Clustering Model

November 3, 2013

1 The Model

Consider a set of potential signals $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J\}$ where these signals represent a m dimensional vectorized spatial field. For a given year t we have a vector $\boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{y}_t^o \\ \boldsymbol{y}_t^u \end{pmatrix}$ where $\boldsymbol{y}_t^o$ is a $n_t \times 1$ vector of observed values at the spatial random field at time t with $n_t << m$ and $\boldsymbol{y}_t^u$ is a $m - n_t \times 1$ vector of unobserved values at the spatial random field at time t. We can write our model statement as

$$\boldsymbol{y}_t = \begin{pmatrix} \boldsymbol{y}_t^o \\ \boldsymbol{y}_t^u \end{pmatrix} = \begin{pmatrix} \boldsymbol{K}_t^o \\ \boldsymbol{K}_t^u \end{pmatrix} \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t = \boldsymbol{K}_t\boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{K}_t$ is a matrix that selects which observations are observed or unobserved, $\boldsymbol{X}$ is a matrix where the j^{th} column is the vector $\boldsymbol{x}_j$, $\boldsymbol{\beta}_t$ is a coefficient vector, and $\boldsymbol{\epsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$. The matrix $\boldsymbol{X}$ can be thought of as a "palette" of potential signals (either endogenous to y_t, exogenous to y_t, or a mixture of endogenous and exogenous signals) that are assumed to be representative of the states of the system. For our purposes, $\boldsymbol{X}$ is the 116 years of PRISM data.

2 Full Dimensional Model Statement

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{K}_t\boldsymbol{X}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \\
\boldsymbol{\alpha}_t &\sim N(\boldsymbol{\mu}_\alpha, \boldsymbol{\Sigma}_\alpha) \\
\boldsymbol{\Sigma}_\alpha &\sim \sigma_\alpha^2 \boldsymbol{I}_{\boldsymbol{\tau}\times\boldsymbol{\tau}} \\
\boldsymbol{\mu}_\beta &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\
\sigma_\alpha^2 &\sim IG(\alpha_\alpha, \beta_\alpha) \\
\sigma_\epsilon^2 &\sim IG(\alpha_\epsilon, \beta_\epsilon)
\end{aligned}$$

where $\boldsymbol{X}$ is a matrix of 116 years of PRISM data, $\boldsymbol{y}_t$ is the fort data at year t, $\boldsymbol{K}_t$ is the locator matrix that ties the observation location to the full grid. Let T represent the number of years of fort data, n_t the number of observations for each year $t = 1, \ldots, T$, N is the number of cells on the spatial grid and τ is the number of years of PRISM data.

3 Reduced Dimensional Model Statement

$$\begin{aligned}
\boldsymbol{y}_t &= \boldsymbol{K}_t\boldsymbol{X}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \\
&= \boldsymbol{K}_t\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \\
&= \boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t
\end{aligned}$$

where $\boldsymbol{UDV}^T$ is the singular value decomposition of $\boldsymbol{X}$, $\tilde{\boldsymbol{X}} = \boldsymbol{XU}$ is the data matrix rotated by the left singular vectors, and $\boldsymbol{\beta} = \boldsymbol{DV}^T\boldsymbol{\alpha}$ is the vector of parameters rescaled by the diagonal matrix $\boldsymbol{D}$ of singular values and rotated by the right singular vector $\boldsymbol{V}^T$.

Since $\boldsymbol{\beta} = \boldsymbol{DV}^T\boldsymbol{\alpha}$ is a linear transformation of $\boldsymbol{\alpha}$

$$\boldsymbol{\beta} \sim N(\boldsymbol{DV}^T\boldsymbol{\mu}_\alpha, \boldsymbol{DV}^T\sigma_\alpha^2\boldsymbol{VD}^T) = N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$$

where $\boldsymbol{DV}^T\boldsymbol{\mu}_\alpha$ is a linear transformation of the mean vector and $\boldsymbol{DV}^T\sigma_\alpha^2\boldsymbol{VD}^T = \sigma_\alpha^2\boldsymbol{DV}^T\boldsymbol{VD}^T = \sigma_\alpha^2\boldsymbol{DD}^T = \sigma_\alpha^2\boldsymbol{\Lambda} = \sigma_\beta^2\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is the matrix of squared singular values on the diagonal.

4 Issues

1. There needs to be shrinkage towards the overall historic mean temperature surface for the years where the number of observations n_t in $\boldsymbol{y}_t$ is small. This can be accomplished using a penalty such as LASSO, by the inclusion of a strong prior on β_t, or using a k-fold cross validation on the mean square prediction error.
2. There is likely to be very high multicollinearity in the $\boldsymbol{X}$ matrix. The use of a singular value decomposition de-correlates $\boldsymbol{X}$ and the discarding of higher order singular value terms of $\tilde{\boldsymbol{X}}$ increases computational efficiency and reduces noise in the temperature reconstructions.

5 Posterior

$$\prod_{t=1}^{T}[\boldsymbol{\beta}_t, \boldsymbol{\mu}_t, \sigma_\beta^2, \sigma_\epsilon^2|\boldsymbol{y}_t] = \prod_{t=1}^{T}[\boldsymbol{y}_t|\boldsymbol{\beta}_t, \sigma_\epsilon^2][\boldsymbol{\beta}_t|\boldsymbol{\mu}_\beta, \sigma_\beta^2][\sigma_\epsilon^2\boldsymbol{I}_{n_t\times n_t}][\sigma_\beta^2][\boldsymbol{\mu}_\beta]$$

6 Full Conditionals

6.1 $\boldsymbol{\beta}_t$

For $t = 1, \ldots, T$,

$$\begin{aligned}
[\boldsymbol{\beta}_t|\cdot] &\propto [\boldsymbol{y}_t|\boldsymbol{\beta}_t, \sigma_\epsilon^2][\boldsymbol{\beta}_t|\boldsymbol{\mu}_\beta, \sigma_\beta^2] \\
&\propto e^{-\frac{1}{2}(\boldsymbol{y}_t - \boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}(\boldsymbol{y}_t - \boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)}e^{-\frac{1}{2}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_\beta)^T(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_\beta)} \\
&\propto e^{-\frac{1}{2}\{\boldsymbol{\beta}_t^T(\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}} + (\sigma_\beta^2\boldsymbol{\Lambda})^{-1})\boldsymbol{\beta}_t - 2\boldsymbol{\beta}_t^T(\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}_t + (\sigma_\beta^2\boldsymbol{\Lambda})^{-1}\boldsymbol{\mu}_\beta)\}}
\end{aligned}$$

which is Normal with mean $= (\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}} + (\sigma_\beta^2\boldsymbol{\Lambda})^{-1})^{-1}(\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}_t + (\sigma_\beta^2\boldsymbol{\Lambda})^{-1}\boldsymbol{\mu}_\beta)$ and variance $= (\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}} + (\sigma_\beta^2\boldsymbol{\Lambda})^{-1})^{-1}$

6.2 $\boldsymbol{\mu}_\beta$

$$\begin{aligned}
[\boldsymbol{\mu}_\beta|\cdot] &\propto \prod_{t=1}^{T}[\boldsymbol{\beta}_t|\boldsymbol{\mu}_\beta, \sigma_\beta^2][\boldsymbol{\mu}_\beta] \\
&\propto e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_\beta)^T(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\mu}_\beta)}e^{-\frac{1}{2}(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_0)^T\Sigma_0^{-1}(\boldsymbol{\mu}_\beta - \boldsymbol{\mu}_0)} \\
&\propto e^{-\frac{1}{2}(\boldsymbol{\mu}_\beta^T(T*(\sigma_\beta^2\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})\boldsymbol{\mu}_\beta - 2\boldsymbol{\mu}_\beta^T(\sum_{t=1}^{T}(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}\boldsymbol{\beta}_t + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0))}
\end{aligned}$$

which is multivariate normal with mean $(T*(\sigma_\beta^2\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}(\sum_{t=1}^{T}(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}\boldsymbol{\beta}_t + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$ and variance $(T*(\sigma_\beta^2\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1}$

6.3 σ_β^2

$$
\begin{aligned}
[\sigma_\beta^2|\cdot] &\propto \prod_{t=1}^{T}[\boldsymbol{\beta}_t|\boldsymbol{\mu}_\beta,\sigma_\beta^2][\sigma_\beta^2] \\
&\propto (\prod_{t=1}^{T}|\sigma_\beta^2\boldsymbol{\Lambda}|^{-\frac{1}{2}})e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)^T(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)}(\sigma_\beta^2)^{-(\alpha_\beta+1)}e^{-\frac{\beta_\beta}{\sigma_\beta^2}} \\
&\propto (\sigma_\beta^2)^{-(\alpha_\beta+\frac{T*|\boldsymbol{\Lambda}|}{2}+1)}e^{-\frac{1}{\sigma_\beta^2}(\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)^T(\boldsymbol{\Lambda})^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)+\beta_\beta)}
\end{aligned}
$$

which is IG$(\alpha_\beta+\frac{T*|\boldsymbol{\Lambda}|}{2},\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)^T(\boldsymbol{\Lambda})^{-1}(\boldsymbol{\beta}_t-\boldsymbol{\mu}_t)+\beta_\beta)$

6.4 σ_ϵ^2

$$
\begin{aligned}
[\sigma_\epsilon^2|\cdot] &\propto \prod_{t=1}^{T}[\boldsymbol{y}_t|\boldsymbol{\beta}_t,\sigma_\epsilon^2][\sigma_\epsilon^2] \\
&\propto (\prod_{t=1}^{T}|\sigma_\epsilon^2 I_{n_t\times n_t}|^{-\frac{1}{2}})e^{-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)^T(\sigma_\epsilon^2 I_{n_t\times n_t})^{-1}(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)}(\sigma_\epsilon^2)^{-\alpha_\epsilon+1}e^{-\frac{\beta_\epsilon}{\sigma_\epsilon^2}} \\
&\propto (\sigma_\epsilon^2)^{-(\frac{1}{2}\sum_{t=1}^{T}n_t+\alpha_\epsilon+1)}e^{-\frac{1}{\sigma_\epsilon^2}(\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)^T(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)+\beta_\epsilon)}
\end{aligned}
$$

which is IG$(\frac{1}{2}\sum_{t=1}^{T}n_t+\alpha_\epsilon,\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)^T(\boldsymbol{y}_t-\boldsymbol{K}_t\tilde{\boldsymbol{X}}\boldsymbol{\beta}_t)+\beta_\epsilon)$

7 Future Work

1. Work on implementing a k-fold cross validation for the selection of the variance component σ_β^2 based on mean square prediction error. This should reduce variability in years with few fort temperature observations.

2. Consider a measurement error term σ_ϵ^2 that varies in time (or even by site). It seems likely that there are inconsistencies in measurement of the fort data temperature that could result in prediction error.

8 Shrinkage

8.1 $\boldsymbol{\beta}_t$

Consider what happens to the full conditional distribution of β_t when $\sigma_\beta^2\to 0$.

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{\beta}_t) &= (\tilde{\boldsymbol{X}}\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}}^T+(\sigma_\beta^2\boldsymbol{\Lambda})^{-1})^{-1} \\
&= (\tilde{\boldsymbol{X}}\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}}^T+1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1})^{-1} \\
&\approx (1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1})^{-1} \\
&= \sigma_\beta^2\boldsymbol{\Lambda} \\
&\to 0
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{\beta}_t) &= (\tilde{\boldsymbol{X}}\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}}^T+(\sigma_\beta^2\boldsymbol{\Lambda})^{-1})^{-1}(\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}_t+(\sigma_\beta^2\boldsymbol{\Lambda})^{-1}\boldsymbol{\mu}_\beta) \\
&= (\tilde{\boldsymbol{X}}\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{K}_t\tilde{\boldsymbol{X}}^T+1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1})^{-1}(\tilde{\boldsymbol{X}}^T\boldsymbol{K}_t^T(\sigma_\epsilon^2\boldsymbol{I})^{-1}\boldsymbol{y}_t+1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1}\boldsymbol{\mu}_\beta) \\
&\approx (1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1})^{-1}(1/\sigma_\beta^2(\boldsymbol{\Lambda})^{-1}\boldsymbol{\mu}_\beta) \\
&= \boldsymbol{\mu}_\beta
\end{aligned}
$$

8.2 $\boldsymbol{\mu}_\beta$

As $\sigma_\beta^2 \to 0$

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{\mu}_\beta) &= (T * (\sigma_\beta^2 \boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} \\
&= (T * 1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} \\
&\approx (T * 1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1})^{-1} \\
&= \sigma_\beta^2 / T * \boldsymbol{\Lambda} \\
&\to 0
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\mu}_\beta) &= (T * (\sigma_\beta^2 \boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} (\sum_{t=1}^{T} (\sigma_\beta^2 \boldsymbol{\Lambda})^{-1} \boldsymbol{\beta}_t + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0) \\
&= (T * 1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1} + \boldsymbol{\Sigma}_0^{-1})^{-1} (1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1} \sum_{t=1}^{T} \boldsymbol{\beta}_t + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0) \\
&\approx (T * 1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1})^{-1} (1/\sigma_\beta^2 (\boldsymbol{\Lambda})^{-1} \sum_{t=1}^{T} \boldsymbol{\beta}_t) \\
&= \sum_{t=1}^{T} \boldsymbol{\beta}_t / T
\end{aligned}$$