta_{obs}|Rest) &\propto p(y|\eta_{obs},\tau^2)p(\eta_{obs}|\beta,\sigma^2,\rho) \\
&\propto \exp\left\{\frac{1}{2\tau^2}(y-\eta_{obs})^T I_n(y-\eta_{obs})\right\}\exp\left\{\frac{1}{2\sigma^2}(\eta_{obs}-X\beta)^T\Gamma(\rho)^{-1}(\eta_{obs}-X\beta)\right\} \\
&\propto \exp\left\{\eta_{obs}^T(\frac{1}{\tau^2}I_n+\frac{1}{\sigma^2}\Gamma(\rho)^{-1})\eta_{obs}-2\eta_{obs}^T(\frac{1}{\tau^2}y+\frac{1}{\sigma^2}\Gamma(\rho)^{-1}X\beta))\right\}
\end{aligned}$$

which we recognize as the kernel of a $MVN(\tilde{m}_{eta_{obs}},\tilde{V}_{eta_{obs}})$ distribution, with

$$\begin{aligned}
\tilde{V}_{eta_{obs}} &= [\frac{1}{\tau^2}I_n+\frac{1}{\sigma^2}\Gamma(\rho)^{-1}]^{-1} \\
\tilde{m}_{eta_{obs}} &= [\frac{1}{\tau^2}I_n+\frac{1}{\sigma^2}\Gamma(\rho)^{-1}]^{-1}[\frac{1}{\tau^2}y+\frac{1}{\sigma^2}\Gamma(\rho)^{-1}X\beta]
\end{aligned}$$

3.5 $\rho|Rest$

$$p(\rho|Rest)\propto p(\eta_{obs}|\beta,\sigma^2,\rho)p(\rho)$$

but there is no way to further simplify the form of this distribution. We'll use a Metropolis Hastings step.

Aside: Another approach, which is what Diggle and Ribeiro do, is to take ρ to take only a finite number of possible values a priori. That is, specify values $\rho_1,\ldots,\rho_k$, with $P(\rho=\rho_k)=\pi_k$ and $\sum_{j=1}^k\pi_j=1$. We can compactly write the PDF as

$$p(\rho)=\sum_{j=1}^k\pi_j I\{\rho=\rho_j\}$$

Then the full conditional

$$\begin{aligned}
p(\rho|Rest)\propto p(\eta_{obs}|\beta,\sigma^2,\rho)p(\rho) &= \sum_{j=1}^k p(\eta_{obs}|\beta,\sigma^2,\rho_j)\pi_j I\{\rho=\rho_j\} \\
&= \sum_{j=1}^k w_j I\{\rho=\rho_j\}
\end{aligned}$$

where $w_j=p(\eta_{obs}|\beta,\sigma^2,\rho_j)\pi_j$. This means that $p(\rho|Rest)$ is another discrete distribution, with $P(\rho=\rho_j|Rest)=w_j/\sum_{j=1}^k w_j$ (since these probabilities must sum to 1).

4 Prediction

To do prediction, we need $\eta_{pred}|\eta_{obs}, \beta, \sigma^2, \rho, y$, which we can get using the multivariate normal conditioning equations (page 35 of the class notes).

$$\eta_{pred}|\eta_{obs}, \beta, \sigma^2, \rho, y \sim MVN(m_{pred}, V_{pred})$$

where

$$\begin{aligned} m_{pred} &= X_{pred}\beta + \gamma(\rho)^T \Gamma(\rho)^{-1}(y - X\beta) \\ V_{pred} &= \sigma^2[\Gamma_{pred}(\rho) - \gamma(\rho)^T \Gamma(\rho)^{-1}\gamma(\rho)] \end{aligned}$$