# CBMS Lecture 7

Alan E. Gelfand
Duke University

## Bayesian Nonparametric Modeling for Spatial Data using Dirichlet Processes

- What are we doing here?
- The Dirichlet Process (DP)
- The Spatial Dirichlet process (SDP) and $SDP_K$
- Comparison between Gaussian Process (GP) and SDP
- The GSDP
- The $GSDP_K$
- Comparison between SDP, GSDP, and $GSDP_K$

# Recall

- Here, point-referenced spatial data
- Often a temporal component (here, replicates)
- Spatial process specification is assumed, usually in the form of spatial random effects
- Typically a Gaussian process which is often assumed stationary

# Nonstationarity

- Kernel convolution (Higdon et al); Paciorek and Schervish extension
- But still Gaussian
- "Nonparametric" modeling for a random spatial surface, e.g., nonparametric regression literature - mean modelling
- Nonparametric variogram approaches - inadequate
- "Deformation" approach - Sampson and Guttorp, Damian et al, Schmidt and O'Hagan
- $\Longrightarrow$ nonparametric specification of the covariance function but still using a Gaussian process

# Bayesian nonparametrics

- Again, not nonparametric modelling of the mean
- Probability models for a random distribution
- Extend to a probability model for a stochastic process of random variables
- Our approach is through the use of Dirichlet processes
- Again, modeling of random effects
- Requires replications in some way
- Other approaches e.g., Gamma processes or, more generally, kernel mixtures of Gamma processes

# Again, the basic spatial data model

Suppose our observations come from a random field $Y(s), s \in D$, $D \in \mathbf{R}^d$, such that

$$Y(s) = \mu(s) + \theta(s) + \varepsilon(s),$$

| | |
|---|---|
| $\mu(s)$ | regression term $(X(s)^T\beta)$ (perhaps a trend surface) |
| $\theta(s)$ | spatial random effect |
| $\varepsilon(s)$ | pure error (noise) term $\varepsilon(s) \sim N(0, \tau^2)$ |

# Customary modeling for $\theta(s)$

- Gaussian process model specification
- Valid covariance function for $\theta(s)$
- Stationary covariance function, $C(s - s'; \phi)$
- Perhaps mixture of Gaussians, perhaps a t-process, or Gaussian/logGaussian process

# The modelling world for this paper

$$
\begin{tikzcd}
 & & \mathbf{GP} \arrow[d, "gen"] & \\
\mathbf{G} \arrow[urr, "gen"] \arrow[r, "gen"] \arrow[dr, "gen"'] & \mathbf{DP} \arrow[r, "gen"] & \mathbf{SDP} \arrow[r, "gen"] & \mathbf{GSDP} \\
 & \mathbf{DP}_k \arrow[u, "lim"'] \arrow[r, "gen"] & \mathbf{SDP}_k \arrow[u, "lim"'] \arrow[r, "gen"] & \mathbf{GSDP}_k
\end{tikzcd}
$$

# Dirichlet Processes

- A growing literature on the use of nonparametric priors, particularly Dirichlet process (DP) priors.
- The Sethuraman representation:
  Let $\theta_1^*, \theta_2^*, \ldots$ be i.i.d. $\sim G_0$.
- Let $q_1, q_2, \ldots$ be indep of $\theta^*$'s and i.i.d. $\sim \text{Beta}(1, \alpha)$.
- $G_0$ can be a distribution over random objects such as vectors, a stochastic process of random variables, or even distributions.
- If $p_1 = q_1$, $p_2 = q_2\,(1 - q_1), \ldots$, $p_k = q_k \prod_{j=1}^{k-1}(1 - q_j), \ldots$ then

$$G(\cdot) = \sum_{k=1}^{\infty} p_k\, \delta_{\theta_k^*}(\cdot)\,,$$

  is said to be distributed according to a DP.

# cont.

- The distribution of $\mathbf{p}^T = (p_1, p_2, \ldots, )$ is usually referred as a "stick-breaking" construction
- Many one dimensional stick-breaking distributions discussed in the literature (Hjort, Ishwaran and colleagues, Pitman)

# cont.

- More generally, consider:

$$G_K(\cdot) = \sum_{k=1}^{K} p_k \, \delta_{\theta_k^*}(\cdot),$$

$K$ is an integer (possibly random, allowed to be infinite), $\theta_k^*$ are i.i.d. from some $G_0$ (possibly atomic), $p_k$ distributed on the simplex $\{\mathbf{p} : \sum_{i=1}^{K} p_k = 1, p_k \geq 0, k = 1, \ldots, K\}$.

- In all of these the stick-breaking is "one-dimensional"; the probability $p_k$ is for the selection of the entire $\theta_k^*$.
- We generalize to multi-dimensional stick-breaking specifications below

# Finite dimensional versions

- *Finite Dimensional Dirichlet Priors* - if $K$ is finite, and $(p_1, \ldots, p_K) \sim Dir(\alpha_{1,K}, \ldots, \alpha_{K,K})$, again with atoms from $G_0$, then $G_K \sim DP_K(\alpha, G_0)$.
- Result (Ishwaran et al): Let $G_K \sim DP_K(\alpha, G_0)$ and $E_{G_k}(h(x)) = \int h(x) G_K(dx)$ denote a random functional of $G_K$, where $h$ is non-negative continuous with compact support. Then:
- If $\alpha_{k,K} = \lambda_K$, where $K\,\lambda_K \to \infty$, then $E_{G_k}(h(x)) \stackrel{p}{\to} E_{G_0}(h(x))$,i.e., a limiting parametric model.
- If $\sum_{k=1}^{K} \alpha_{k,K} \to \alpha > 0$ and $\max \alpha_{1,K}, \ldots, \alpha_{K,K} \to 0$ as $K \to \infty$, $E_{G_k}(h(x)) \stackrel{\mathcal{D}}{\to} E_G(h(x))$, where $G \sim \mathrm{DP}(\alpha G_0)$.

# Bringing in space; the *SDP* and $SDP_K$

- A different nonparametric spatial modeling approach using the Dirichlet process (DP).
- According to the atoms, DPs provide random univariate (and multivariate) distributions.
- A random "distribution" for a stochastic process of random variables.
- Specified through arbitrary finite dim distributions.
- Resulting process is nonstationary, resulting joint distributions are not normal.
- For $n = 1$ we have $\{F(Y(s)) : s \in D\}$. Want the $F(Y(s))$ to be dependent and, as $s \to s_0$, we want the realized $F(Y(s))$ to converge to the realized $F(Y(s_0))$. Spatial **prediction/kriging** for distributions!

# cont.

- Extend $\theta_l$ to $\theta_{l,D} = \{\theta_l(s) : s \in D\}$. For instance, $G_0$ might be a stationary GP with each $\theta_{l,D}$ being a realization from $G_0$, i.e., a surface over $D$.
- The resulting random distribution, $G$, for $\theta_D$ is called a spatial DP (SDP), denoted by $\sum_{l=1}^{\infty} \omega_l \delta_{\theta^*_{l,D}}$.
- Interpretation: $G$ induces a random probability measure $G^{(\mathbf{s}^{(n)})}$ on the space of distribution functions for the set $(\theta(s_1), ..., \theta(s_n))$.

## cont.

- Given $G$, $\mathrm{E}(\theta(s) \mid G) = \sum \omega_l \theta_l^*(s)$ and $\mathrm{Var}(\theta(s) \mid G) = \sum \omega_l \theta_l^{*2}(s) - \{\sum \omega_l \theta_l^*(s)\}^2$.
  For a pair of sites $s_i$ and $s_j$,
  $$\mathrm{Cov}(\theta(s_i), \theta(s_j) \mid G) =$$
  $$\sum \omega_l \theta_l^*(s_i)\theta_l^*(s_j) - \left\{\sum \omega_l \theta_l^*(s_i)\right\}\left\{\sum \omega_l \theta_l^*(s_j)\right\}$$
  .
- Use DP mixing to overcome the a.s. discreteness of $G$
- That is, $\theta_D$ given $G$ is a realization from $G$ and $\mathbf{Y}_D - \theta_D$ is a realization from a pure error process.

# DP mixing

- Then, **formally** a convolution,

$$F\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int \mathcal{K}\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) G\left(d\theta_D\right).$$

- *Differentiating* to densities,

$$f\left(\mathbf{Y}_D \mid G, \tau^2\right) = \int k\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) G\left(d\theta_D\right).$$

- So, $Y(s) = \theta(s) + \epsilon(s)$ where $\theta(s)$ follows a spatial SDP and $\epsilon(s)$ is $\mathrm{N}(0, \tau^2)$, a pure error (nugget) component
- Apart from mean, usual partitioning of residual
- Convolving distributions rather than convolving process variables to create a process.
- Replacing countable sums with finite sums and a Dirichlet distribution for the weights yields the $SDP_K$

# Finite dimensional joint distributions

- Joint density of $\mathbf{Y} = (Y(s_1), ..., Y(s_n))'$ given $\tau^2$ and $G^{(n)}$, where $G^{(n)} \sim DP(\nu G_0^{(n)})$, is

$$f\left(\mathbf{Y} \mid G^{(n)}, \tau^2\right) = \int N_n\left(\mathbf{Y} \mid \theta, \tau^2 I_n\right) G^{(n)}(d\theta)$$

.

- Again, the a.s. representation of $G^{(n)}$ yields that $f(\mathbf{Y} \mid G^{(n)}, \tau^2)$ is a.s. of the form $\sum_{l=1}^{\infty} \omega_l N_n(\mathbf{Y} \mid \theta_l^*, \tau^2 I_n)$, i.e., a countable location mixture of normals.
- Usually add a regression term $X'\beta$, to the kernel of the mixture model.

# The hierarchical model

- The following semiparametric hierarchical model emerges

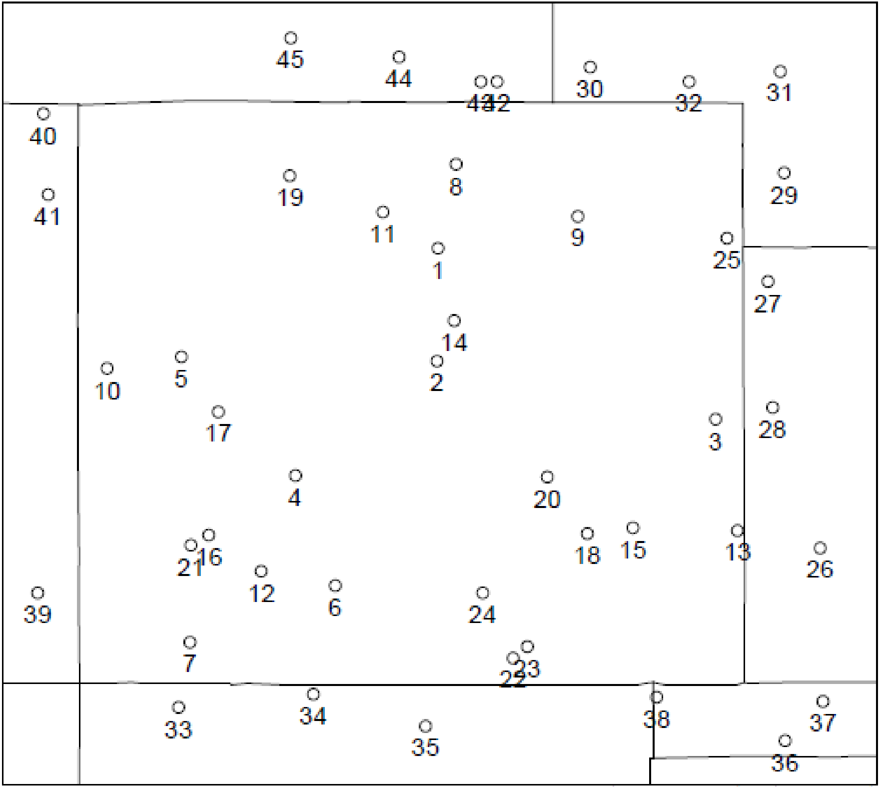
$$\begin{array}{ccc} \mathbf{Y}_t \mid \theta_t, \boldsymbol{\beta}, \tau^2 & \overset{ind.}{\sim} & N_n(\mathbf{Y}_t \mid X_t' \boldsymbol{\beta} + \theta_t, \tau^2 I_n) \\ \theta_t \mid G^{(n)} & \overset{i.i.d.}{\sim} & G^{(n)}, t = 1, \ldots, T \\ G^{(n)} \mid \nu, \sigma^2, \phi & \sim & DP(\nu G_0^{(n)}) \\ \boldsymbol{\beta}, \tau^2 & \sim & N_p(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \Sigma_{\boldsymbol{\beta}}) \mathrm{IG}(\tau^2 \mid a_\tau, b_\tau) \\ \nu, \sigma^2, \phi & \sim & \mathrm{G}(\nu \mid a_\nu, b_\nu) \mathrm{IG}(\sigma^2 \mid a_\sigma, b_\sigma) [\phi] \end{array}$$

- $G_0^{(n)}(\cdot \mid \sigma^2, \phi) = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi))$
- Model fitting is standard using "marginalization over $G$" for DP models.

# Comparing the GP with the *SDP* and $SDP_K$

- Compare the behavior of the GP, the SDP and the $SDP_K$ using data collected at 45 weather stations in Colorado
- Average monthly temperatures and precipitation data throughout 40 years (1958-1997) from NCAR.
- Use average monthly temperature for July to achieve approximate independence.
- Embedding within a dynamic model could also be done.

# The sites



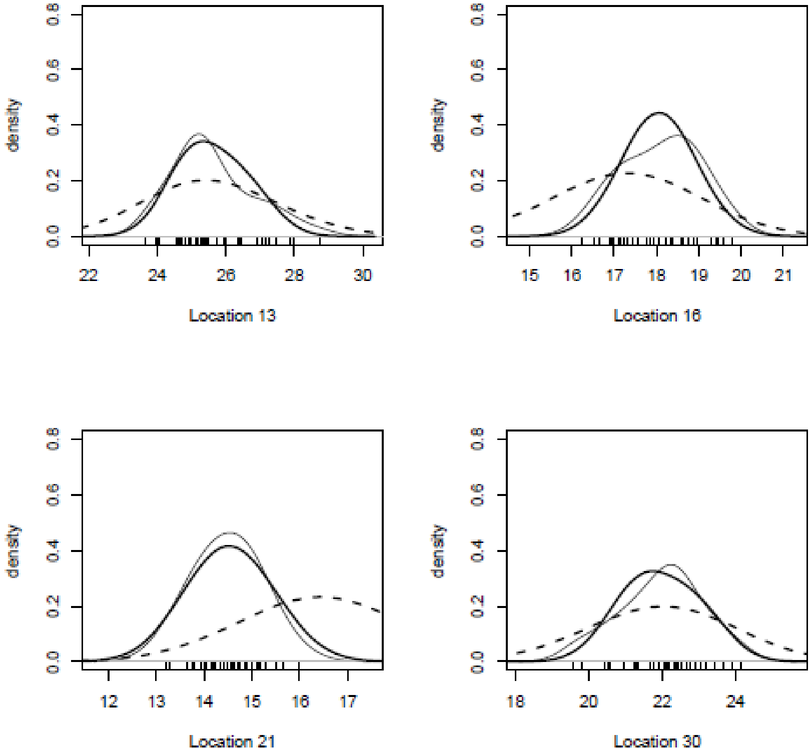
45
44
43
42
30
32
31
40
19
8
29
41
11
9
25
1
27
14
10
5
2
17
3
28
4
20
18
15
13
26
21
16
12
6
24
39
7
23
22
33
34
35
38
37
36

# Model details

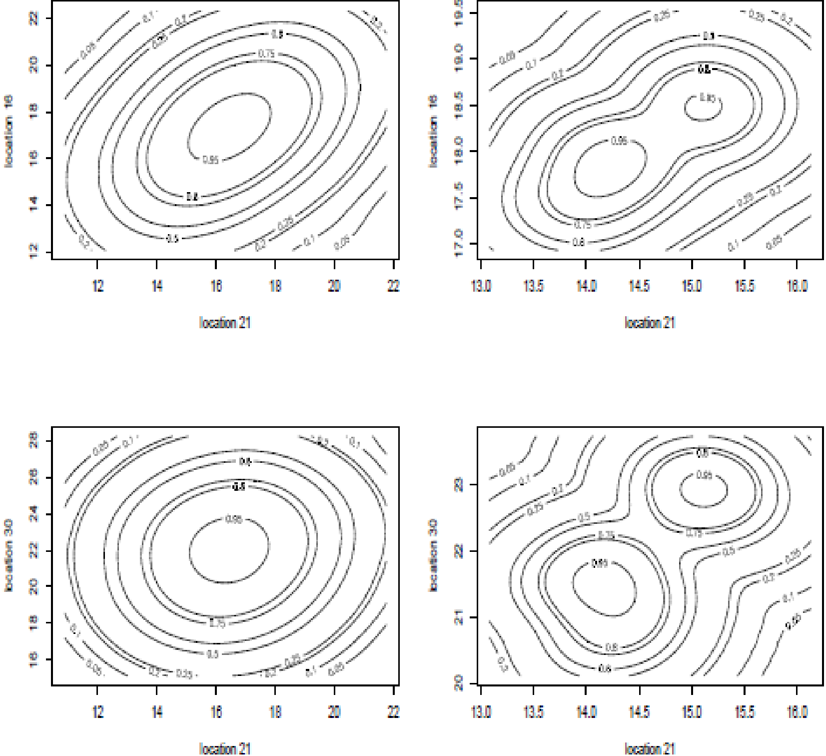
- 40 replications over 45 locations so $SDP$ and the $SDP_K$ can be fitted.
- $Y_t(s)$ is the average temperature and $\mu_t(s) = \beta_0 + \beta_1^T X_t(s)$, with $X_t(s)$ associated precipitation.
- Exponential correlation function for base process/multivariate normal distribution, $\rho(s - s') = \sigma^2 \exp\{-\phi \,||s - s'||\}$.
- To facilitate comparison with the $SDP_K$ we fix $\alpha = 10$ (trials with random $\alpha$ in the SDP didn't change the results significantly).
- $K = 10$ and $\alpha_{k,K} = \alpha/K$ implies in the $SDP_K$ that the $p_k \sim$ uniform Dirichlet.

# Comparison

- We consider the model for $\theta_t(s) + \epsilon_t(s), t = 1, 2, ..., T$
- The two extreme cases: $\alpha \to \infty$ (where the $\theta_t(s)$ are all distinct) and $\alpha \to 0$ (where $\theta_t(s) = \theta(s)$).
- From this perspective, the *SDP* (and the $SDP_K$) is in between since it permits $\alpha \in (0, \infty)$.
- $\theta_t(s) + \epsilon_t(s)$ has dependence within a replication but indep across replications ("known" mixing dist.)
- $\theta(s) + \epsilon_t(s)$ has dependence both within and across replications
- The simple GP ($\alpha \to 0$) is unable to capture the variability of multimodal data. PMSE for GP is $\approx 1600$ while for *SDP* and $SDP_K$ PMSE is $\approx 950$.
- If number of components is small relative to $K$, not much difference between the $SDP_K$ and *SDP*.




Posterior predictive densities, $[Y_{new}(s)|\text{data}]$




Contour plots of the post dist of the mean - GP and SDP

# Generalized SDP

- Motivation - Clustering using the DP is attractive, perhaps more elegant than finite mixture models, e.g., in species sampling, a mechanism that enables new species types (new classes, in general)
- But still DP can be inefficient. Suppose species are defined through a vector of traits and a new species is a *hybrid*. It would be efficient to allow different components of the vector to be drawn from different components of the $\theta_k^*$'s
- Fewer clusters would be needed, a simpler story for speciation results

# GSDP cont.

- In fact, our goal is a bit more ambitious.
- Local surface selection among the process realizations that define the *SDP* or $SDP_k$.
- Need to provide such selection for any number of and choice of locations.
- With spatial structure to such selection. The closer two locations are the more likely they are to select the same surface
- An example: in brain imaging (neurological activity level) - healthy brain images (surfaces) as well as impaired brain images (surfaces)
- Only a portion of the brain is impaired suggests surface selection according to where the brain is damaged.

# Some details

- A base random field $G_0$, say, stationary and Gaussian, with $\theta^*_{l,D} = \{\theta^*_l(s), s \in D\}$ a realization from $G_0$.
- A random probability measure $G$ on the space of surfaces over $D$ whose finite dimensional distributions have a.s. the following representation: ($K$ can be $\infty$)

$$pr\{\theta(s_1) \in A_1, \ldots, \theta(s_n) \in A_n\}$$

$$= \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} p_{i_1,\ldots,i_n} \, \delta_{\theta^*_{i_1}(s_1)}(A_1) \ldots \delta_{\theta^*_{i_n}(s_n)}(A_n)$$

# cont.

- The $\theta_j^*$'s are independent and identically distributed as $G_0^{(n)}$ and independent of the weights $\{p_{i_1,\ldots,i_n}\}$
- $i_j$ denotes $i(s_j)$, $j = 1, 2, \ldots, n$, and the $\{p_{i_1,\ldots,i_n}\}$ are distributed on the simplex $\mathbb{P} = \{p_{i_1,\ldots,i_n} \geq 0 : \sum_{i_1=1}^{K} \cdots \sum_{i_n=1}^{K} p_{i_i,\ldots,i_n} = 1\}$
- The collection of probabilities is really a process ($s$'s suppressed). Require a continuity property (essentially, Kolmogorov consistency of the finite dimensional laws); for $s_1$ and $s_2$, as $s_1 \to s_2$, $p_{i_1,i_2} = pr\{\theta(s_1) = \theta_{i_1}^*(s_1),\, \theta(s_2) = \theta_{i_2}^*(s_2)\}$, tends to the marginal probability $p_{i_2} = pr\{\theta(s_2) = \theta_{i_2}^*(s_2)\}$ when $i_1 = i_2$, and to 0 otherwise.
- Extension to $n$ locations is clear; this property is interpreted as almost sure continuity of the weights

# Digression

- The dependent Dirichlet process (DDP) in the spatial setting specifies the random distribution of $\theta(s)$ as $F_s$, yielding a collection of random distributions indexed by location
- What about the joint distribution of say $\theta(s_1), \theta(s_2)$? Suppose $pr(\theta(s_1) = \theta^*_l(s_1), \theta(s_2) = \theta^*_{l'}(s_2)) = p_l(s_1)p_{l'}(s_2)$
- Conditional independence given $F_{s_1}$ and $F_{s_2}$
- $|F_{s_1}(\cdot) - F_{s_2}(\cdot)| \to 0$ as $||s_1 - s_2|| \to 0$. Distributions become close but not realizations from the distributions
- We are constructing joint distributions, i.e., $pr(\theta(s_1) = \theta^*_l(s_1), \theta(s_2) = \theta^*_{l'}(s_2)) = p_{l,l'}(s_1, s_2)$

# Labels

- So, we are assigning local "labels"
- We can imagine a label $L(s)$ at every $s \in D$
- We need to build a labeling process
- Again, the $SDP$ and $SDP_K$ provide a constant label across all locations
- To build a labeling process we need to specify finite dimensional distibutions, again $P(L(s_1 = l_1, L(s_2) = l_2, ..., L(s_n) = l_n)$
- Can build in several ways (below); will call this a generalized spatial Dirichlet process (GSDP)
- A simple idea is a *partition* process. Partition $D$ into say $m$ regions and assign a common label to all $s$ in the same region
- If we restrict the number of atoms to $K$, hence, the number of labels to $K$, call it a $GSDP_K$

# Properties

- Can calculate moments, as with SDP
- With almost surely continuous realizations from the base process and also of the weights, weak conv of $\theta(s)$ to $\theta(s_0)$ and the GSDP is mean square cont.
- As with the SDP, the process $\theta(s)$ has heterogenous variance and is nonstationary.
- If we marginalize over $G$ with $G_0$ a $GP(0, \sigma^2 \rho_\phi(s_i - s_j))$,

$$cov\{\theta(s_i), \theta(s_j)\} = \sigma^2 \rho_\phi(s_i - s_j) \sum_{l=1}^{K} E\{p_{ll}(s_i, s_j)\}.$$

- $\sum_{l=1}^{K} E\{p_{ll}(s_i, s_j)\} < 1$, unless $p_{ll'}(s_i, s_j) = 0$, $l \neq l'$,

# GSDP through a latent spatial process

- Latent variable process determines surface selection.
- Employ Gausssian thresholding to provide binary outcomes, i.e., assume that $\{Z_l(s),\ s \in D,\ l = 1, 2, \ldots\}$ are indep $GP(\mu_l(s), \rho_Z(\cdot, \eta))$
- Then $q_{l,u_1,\ldots,u_n}(s_1, \ldots, s_n) =$

$$pr\{\delta^*_{\{Z_l(s_1)\geq 0\}} = u_1, \ldots, \delta^*_{\{Z_l(s_n)\geq 0\}} = u_n |\, \mu_l(s_1), \ldots, \mu_l(s_n)\}$$

- At any location $s$ we obtain

$$q_{l,1}(s) = pr\{Z_l(s) \geq 0\} = 1 - \Phi\left\{-\mu_l(s)\right\} = \Phi\left\{\mu_l(s)\right\},$$

- If the $\mu_l(s)$ such that $\Phi\{\mu_l(s)\}$ are indep Beta$(1, \nu)$, then for each $s$, $\theta(s)$ is a DP, probabilities vary with location.

## Comparing the SDP's and GSDP's

- Compare using a simulated data set
- Data are generated from a finite mixture model of GPs. Let $\mathbf{Y}_t = (Y_t(s_1), \ldots, Y_t(s_n))^T$
- $Y_t(s)$ arises from a mixture of two GPs, $G_0^1(\xi_1, \sigma_1^2 \rho_{\psi_1})$ and $G_0^2(\xi_2, \sigma_2^2 \rho_{\psi_2})$ such that

$$Y_t(s) \sim \alpha(s)\, G_0^1 + (1 - \alpha(s))\, G_0^2.$$

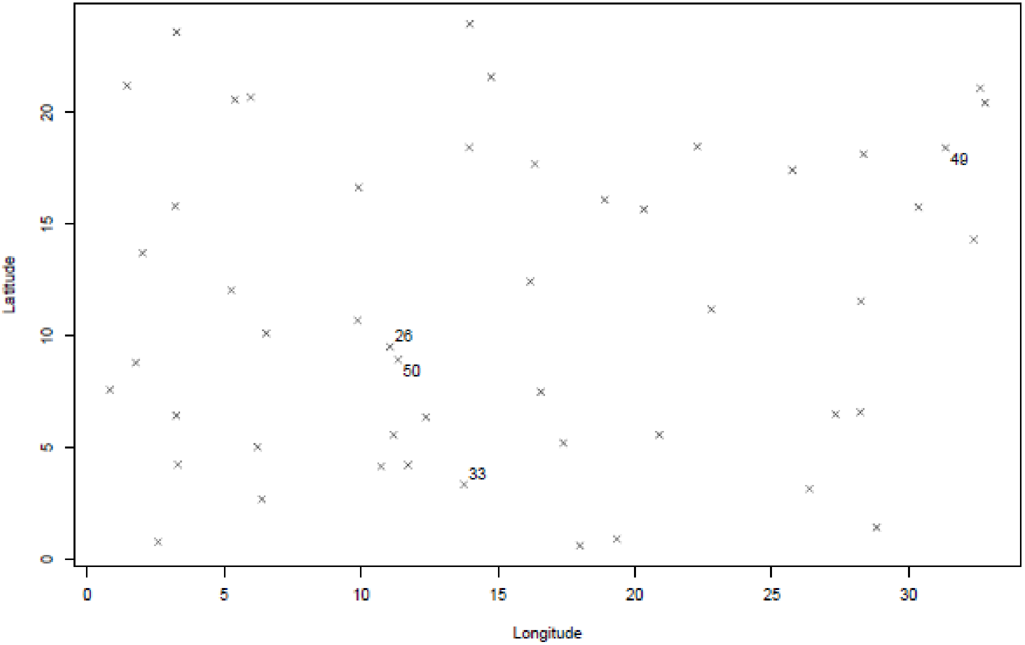
- Marginal weights $\alpha(s) = P(Z(s) > 0)$, where $Z(s)$ is a mean zero stationary GP with cov function $\rho_\eta(s - s')$.
- The joint distribution for $s, s'$ in $D$ is

$$\begin{aligned}(Y_t(s), Y_t(s')) \sim\, & \alpha_{1,1}(s,s')\, G^1_{0,s,s'} + \alpha_{2,2}(s,s')\, G^2_{0,s,s'} + \\ & + \alpha_{1,2}(s,s')\, G^1_{0,s}\, G^2_{0,s'} + \alpha_{2,1}(s,s')\, G^2_{0,s}\, G^1_{0,s'}\end{aligned}$$

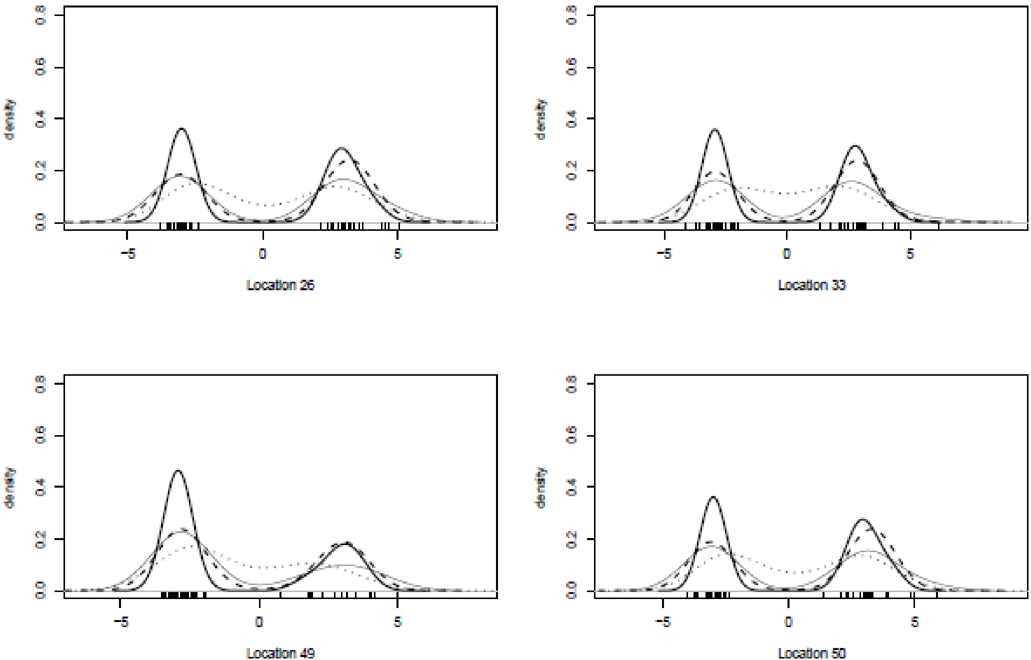
where $\alpha_{i,j} = P\left((-1)^{i+1} Z(s) > 0,\, (-1)^{j+1} Z(s') > 0\right)$, $i, j = 1, 2$.

# Specifications

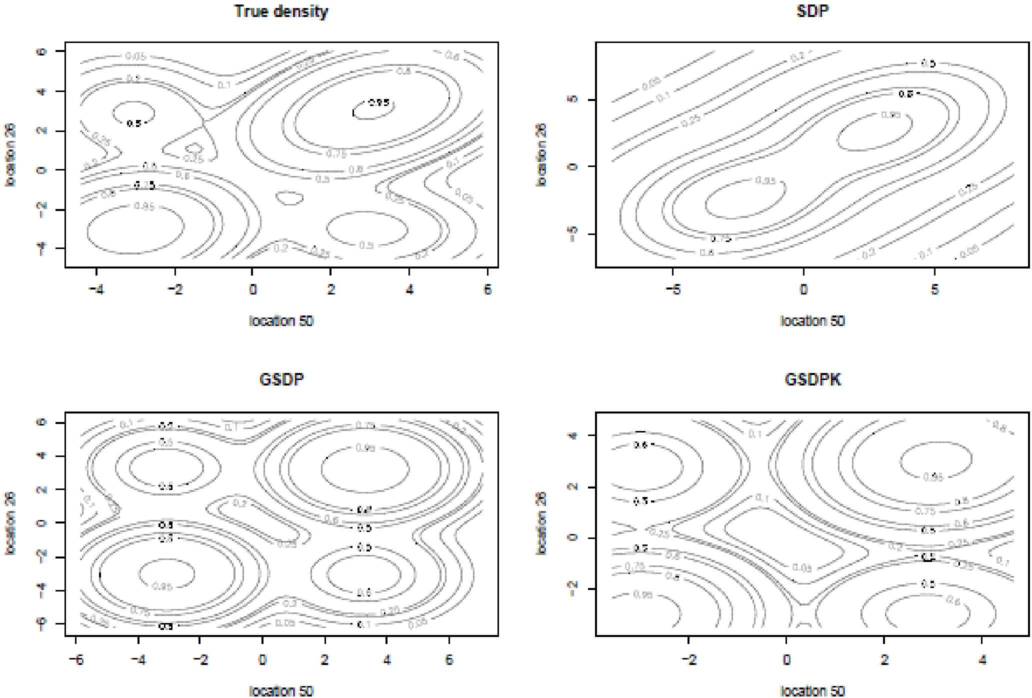
- $n = 50$ and $T = 40$.
- Also, $\xi_1 = -\xi_2 = 3$, $\sigma_1 = 2\,\sigma_2 = 2$, $\phi_1 = \phi_2 = 0.3$, and $\eta = 0.3$.
- We fit the SDP model, the GSDP and the $GSDP_K$ with $K = 20$.
- To focus on the modeling of the spatial association, we assume $\mu(s) = 0$



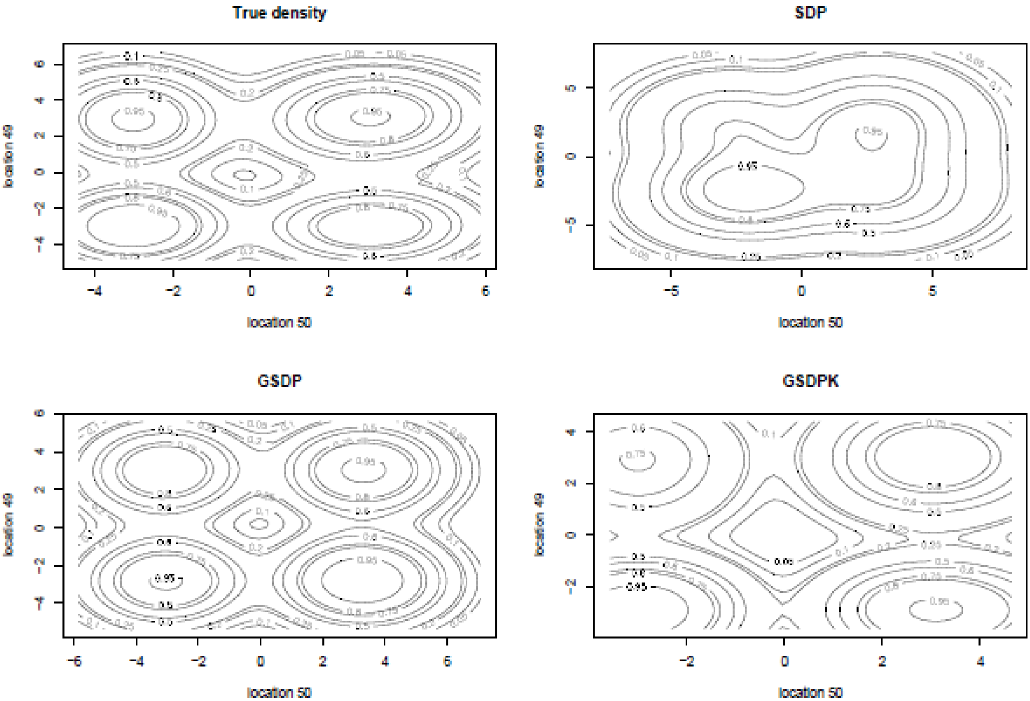

The Design Locations for the Simulation Example




True density, predictive posterior density for SDP, GSDP, and $GSDP_K$




Contour plots for locations 26 and 50.




Contour plots for locations $49$ and $50$.

# Ongoing work

- Time dependent replications - embed the GSDP's within a dynamic model
- Multivariate spatial data has not been addressed. Coregionalization (random linear transformation) approach. Random transformation introduced into the base measure or random linear transformation of SDP realizations
- Functional data analysis (FDA). Replace geographic space $s \in D$ with covariate space $z \in Z$. Atoms in DP are random functions.
- Multivariate FDA, e.g., an ensemble of functional data for a patient over time
- Finally, spatial FDA. Use DP specifications to handle both functional and spatial aspects of the modelling.