Review of the Spatial Dirichlet Process [1]

UCSC AMS 241 Course Project
9 December 2015
Arthur Lui

This project is a review of the spatial Dirichlet process (SDP) developed by Gelfand et al. (2005). I will first discuss how to model using the SDP, then examine properties of the model through a data analysis.

1 Spatial Dirichlet Process Modeling

To denote realizations from point-referenced spatial data, we use $\{y(s) : s \in S\}, S \subset R^d$, where d is the dimension of S. We observe data only at a subset of all possible points in S, $\mathbf{s}^{(n)} = (s_1, ..., s_n)$. Typically, this type of data is modeled by a Gaussian process (GP). However, the assumption that the data arises from a GP is often a restriction. we may want to allow deviation from a Gaussian random field. An SDP prior can be put on the random field and have a Gaussian process as the baseline distribution. Required for the model are replicates at each point. That is we need the full dataset to consist of a collection of vectors $\mathbf{y}_t = (y(s_1), ..., y(s_n))$, $t = 1, , , .T$. Note that the points s_i can be a pair of latitudes and longitudes.

We can construct the model as follows:

$$
\begin{array}{rcllll}
\mathbf{y}_t & | & \boldsymbol{\theta}_t, \beta, \tau^2 & \overset{ind.}{\sim} & \text{N}_n(\boldsymbol{\theta}_t + \mathbf{1_n}\beta,\ \tau^2\mathbf{I}_n), & {\scriptstyle t=1,\dots,T} \\
\boldsymbol{\theta}_t & | & G^{(n)} & \overset{i.i.d.}{\sim} & G^{(n)}, & {\scriptstyle t=1,\dots,T} \\
G^{(n)} & | & \alpha, \sigma^2, \phi & \sim & \text{DP}(\ \alpha, G_0^{(n)}\ \text{N}_n(\mathbf{0}_n, \sigma^2 H_n(\phi))\ ) & \\
 & & & & & \\
 & & \beta, \tau^2 & \sim & \text{N}(m, s^2) \times \text{IGamma}(a_{\tau^2} = 2, b_{\tau^2}) & \\
 & & \alpha & \sim & \text{Gamma}(a_\alpha, b_\alpha) & \\
 & & \sigma^2 & \sim & \text{IGamma}(b_{\sigma^2} = 2, b_{\sigma^2}) & \\
 & & \phi & \sim & \text{Uniform}(0, b_\phi) &
\end{array}
$$

where $H_n(\phi)$ is a covariance function, for example, the exponential covariance function with decay parameter ϕ. (i.e. $(H_n(\phi))_{ij} = \exp\{-\phi\ \|s_i - s_j\|\}$.)

Here, $\boldsymbol{\theta}_t$ are the location-specific mean deviations from a grand mean β across the n spatial locations. Notice that clustering can result among the $\boldsymbol{\theta}_t$'s. This may be useful when our θ_t's are indexed by time and we want to learn how the observations are clustered in time. Also notice that if we replaced the prior for the $\boldsymbol{\theta}$ the baseline distribution used, we get a Gaussian process. And as $\alpha \to \infty$, $\boldsymbol{\theta}_t$ become i.i.d. $G_0^{(n)}$ conditional on the hyperparameters. We assume $\mathbf{y}_t$ to be independent and multivariate normal with no correlation. That is, the covariance is $\tau^2\mathbf{I}_n$. The $\mathbf{y}_t$'s are simply modeled as a mixture of multivariate normals.

[1] `https://github.com/luiarthur/bnp_hw/project/bnp_spatialDP`

1.1 Prior Specification

The overall mean β modeled with a normal prior. In modeling maximum temperatures, it is appropriate to use a reasonably informative prior. In the following data analysis, a prior mean of 30 and a prior variance of 5 were used. Gelfand et al. (2005) suggests for priors for σ^2 and τ^2 Inverse-Gamma priors with parameters $(2, b_{\sigma^2}$ and $(2, b_{\tau^2})$ respectively. This corresponds to a prior mean of b_{σ^2} and b_{τ^2} and infinite variance for the two parameters. For the data analysis below, I used fixed variances as the model could not learn their values. With a T of only 20, it appears that posterior learning could not occur. Gelfand et al. (2005) states that "practical experience... suggests that there is posterior learning for $[\alpha]$ when sampling sizes are moderate to large". Again ϕ, which determines the rate of decay in the exponential decay function, was fixed. The prior for α chosen such that the prior mean and variance were 1 and and 100 respectively.

Note that all the above priors except that for ϕ are conjugate. And yields the following complete conditionals that can be used in a Gibbs sampler.

$$\begin{array}{rcllcl}
\beta & \mid & \mathbf{y}, \boldsymbol{\theta}, \tau^2 & \sim & \mathrm{N}(\frac{\tau^2 m + s^2 \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{it} - \theta_{it})}{\tau^2 + Tns^2}, \frac{s^2\tau^2}{\tau^2 + Tns^2}) \\
\tau^2 & \mid & \mathbf{y}, \boldsymbol{\theta}, \beta & \sim & \mathrm{IG}(a_{\tau^2} + \frac{nT}{2}, b_{\tau^2} + \frac{\sum_{t=1}^{T} (\boldsymbol{\mu}_t - \beta \mathbf{1}_n)'(\boldsymbol{\mu}_t - \beta \mathbf{1}_n)}{2}) \\
\sigma^2 & \mid & \boldsymbol{\theta}^*, T^*, \mathbf{y}, \sigma^2 & \sim & \mathrm{IG}(a_{\sigma^2} + \frac{nT^*}{2}, b_{\sigma^2} + \frac{\sum_{t=1}^{T^*} \boldsymbol{\theta}_t^{*\prime} H_n^{-1}(\phi) \boldsymbol{\theta}_t^*}{2}) \\
p(\phi & \mid & \boldsymbol{\theta}^*, T^*, \mathbf{y}, \sigma^2) & \propto & [\phi] |H_n(\phi)|^{-T^*/2} \exp\left(- \frac{\sum_{t=1}^{T^*} \boldsymbol{\theta}_t^{*\prime} H_n^{-1}(\phi) \boldsymbol{\theta}_t^*}{2\sigma^2} \right) \\
\eta & \mid & \alpha, \mathbf{y} & \sim & \mathrm{Beta}(\alpha + 1, T) \\
p(\alpha & \mid & T^*, \mathbf{y}) & = & (\epsilon)\ \gamma(\alpha | a_\alpha + T^*, b_\alpha - \log(\eta)) + \\
 & & & & (1 - \epsilon)\ \gamma(\alpha | a_\alpha + T^* - 1, b_\alpha - \log(\eta))
\end{array}$$

where $\boldsymbol{\mu}_t = \mathbf{y_t} - \boldsymbol{\theta}_\mathbf{t}$, η is an auxiliary variable introduced to make the prior for α conjugate, γ is the gamma density function with the mean and rate parameterization, and $\epsilon = \dfrac{a_\alpha + T^* - 1}{n(b_\alpha - \log(\eta)) + a_\alpha + T^* - 1}$.

As $\boldsymbol{\theta}_t$ has prior conjugacy, we can use the algorithm provided by Escobar and West (1995)

$$\begin{array}{rclcl}
\boldsymbol{\theta}_t & \mid & y_t, \beta, \tau^2, \sigma^2, \phi & \sim & \mathrm{N}_n(\tau^{-2} \boldsymbol{\Lambda}(\mathbf{y_t} - \mathbf{1_n}\beta), \boldsymbol{\Lambda}) \\
 & & q0 & = & |\Lambda|^{1/2} \\
 & & & & \times \exp\left\{ \frac{(\mathbf{y_t} - \mathbf{1_n}\beta)'(\mathbf{I_n} - \tau^{-2}\boldsymbol{\Lambda})(\mathbf{y_t} - \mathbf{1_n}\beta)}{2\tau^2} \right\} \\
 & & & & \times [(2\pi\tau^2\sigma^2)^{n/2} |H_n(\phi)|^{1/2}]^{-1}
\end{array}$$

where $\boldsymbol{\Lambda} = [\tau^{-2}\mathbf{I}_n + \sigma^{-2} H_n^{-1}(\phi)]^{-1}$.

1.2 Prediction at Unobserved Locations

Often, spatial modeling is used for prediction at unobserved locations (kriging). If we were able to observe the replicates at new locations, we get the following full Bayesian model

$$\prod_{t=1}^{T} [\mathbf{y}_t | \boldsymbol{\theta}_t, \beta, \tau^2] \prod_{t=1}^{T} [\tilde{\mathbf{y}}_t | \tilde{\boldsymbol{\theta}}_t, \beta, \tau^2] \prod_{t=1}^{T} [(\boldsymbol{\theta}_t, \tilde{\boldsymbol{\theta}}_t) | G^{(n+m)}][G^{(n+m)} | \alpha, \sigma^2, \phi][\tau^2][\alpha][\sigma^2][\phi]$$

where $\tilde{\mathbf{y}}_t$ and $\tilde{\boldsymbol{\theta}}_t$ are the **unobserved** replicate and it's location specific mean at new locations $(y_t(\tilde{s_1}),...,y_t(\tilde{s_m}))$, for $t=1,...,T$. The second term can be integrated from the multivariate normal likelihood (this is done by simply taking the corresponding terms in our likelihood away from the model). Also, after marginalizing over $G^{(n+m)}$ we obtain the following expression

$$\left(\prod_{t=1}^{T}[\mathbf{y}_t|\boldsymbol{\theta}_t,\beta,\tau^2]\right)[(\theta_1,\tilde{\theta_1}),...,(\theta_T,\tilde{\theta_T})|\alpha,\sigma^2,\phi][\tau^2][\alpha][\sigma^2][\phi].$$

This expression can be further rewritten as

$$\left(\prod_{t=1}^{T}[\mathbf{y}_t|\boldsymbol{\theta}_t,\beta,\tau^2]\right)[(\theta^*,\tilde{\theta^*})|T^*,\sigma^2,\phi][\mathbf{w},T^*|\alpha,\sigma^2,\phi][\tau^2][\alpha][\sigma^2][\phi],$$

where the $(\theta^*,\tilde{\theta^*})$ are the unique $\boldsymbol{\theta}_t$ which arise i.i.d. from $G_0^{(n+m)}$. That is $[(\theta^*,\tilde{\theta^*})|T^*,\sigma^2,\phi] = \prod_{j=1}^{T^*}\mathrm{N}_{n+m}(\theta_j^*,\tilde{\theta_j^*}\mid\mathbf{0}_{n+m},\sigma^2H_{n+m}(\phi))$. Finally, the model can be expressed as

$$\left(\prod_{t=1}^{T}[\mathbf{y}_t|\boldsymbol{\theta}_t,\beta,\tau^2]\right)\left(\prod_{t=1}^{T^*}[\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi]\right)\left(\prod_{t=1}^{T^*}\mathrm{N}_n(\theta_j^*\mid\mathbf{0}_n,\sigma^2H_n(\phi))\right)[\mathbf{w},T^*|\alpha,\sigma^2,\phi][\tau^2][\alpha][\sigma^2][\phi],$$

or equivalently

$$\left(\prod_{t=1}^{T^*}[\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi]\right)\left(\prod_{t=1}^{T}[\mathbf{y}_t|\boldsymbol{\theta}_t,\beta,\tau^2]\right)\left(\prod_{t=1}^{T^*}\mathrm{N}_n(\theta_j^*\mid\mathbf{0}_n,\sigma^2H_n(\phi))\right)[\mathbf{w},T^*|\alpha,\sigma^2,\phi][\tau^2][\alpha][\sigma^2][\phi].$$

Based on the expression above, the joint posterior $[(\boldsymbol{\theta}^*,\tilde{\boldsymbol{\theta}^*}),\mathbf{w},T^*,\beta,\tau^2,\alpha,\sigma^2,\phi|\text{data}]$ can be decomposed into

$$\left(\prod_{t=1}^{T^*}[\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi]\right)[\ \boldsymbol{\theta}^*,\mathbf{w},T^*,\beta,\tau^2,\alpha,\sigma^2,\phi|\text{data}\ ].$$

The second term can be is the joint posterior in the SDP model. So, we can employ the following algorithm to the MCMC outputs to obtain posterior draws for $\prod_{t=1}^{T^*}[\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi]$: for each $b=1,...,B$, sample from $[\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi]$ for each $j=1,...,T^*$. Notice that

$$\tilde{\theta_j^*}|\theta_j^*,\sigma^2,\phi\sim\mathrm{N}_m(\mathbf{0}_m+\mathbf{S}_{12}\mathbf{S}_{22}^{-1}(\theta_j-\mathbf{0}_n),\quad\mathbf{S}_{11}-\mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21})$$

where $\mathbf{S}=\sigma^2H_{n+m}(\phi)$, and the subsequent submatrices are defined as follows: $\mathbf{S}_{11}=\sigma^2H_{1:m,\ 1:m}(\phi)$, $\mathbf{S}_{12}=\sigma^2H_{1:m,\ m+1:m+n}(\phi)$, $\mathbf{S}_{22}=\sigma^2H_{m+1:m+n,\ m+1:m+n}(\phi)$, $\mathbf{S}_{21}=\sigma^2H_{m+1:m+n,\ 1:m}(\phi)$.
To obtain posterior predictives at the new locations, we further draw values for $(\tilde{\mathbf{y}_{0b}},\mathbf{y}_{0b})$ from $\mathrm{N}_{m+n}(\beta_b\mathbf{1}_{m+n}+(\tilde{\theta}_{0b},\theta_{0b}),\tau^2\mathbf{I}_{m+n})$.

2 Data Analysis

The data used for this study was gathered from The Atmospheric Science Data Center at NASA. The script used to parse this dataset can be found at the referenced website [2]. The data consists

[2] https://github.com/luiarthur/bnp_hw/tree/master/project/bnp_spatialDP/data/retrieve

of daily averaged maximum temperatures at 10 meters for every July from 1985 to 2004, at 100 locations. (There was available every month average within the years and for more locations, but to save on computation, only a subset of the data was retrieved and used.) Figure 1 shows the average maximum daily temperatures at 100 locations in July 1989, with hotter regions in dark red and cooler regions in dark blue. It can be seen and expected that in the northern regions and along the coast, temperatures are lower; and in the southern and inland regions, temperatures are higher. We fit the mentioned model to our data using the suggested prior specification (with the exception of fixing the variances).

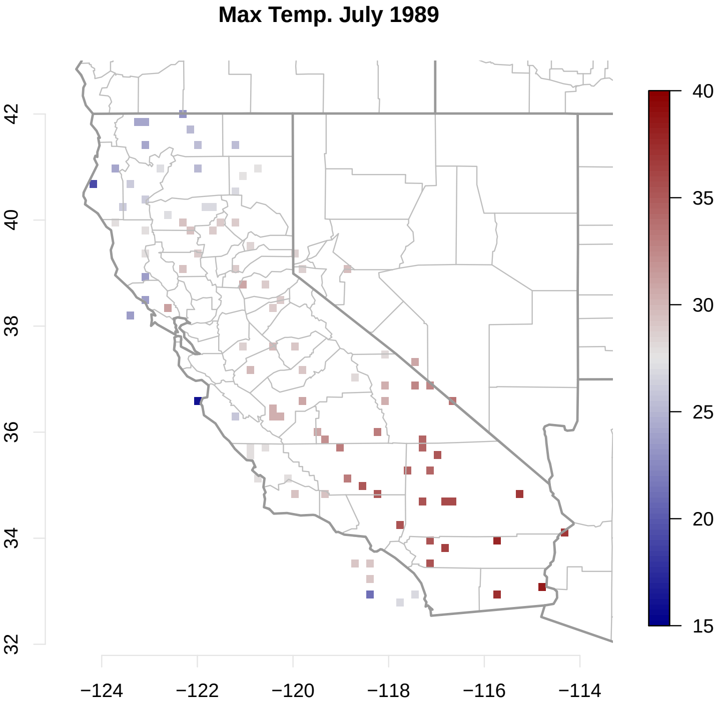

Figure 1: Average maximum daily temperatures at various locations in the state of California in 1989. Warmer areas are dark red and cooler locations are dark blue.

We plot the mean of the data (mean temperatures at each location) side by side with the posterior predictive mean in Figure 2. We see that the model smoothly interpolates the temperatures between locations of known temperature. The model borrows information from closer locations and less information from grid points far away. Far from the locations with data, the temperatures return the the mean β.

The posterior predictive variance in Figure 3. The variance in the posterior predictive is large (about 20 everywhere). I suspect this is due to the small amount of data ($T = 20$). With more data, which is available this can be remedied. It should be noted that in $C++$ fitting the model with 2500 MCMC iterations took 3 minutes. Fitting the model with more data can be easily accommodated and quick. In predicting at new locations, matrix inversions can take a long time

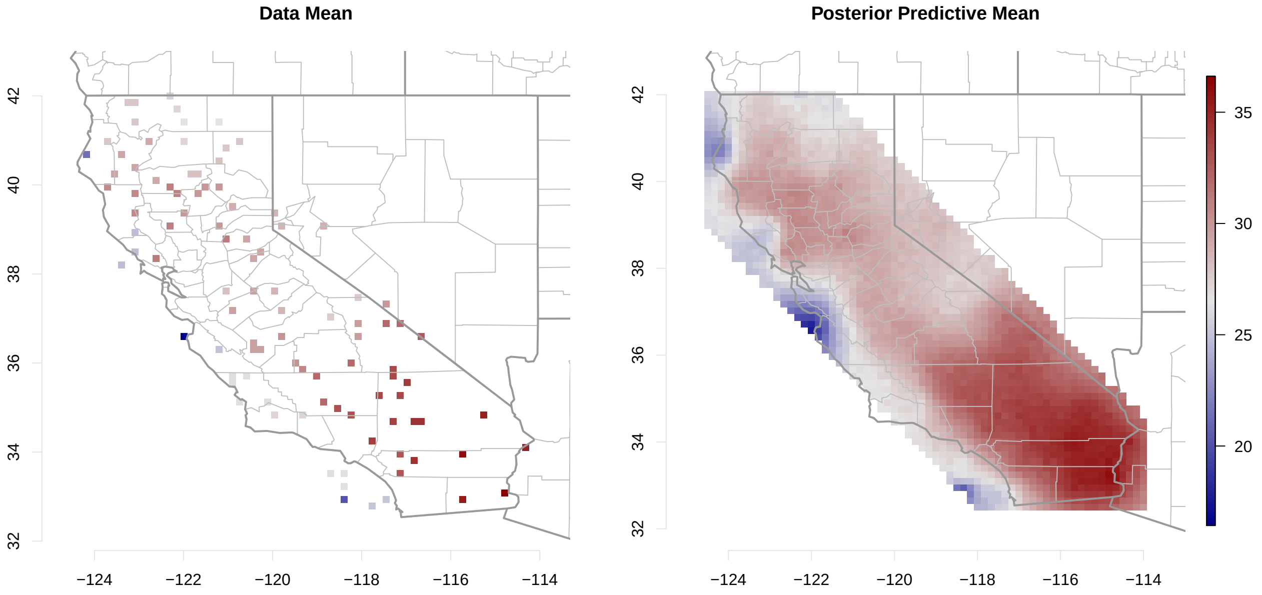

Figure 2: Left: Average of the average maximum daily temperatures at recorded locations in California. Right: Predicted average maximum daily temperatures at various locations in the state of California for a new year. Warmer areas are dark red and cooler locations are dark blue.

with large matrices, but this can be sped up by parallelization on larger systems.

3 Conclusions

The spatial Dirichlet process provides a flexible framework for modeling spatial data when stationarity and Gaussianity is not desireable. Clustering can be induced (though not discussed here). Posterior predictive for new locations can be done easily *after* fitting the model with only the locations in the data. So, predictions at new locations can be done in parallel after fitting the model. It is worth noting that further information can be harvested from the data by modeling the temporal structure. This topic has been explored using the SDP by Kottas et al. (2008).

References

Escobar, M. D., and West, M. (1995), “Bayesian density estimation and inference using mixtures,” *Journal of the american statistical association*, 90, 577–588.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), “Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing,” *Journal of the American Statistical Association*, 100, 1021–1035.

Kottas, A., Duan, J. A., and Gelfand, A. E. (2008), “Modeling disease incidence data with spatial and spatio temporal Dirichlet process mixtures,” *Biometrical Journal*, 50, 29–42.

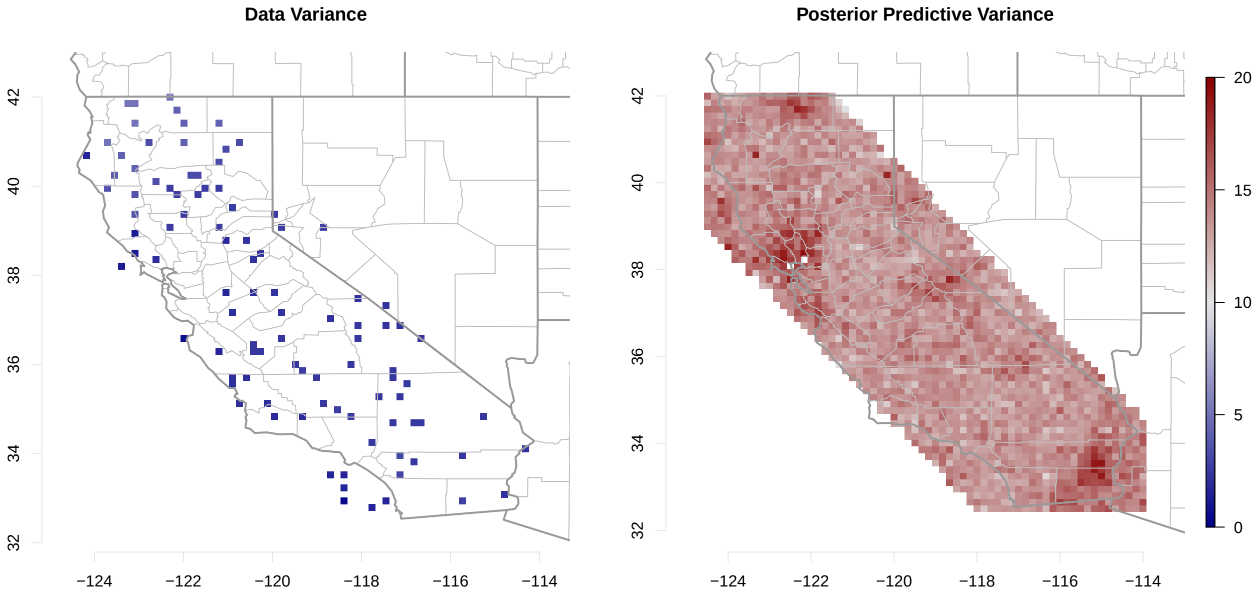

Figure 3: Left: Variance of the average maximum daily temperatures at recorded locations in California. Right: Predicted variance of the average maximum daily temperatures at various locations in the state of California for a new year. Areas with higher variance are dark red and areas with lower variance would be in dark blue. Almost all locations have high variance.