Bayesian Conditional Density Estimation using Partition Models

Bani K. Mallick

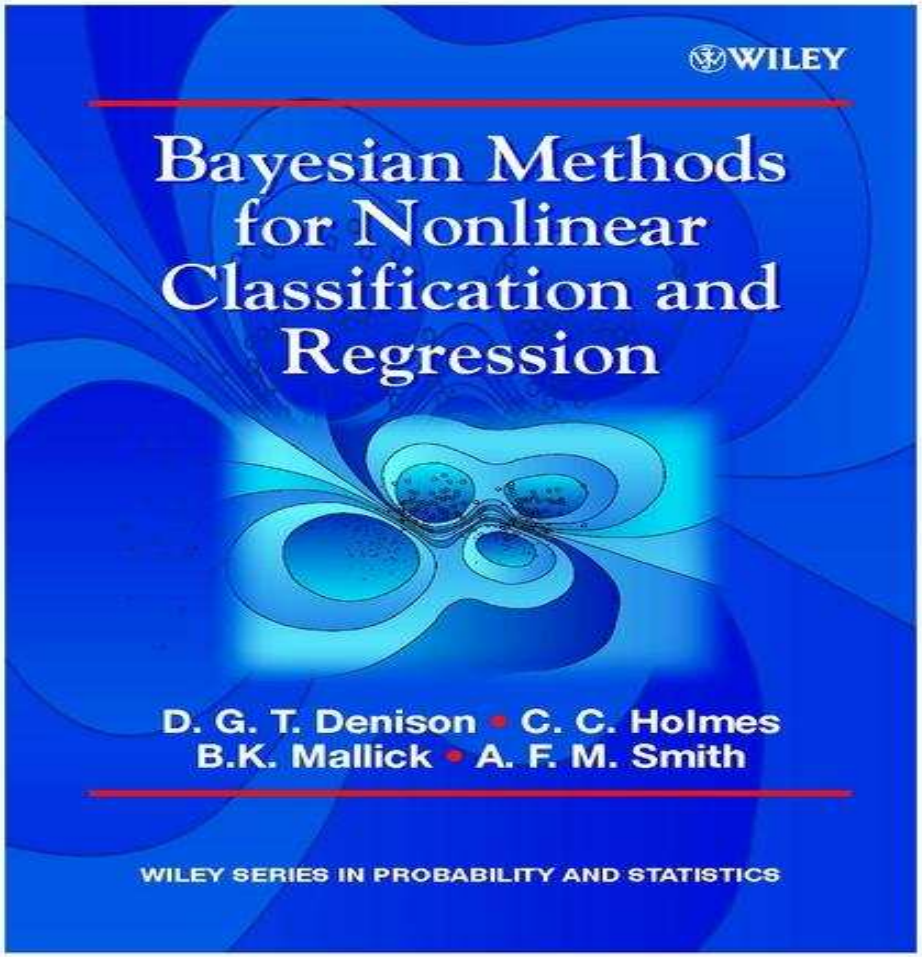
May 10, 2018

STATISTICS
TEXAS A&M UNIVERSITY

Introduction

Modeling with Covariates
Y: response
X: Covariates
Regression
Classification
Supervised Learning

Book


WILEY
Bayesian Methods for Nonlinear Classification and Regression
D. G. T. Denison • C. C. Holmes
B.K. Mallick • A. F. M. Smith
WILEY SERIES IN PROBABILITY AND STATISTICS

Definition

What is density regression or conditional density estimation?
Regress the whole density function with covariates

Definition

What is density regression or conditional density estimation?
Regress the whole density function with covariates

$$p(y \mid X = x)$$

Regression is a special case:

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \\ p(y \mid X = x) &\equiv N(x\beta, \sigma^2) \end{aligned}$$

Definition

What is density regression or conditional density estimation?
Regress the whole density function with covariates

$$p(y \mid X = x)$$

Regression is a special case:

$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \\ p(y \mid X = x) &\equiv N(x\beta, \sigma^2) \end{aligned}$$

Many datasets don't fit these strong parametric assumptions.

Definition

What is density regression or conditional density estimation?
Regress the whole density function with covariates

$$p(y \mid X = x)$$

Regression is a special case:

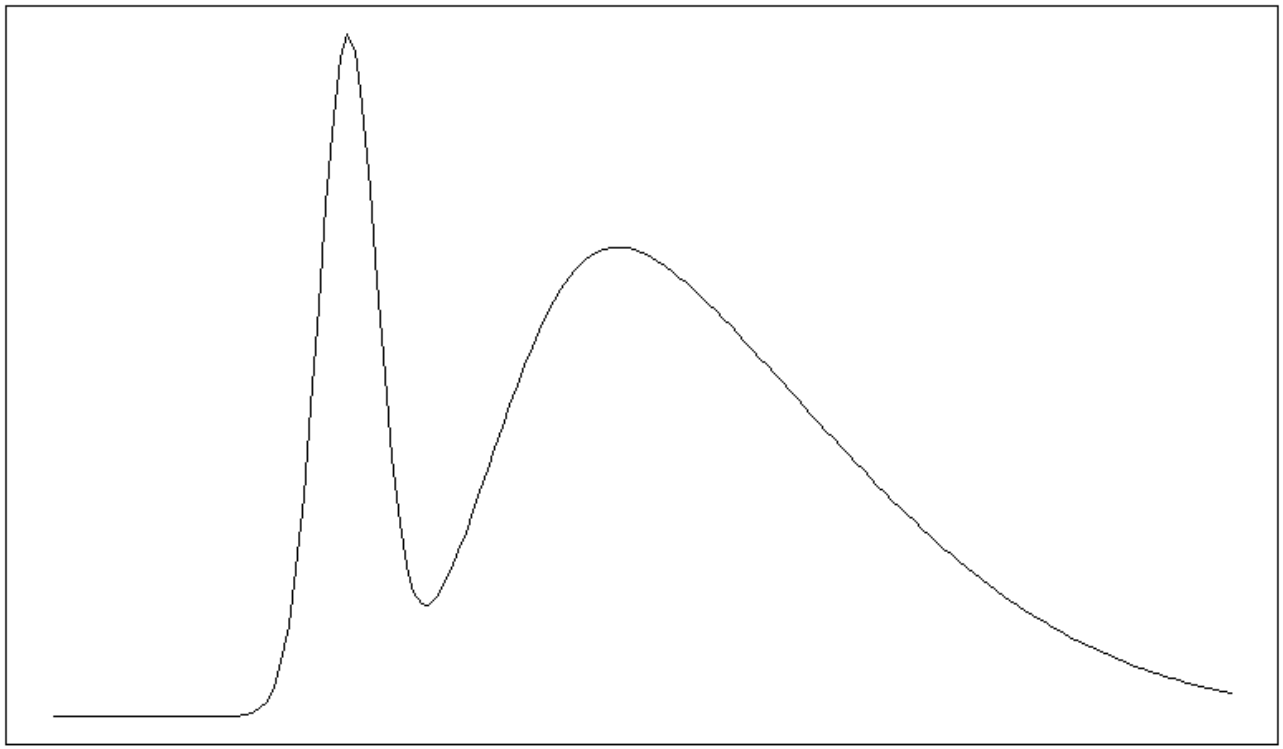
$$\begin{aligned} Y &= X\beta + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \\ p(y \mid X = x) &\equiv N(x\beta, \sigma^2) \end{aligned}$$

Many datasets don't fit these strong parametric assumptions.
Some methods seek to be more flexible by:

- Modeling the mean flexibly (e.g. splines)
- Modeling $\epsilon(x)$
- Combining the two above
- Other approaches

Simple Model with one covariate: Sex

$$\begin{aligned} Y &= \beta_0 + \epsilon \\ \epsilon &\sim \textit{Dirichlet Process} \end{aligned}$$



Motivation

When is conditional density estimation useful?

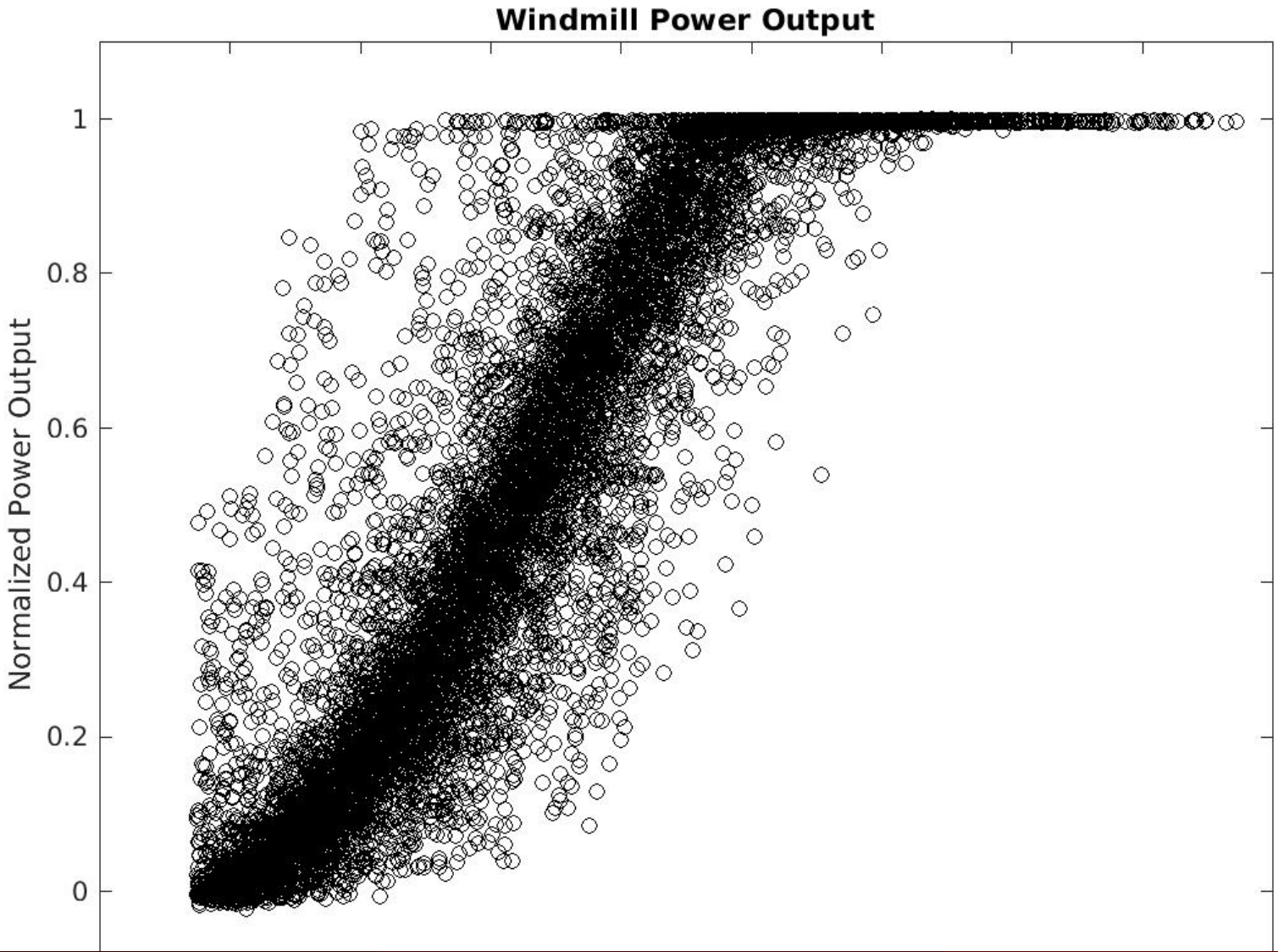
Motivation

When is conditional density estimation useful?

Windmill Data

- Power output from windmills
- Wind energy is one of the fastest growing renewable energy sources
- In wind industry the power curve measures the relationship between power output of a turbine and wind speed

Wind Curve


Windmill Power Output
Normalized Power Output
1
0.8
0.6
0.4
0.2
0

Wind Curve

- A turbine starts to produce power after the wind reaches the cut-in speed
- A nonlinear relation between power output and wind speed then ensue, until wind reaches the rated speed
- When the wind speed is beyond rated speed, the turbine output power will be restricted at the rated power output using control mechanism
- The turbine will be halted when the wind reaches cut-off speed

Other variables

- Though power curve follow the general trend, there appears to be a considerable amount of information that cannot be accounted.
- Wide array of sensors measure other variables too
- wind speed
- wind direction
- air density
- wind sheer
- turbulence intensity

Current Practice

- Nonparametric approach known as the binning method (IEC 2005)
- Discretize the domain of wind speed into finite number of bins and take sample averages of the power output within each bin
- Usual bin width .5m/s
- All other variables are ignored
- Our technical objective is to estimate the conditional density $p(y \mid \mathbf{x})$

Existing (Bayesian) Conditional Density Estimation

- Use mixture models (e.g. Dirichlet process) for the conditional distribution of $p(y \mid \mathbf{x})$ and allow the mixing weights as well as the parameters to depend on the covariates (Chung and Dunson, 2009; Dunson and Park, 2008; Dunson et al., 2007; Griffin and Steel, 2006).

Existing (Bayesian) Conditional Density Estimation

- Use mixture models (e.g. Dirichlet process) for the conditional distribution of $p(y \mid \mathbf{x})$ and allow the mixing weights as well as the parameters to depend on the covariates (Chung and Dunson, 2009; Dunson and Park, 2008; Dunson et al., 2007; Griffin and Steel, 2006).
- Kernels, splines, mixtures of experts (Fan et al., 1996; Fu et al., 2011; Kooperberg and Stone, 1991; Stone et al., 1997; Jacobs et al., 1991).

Existing (Bayesian) Conditional Density Estimation

- Use mixture models (e.g. Dirichlet process) for the conditional distribution of $p(y \mid \mathbf{x})$ and allow the mixing weights as well as the parameters to depend on the covariates (Chung and Dunson, 2009; Dunson and Park, 2008; Dunson et al., 2007; Griffin and Steel, 2006).
- Kernels, splines, mixtures of experts (Fan et al., 1996; Fu et al., 2011; Kooperberg and Stone, 1991; Stone et al., 1997; Jacobs et al., 1991).
- Subspace projection (Tokdar et al., 2010)
- Latent variable methods (Kundu and Dunson, 2011; Bhattacharya and Dunson, 2010)

Windmill Data

All the referred methods are suitable when the density of y changes smoothly over the covariate space.
It is known that there are sharp changes in the density of y over the covariates.

Windmill Data

All the referred methods are suitable when the density of y changes smoothly over the covariate space.
It is known that there are sharp changes in the density of y over the covariates.

- Pitch control at high wind speeds
- To protect generator under high wind
- Wind speed is very high, the turbine blades turned parallel to the wind to reduce the energy absorption capability
- Hence, the power output is concentrated near maximum power level

Windmill Data

- Terrain: Not flat and smooth the sharp change in power output
- Wake effects: Downwind from another turbine creates sharp change in power output

Partition Model

- A partition model is made up by splitting the covariate space $\mathbf{X}$ in M disjoint regions, say $R_1 \cdots R_M$
- The response Y in each region is assumed to be exchangeable, generated from a common density
- The distribution of Y is independent among the partition
- Motivation: points nearby in covariate space come from the same local distribution
- Naturally models sharp changes in covariate space
- Obtain the partition adaptively, hence automatically finds important change points

Partition model

Partition Model

- Partitions the data into M pieces
- Fits a separate model for each piece

Partition model

Partition Model

- Partitions the data into M pieces
- Fits a separate model for each piece

$$p(\mathbf{y} \mid \text{Partition}) = \prod_{i=1}^{M} p(\mathbf{y}_i)$$

Partition model

Partition Model

- Partitions the data into M pieces
- Fits a separate model for each piece

$$p(\mathbf{y} \mid \text{Partition}) = \prod_{i=1}^{M} p(\mathbf{y}_i)$$

- Can assume i.i.d. structure of the data (covariates only influence partitioning), or not.

Voronoi Partition

How to define regions so that points close together have the same distribution?

Let $y_1, \ldots, y_n \in \mathcal{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ represent the observed response and covariate vectors.

A tessellation with M regions is determined by choosing M centers, $\mathbf{c}_1, \ldots, \mathbf{c}_M$ and a weight vector, $\mathbf{w}$, $\sum w_k = 1$.

Voronoi Partition

How to define regions so that points close together have the same distribution?

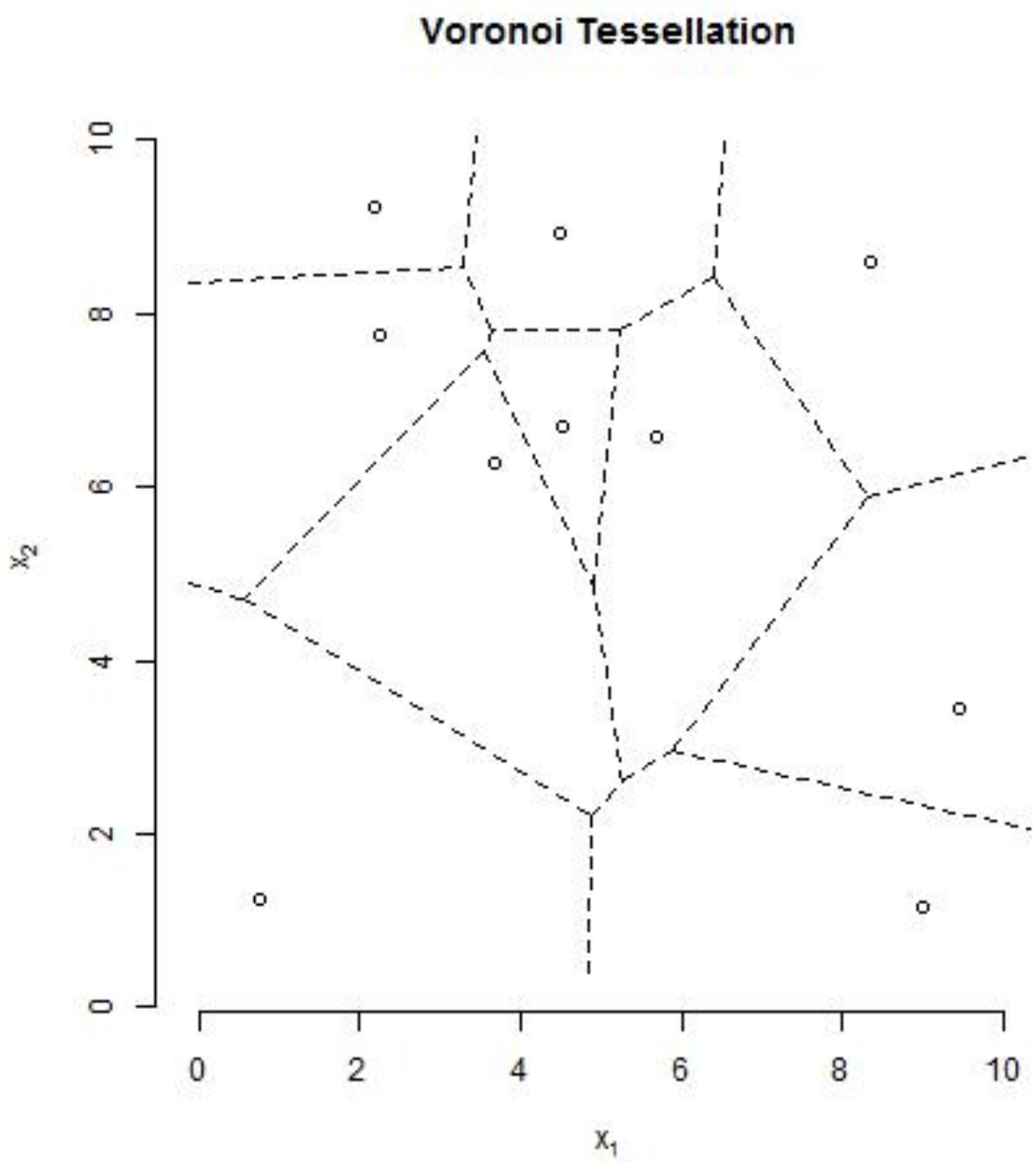
Let $y_1, \ldots, y_n \in \mathcal{Y}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ represent the observed response and covariate vectors.

A tessellation with M regions is determined by choosing M centers, $\mathbf{c}_1, \ldots, \mathbf{c}_M$ and a weight vector, $\mathbf{w}$, $\sum w_k = 1$.

A region of the tessellation is defined as all the points in $\mathcal{X}$ that are closest to $\mathbf{c}_i$, i.e.

$$R_i = \{\mathbf{x} \in \mathcal{X} : ||\mathbf{x} - \mathbf{c}_i|| < ||\mathbf{x} - \mathbf{c}_j|| \ \forall \ i \neq j\}$$

where $||\mathbf{x}|| = ||(x_1, \ldots, x_p)|| = \sum_{i=1}^{p} w_i x_i^2$.


Voronoi Tessellation
x_2
x_1
0
2
4
6
8
10

Voronoi Partition Prior

$$p(\mathbf{c}, M, \mathbf{w}) \quad = \quad p(\mathbf{c} \mid M)p(M)p(\mathbf{w})$$

Voronoi Partition Prior

$$
\begin{aligned}
p(\mathbf{c}, M, \mathbf{w}) &= p(\mathbf{c} \mid M) p(M) p(\mathbf{w}) \\
\\
p(M) &= \mathrm{DU}(M \mid 1, \ldots, M_{\max}) \\
p(\mathbf{c} \mid M) &= \mathrm{DU}\left(\mathbf{c} \mid 1, \ldots, \binom{n}{M}\right) \\
p(\mathbf{w}) &= \mathrm{Di}(\mathbf{w} \mid 1, \ldots, 1)
\end{aligned}
$$

DU: Discrete Uniform
DI: Dirichlet Distribution

Distribution within a Partition

- Regression: Gaussian Model with conjugate prior (Denison eta l.)
- Classification: Binary Model with conjugate prior (Denison et al.)
- Density regression: **Need a flexible model for Density Functions**

Logistic Gaussian Process

Logistic Gaussian process

$$p(y) = \frac{\exp(f(y))}{\int_{\mathcal{Y}} \exp(f(s))ds}$$

Logistic Gaussian Process

Logistic Gaussian process

$$p(y) = \frac{\exp(f(y))}{\int_{\mathcal{Y}} \exp(f(s))ds}$$

- Clearly $p()$ defines a stochastic process whose realizations satisfy $p \geq 0$ and $\int_0^1 p() = 1$: properties which define a density function
- We model unrestricted $f(\cdot)$ as a Gaussian process
- How can we bring covariates $\mathbf{x}$ in this modeling?

Partitioned Logistic Gaussian Process

Partition the covariates $\mathbf{x}$ in separate regions
We assume that the ith density in region R_i, $i = 1, \ldots, M$ can be modeled using a logistic Gaussian process

$$p_i(y) = \frac{\exp(f_i(y))}{\int_{\mathcal{V}_i} \exp(f_i(s))ds}$$

Partitioned Logistic Gaussian Process

Partition the covariates $\mathbf{x}$ in separate regions
We assume that the ith density in region R_i, $i = 1, \ldots, M$ can be modeled using a logistic Gaussian process

$$p_i(y) = \frac{\exp(f_i(y))}{\int_{\mathcal{V}_i} \exp(f_i(s))ds}$$

- We model the density of y over a compact set, $\mathcal{V}_i \in \mathbb{R}$
- $f_i(\cdot)$ is a Gaussian process

GP Prior on $f_i(\cdot)$

$$f_i(\cdot) = \mu_i(\cdot) + g_i(\cdot)$$
$$g_i(\cdot) \sim GP(0, \kappa(\cdot, \cdot))$$

GP Prior on $f_i(\cdot)$

$$
\begin{aligned}
f_i(\cdot) &= \mu_i(\cdot) + g_i(\cdot) \\
g_i(\cdot) &\sim GP(0, \kappa(\cdot, \cdot)) \\
\mu_i(\cdot) &= \mathbf{h}(\cdot)^T \boldsymbol{\beta}_i
\end{aligned}
$$

GP Prior on $f_i(\cdot)$

$$
\begin{aligned}
f_i(\cdot) &= \mu_i(\cdot) + g_i(\cdot) \\
g_i(\cdot) &\sim GP(0, \kappa(\cdot, \cdot)) \\
\mu_i(\cdot) &= \mathbf{h}(\cdot)^T \boldsymbol{\beta}_i \\
\mathbf{h}(y) &= (y, y^2)^T
\end{aligned}
$$

GP Prior on $f_i(\cdot)$

$$
\begin{aligned}
f_i(\cdot) &= \mu_i(\cdot) + g_i(\cdot) \\
g_i(\cdot) &\sim GP(0, \kappa(\cdot, \cdot)) \\
\mu_i(\cdot) &= \mathbf{h}(\cdot)^T \boldsymbol{\beta}_i \\
\mathbf{h}(y) &= (y, y^2)^T \\
\boldsymbol{\beta}_i &\sim N(\mathbf{b}, B)
\end{aligned}
$$

GP Prior on $f_i(\cdot)$

$$
\begin{aligned}
f_i(\cdot) &= \mu_i(\cdot) + g_i(\cdot) \\
g_i(\cdot) &\sim GP(0, \kappa(\cdot, \cdot)) \\
\mu_i(\cdot) &= \mathbf{h}(\cdot)^T \boldsymbol{\beta}_i \\
\mathbf{h}(y) &= (y, y^2)^T \\
\boldsymbol{\beta}_i &\sim N(\mathbf{b}, B)
\end{aligned}
$$

$\boldsymbol{\beta}_i$ can be integrated out to yield the marginal prior for $f_i(\cdot)$

$$
f_i(\cdot) \sim GP\left(\mathbf{h}(\cdot)^T \mathbf{b},\ \kappa(\cdot, \cdot) + \mathbf{h}(\cdot)^T B \mathbf{h}(\cdot)\right)
$$

GP Prior on $f_i(\cdot)$

$$
\begin{aligned}
f_i(\cdot) &= \mu_i(\cdot) + g_i(\cdot) \\
g_i(\cdot) &\sim GP(0, \kappa(\cdot, \cdot)) \\
\mu_i(\cdot) &= \mathbf{h}(\cdot)^T \boldsymbol{\beta}_i \\
\mathbf{h}(y) &= (y, y^2)^T \\
\boldsymbol{\beta}_i &\sim N(\mathbf{b}, B)
\end{aligned}
$$

$\boldsymbol{\beta}_i$ can be integrated out to yield the marginal prior for $f_i(\cdot)$

$$f_i(\cdot) \sim GP\left(\mathbf{h}(\cdot)^T \mathbf{b},\ \kappa(\cdot, \cdot) + \mathbf{h}(\cdot)^T B \mathbf{h}(\cdot)\right)$$

We choose $\mathbf{b} = \mathbf{0}$ and $B = I\lambda^2$ for the examples later.

Discretization and Likelihood

The integral in the denominator of Logistic Gaussian Process involves the entire sample path, which is an infinite dimensional object which makes it infeasible to carry out any likelihood-based computation

For perspective of efficient computing, we make finite dimensional approximation through discretization

The spacing between the grid points decreases, the Kullback-Leibler divergence from an infinite-dimensional model to the finite-dimensional approximation converges to zero

Discretization and Likelihood

We choose m points on a regular grid $(z_1, \ldots, z_m) \in \mathcal{Y}$ on which we evaluate $\mathbf{f}_i = (f_i(z_1), \ldots, f_i(z_m))^T$ which yields a discretized version of the density.

$$p(\mathbf{y}_i \mid \mathbf{f}_i) = \exp\left\{ \mathbf{y}_i^{\star T} \mathbf{f}_i - n_i \log\left(\sum_{j=1}^{m} \exp(\mathbf{f}_{ij}) \right) \right\}$$

where $\mathbf{y}_i$ is the collection of n_i observed responses in region R_i, $\mathbf{y}_i^{\star}$ is a column vector of length m with the jth element as the number of elements of $\mathbf{y}_i$ that fall into the subregion centered at z_j.

Posterior Tessellation

We are interested in searching the posterior of the tessellation structure, $T = \{\mathbf{c}, M, \mathbf{w}\}$,

$$p(T \mid \mathbf{y}) \propto p(T)p(\mathbf{y} \mid T)$$

Posterior Tessellation

We are interested in searching the posterior of the tessellation structure, $T = \{\mathbf{c}, M, \mathbf{w}\}$,

$$p(T \mid \mathbf{y}) \propto p(T)p(\mathbf{y} \mid T) = p(T) \prod_{i=1}^{M} \int p(\mathbf{y}_i \mid \mathbf{f}_i, T)p(\mathbf{f}_i)d\mathbf{f}_i$$

Posterior Tessellation

We are interested in searching the posterior of the tessellation structure, $T = \{\mathbf{c}, M, \mathbf{w}\}$,

$$p(T \mid \mathbf{y}) \propto p(T)p(\mathbf{y} \mid T) = p(T) \prod_{i=1}^{M} \int p(\mathbf{y}_i \mid \mathbf{f}_i, T)p(\mathbf{f}_i)d\mathbf{f}_i$$

- Often likelihoods and priors are chosen such that the integral has an analytical form.

Posterior Tessellation

We are interested in searching the posterior of the tessellation structure, $T = \{\mathbf{c}, M, \mathbf{w}\}$,

$$p(T \mid \mathbf{y}) \propto p(T)p(\mathbf{y} \mid T) = p(T) \prod_{i=1}^{M} \int p(\mathbf{y}_i \mid \mathbf{f}_i, T)p(\mathbf{f}_i)d\mathbf{f}_i$$

- Often likelihoods and priors are chosen such that the integral has an analytical form.
- In our case, there is no analytical form, so we employ a Laplace approximation to estimate each integral.

Laplace Approximation

The Laplace approximation performs a Taylor expansion around $\log[p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)]$ and then uses a multivariate-normal density centered at $\hat{\mathbf{f}}_i = \arg\max_{\mathbf{f}_i} p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)$.

Laplace Approximation

The Laplace approximation performs a Taylor expansion around $\log[p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)]$ and then uses a multivariate-normal density centered at $\hat{\mathbf{f}}_i = \arg\max_{\mathbf{f}_i} p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)$.

- $\hat{\mathbf{f}}_i$ is obtained efficiently using Newton's method

Laplace Approximation

The Laplace approximation performs a Taylor expansion around $\log[p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)]$ and then uses a multivariate-normal density centered at $\hat{\mathbf{f}}_i = \arg\max_{\mathbf{f}_i} p(\mathbf{y}_i \mid \mathbf{f}_i)p(\mathbf{f}_i)$.

- $\hat{\mathbf{f}}_i$ is obtained efficiently using Newton's method
- Laplace approximation has a closed form

Selection of Hyperparameters

In each partition element, we have to select the covariance function parameters for the prior on $\mathbf{f}_i$.
We select squared exponential covariance function.

$$\kappa(z, z') = \sigma_i^2 \exp\left(-\frac{1}{2l_i^2}(z - z')^2\right)$$

Use empirical Bayes to find σ_i^2, l_i.

Consistency Results

If the proposed model is true, then as n goes to infinity, the posterior density concentrates near a small total variation neighborhood around the true density.

First, we showed the consistency under the true model where the target density has partition form.

Consistency Results

- First show that the partition formed by a Voronoi tessellation can adequately approximate the true partition
- Then establish that we have sufficient prior probability for the approximating partitioning and around any small neighborhood of the true Gaussian process path in supremum norm
- Finally, if we have sufficient prior mass around the true density, the likelihood pulls the posterior density towards the data generating density under the true model

Consistency Results

Theorem

Let $U_{\epsilon'} = \{p : \int |p(y|\mathbf{x}) - p^*(y|\mathbf{x})| dH(\mathbf{x}) dy < \epsilon'\}$, $\epsilon' > 0$ *and* $p^*(y|\mathbf{x})$ *denote the true conditional density. Then, under some condition* $\Pi(U_{\epsilon'}|\cdot) \rightarrow 1$ *with probability one, as n the number of observations goes to infinity.*

We show similar consistency results under model misspecification where the true conditional density is Lipschitz continuous.

Simulation

500 observations were simulated from the following model:

$$Y \mid X_1, X_2 \sim \begin{cases} \text{Gamma}(10, 2) & \text{if } X_1 > X_2,\ X_2 < .75 \\ .5N(1,1) + .5N(5,1) & \text{if } X_1 < X_2,\ X_1 < .75 \\ N(1, \sqrt{.5}) & \text{if } X_1 > .75,\ X_2 > .75 \end{cases}$$

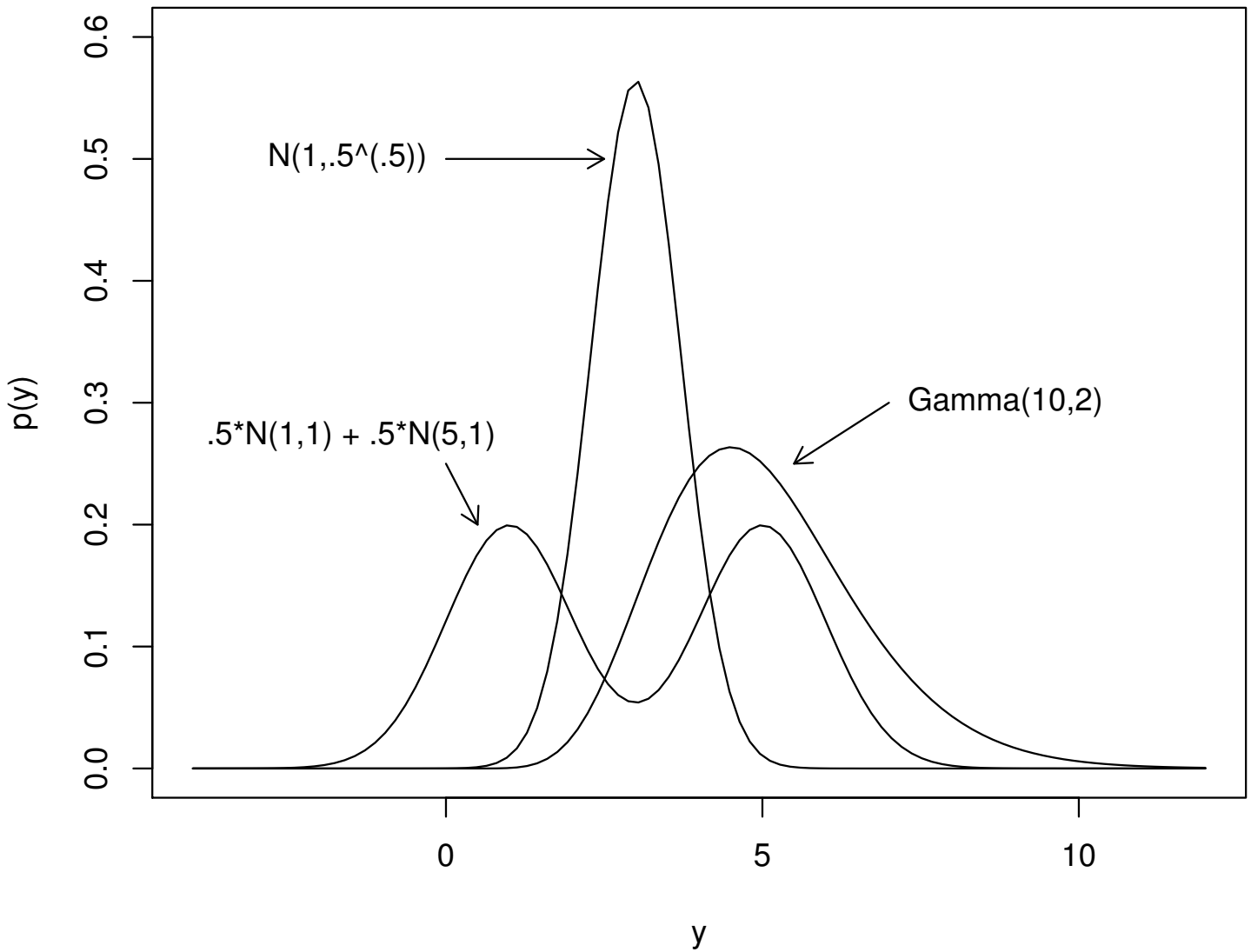
Simulation

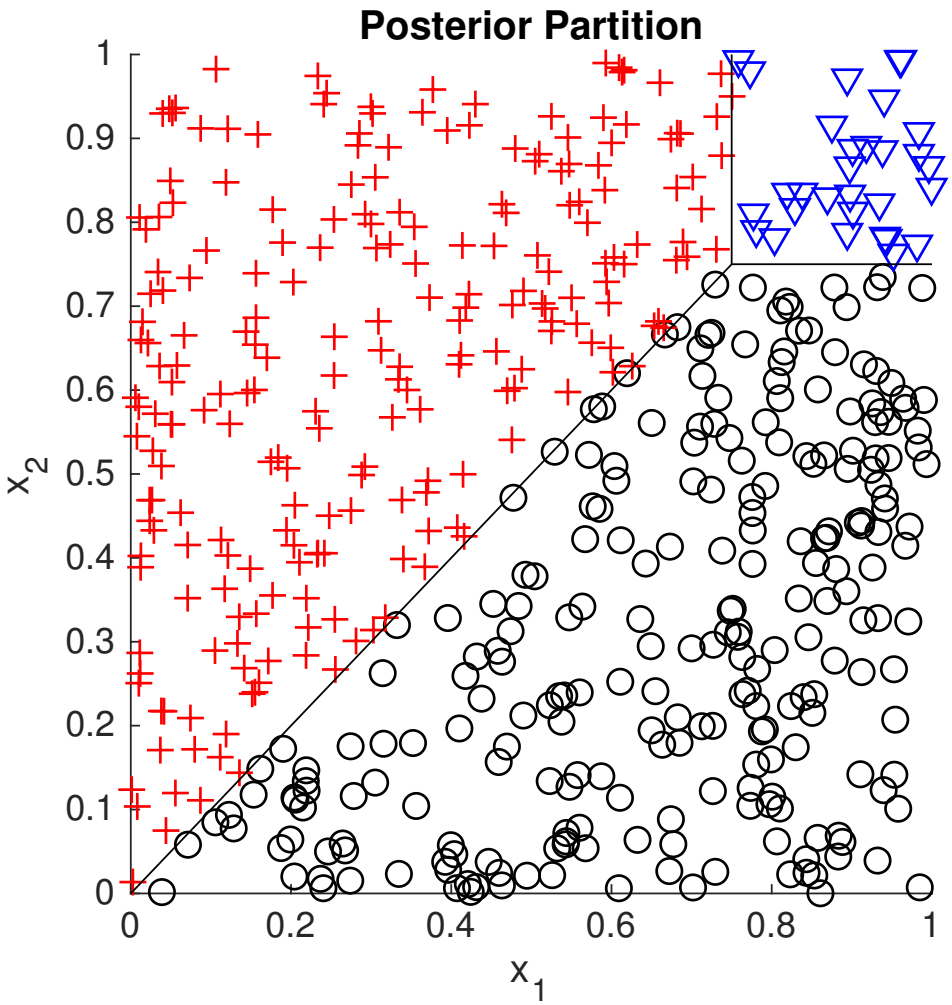
500 observations were simulated from the following model:

$$Y \mid X_1, X_2 \sim \begin{cases} \text{Gamma}(10, 2) & \text{if } X_1 > X_2,\ X_2 < .75 \\ .5N(1,1) + .5N(5,1) & \text{if } X_1 < X_2,\ X_1 < .75 \\ N(1, \sqrt{.5}) & \text{if } X_1 > .75,\ X_2 > .75 \end{cases}$$

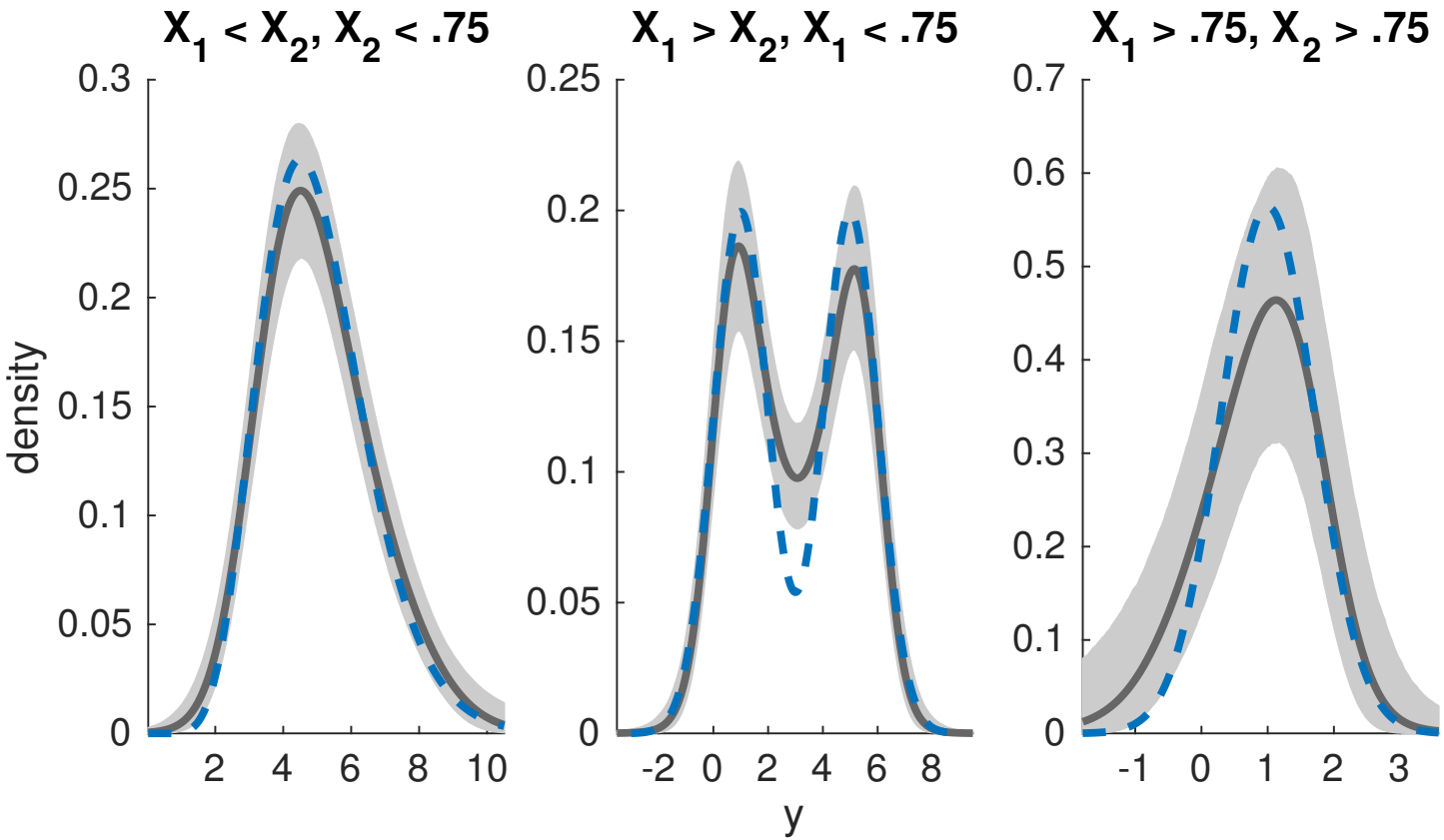
Comparison with:

- Dependent Bernstein polynomials (Barrientos et al., 2017)
- Linear dependent tail-free processes (Jara and Hanson, 2011)
- Dirichlet process mixtures of normals (Müller et al., 1996)
- Voronoi partition assuming normality (Denison et al., 2002)


Partition Densities
p(y)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
N(1,.5^(.5))
.5*N(1,1) + .5*N(5,1)
Gamma(10,2)
0
5
10
y



The partition with the highest posterior probability from the simulated dataset. The true partition boundaries are denoted by the black lines, and color and shape of the points designate which partition element they belong to.



The posterior mean (solid) and 95% credible intervals (shaded) of the data in each partition element from the partition model. The true density is also shown (blue, dashed).

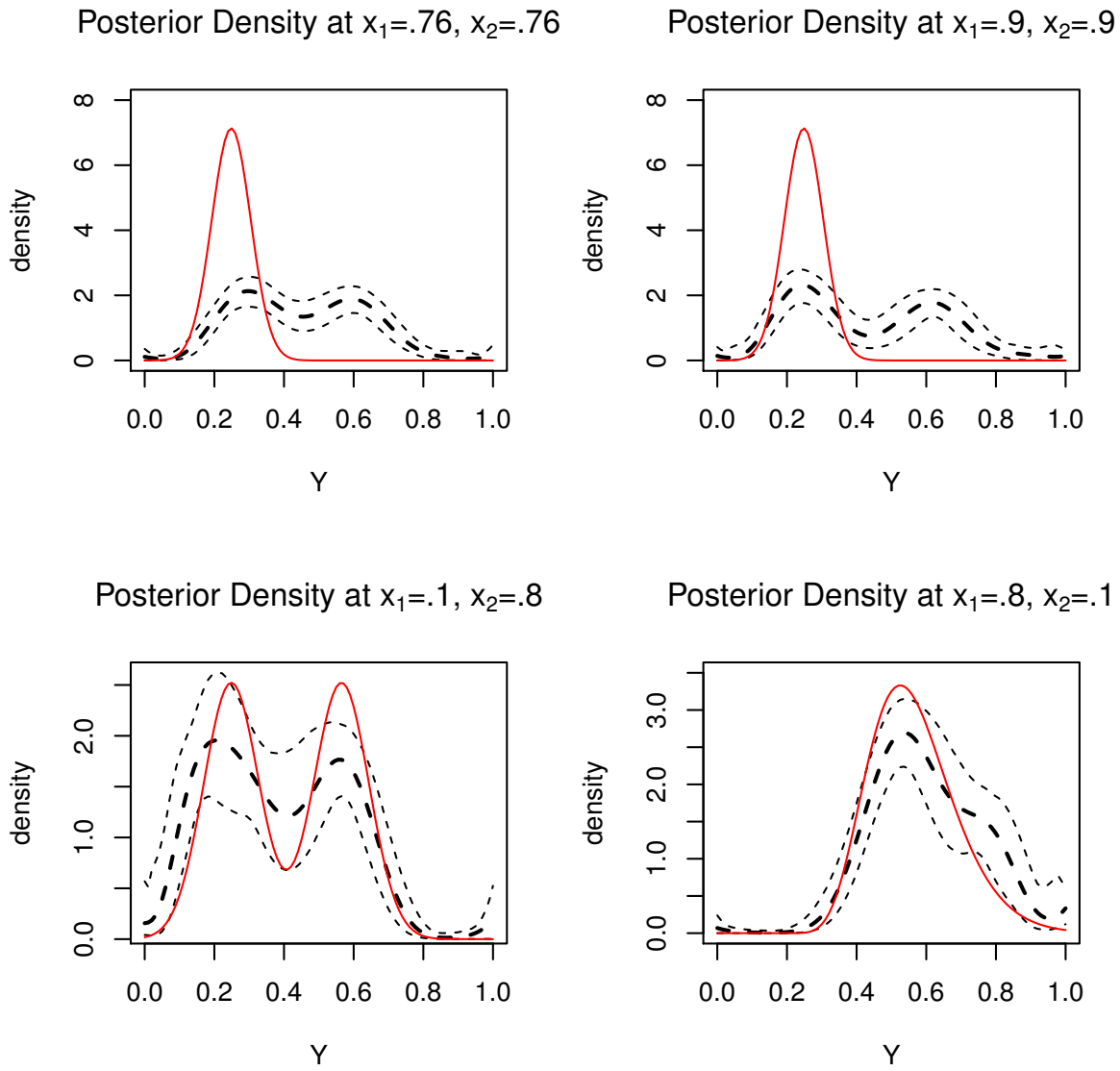
Comparison with Existing methods

- Linear Dependent Tail Free processes [Jara and Handson, 2011]
- Bernstein Polynomials [Barrientos et al. 2017]
- Dirichlet Process Mixtures of Normals [Muller et al, 1996]
- **They assume that the density of y changes smoothly as a function of the covariates**.
- We also compare with Gaussian Partition Model [Denison et al.,2002] which allows discontinuity but restricted to Gaussian distribution.

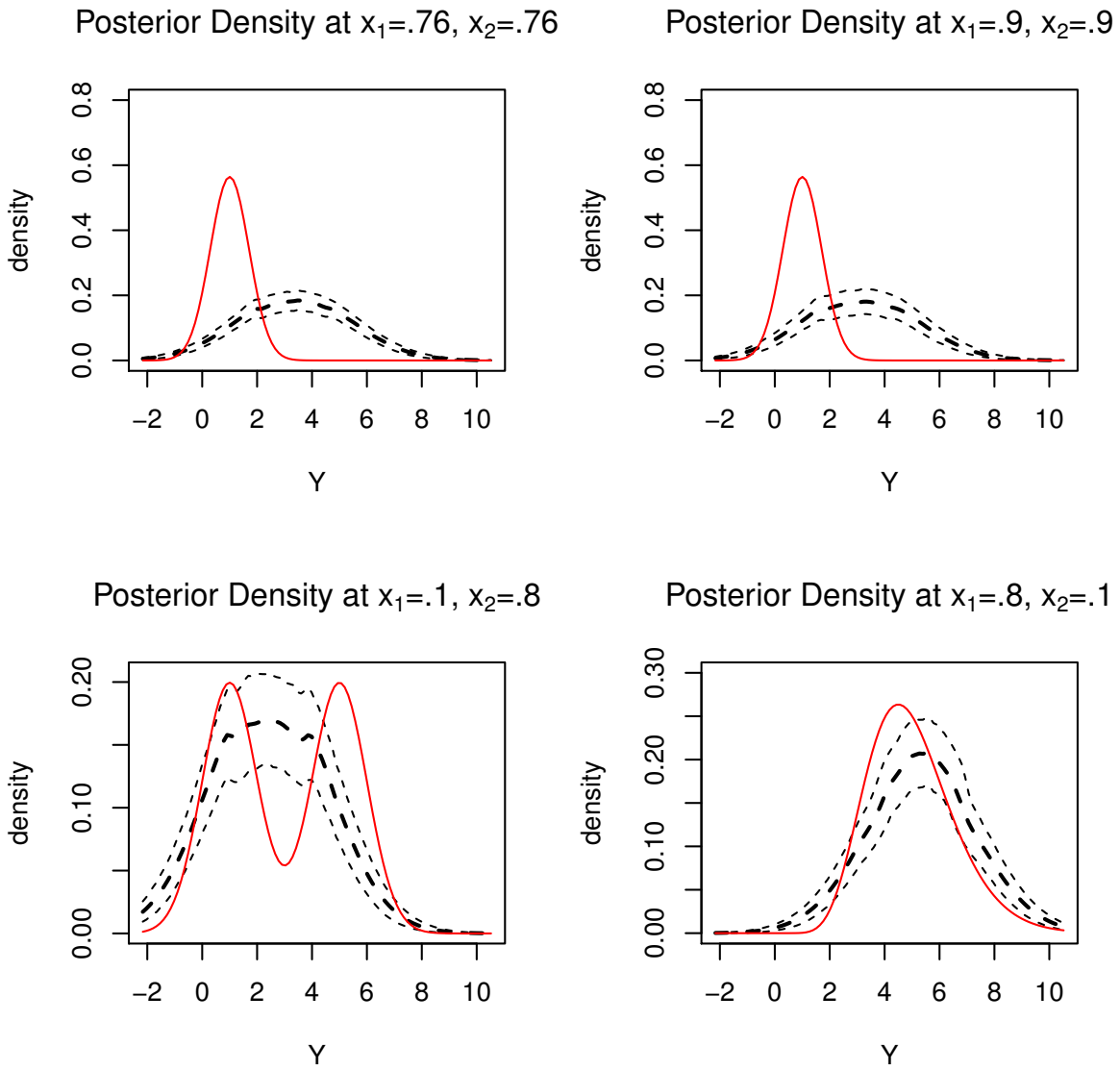
Comparison with Existing methods

- We evaluate posterior density at four different covariate locations including one on the boundary
- None of them can compete with our partition model
- Dirichlet Process mixture of Normals performed the best among others
- That methods failed at the boundary
- Adequate in the middle though with wide credible intervals at the away-from-boundary points

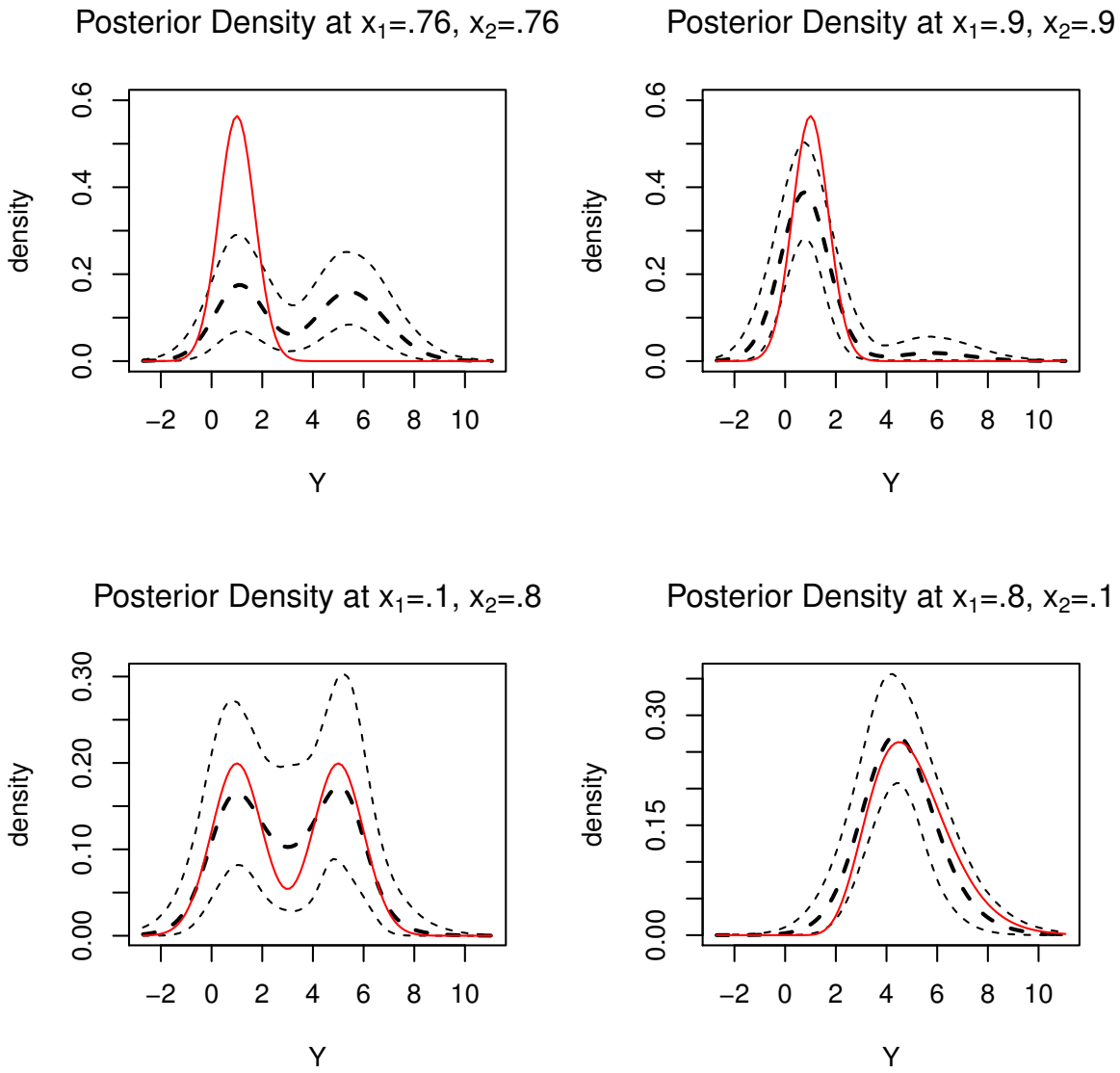
Dependent Bernstein Polynomials



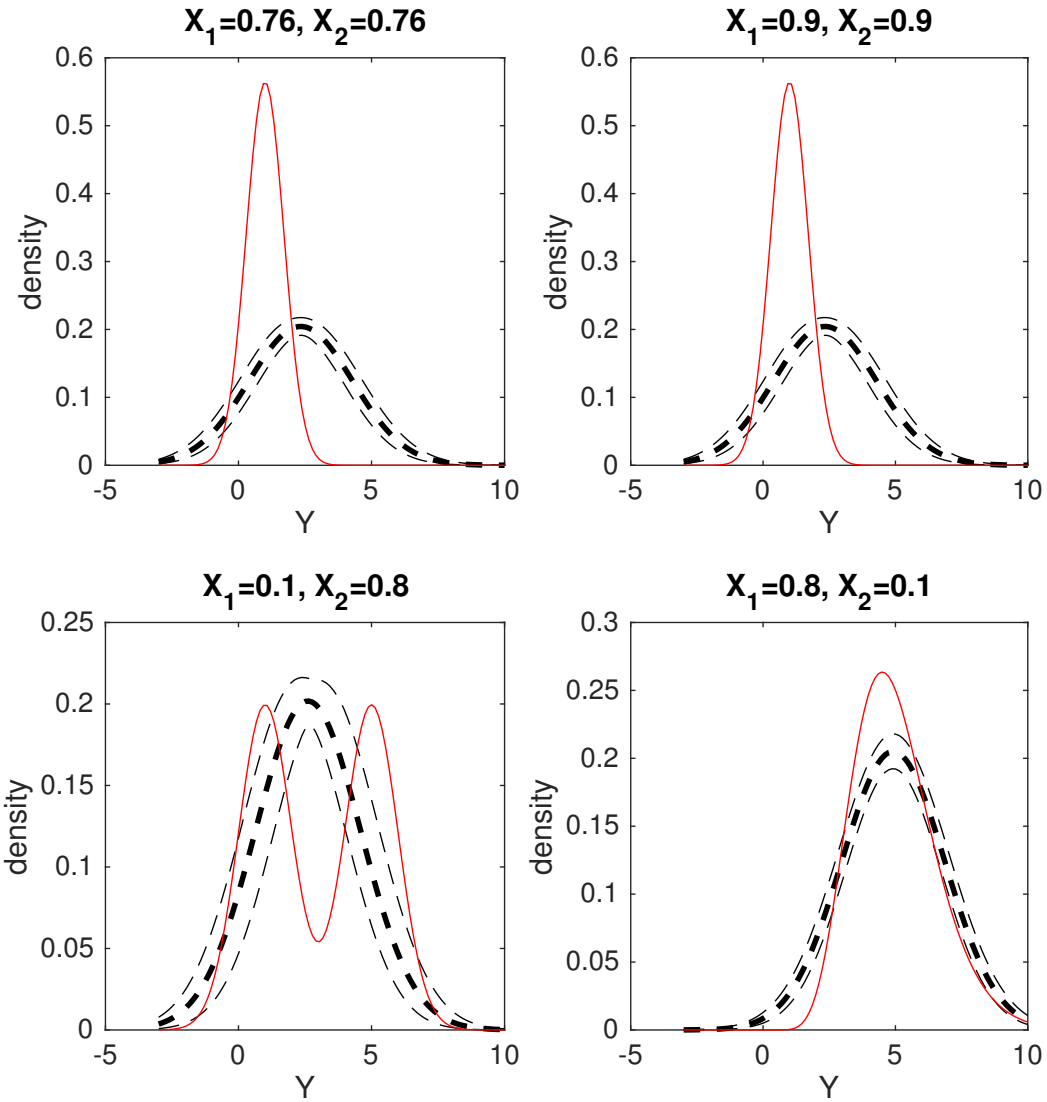
Linear Dependent Tail-free Processes



Dirichlet Process Mixture

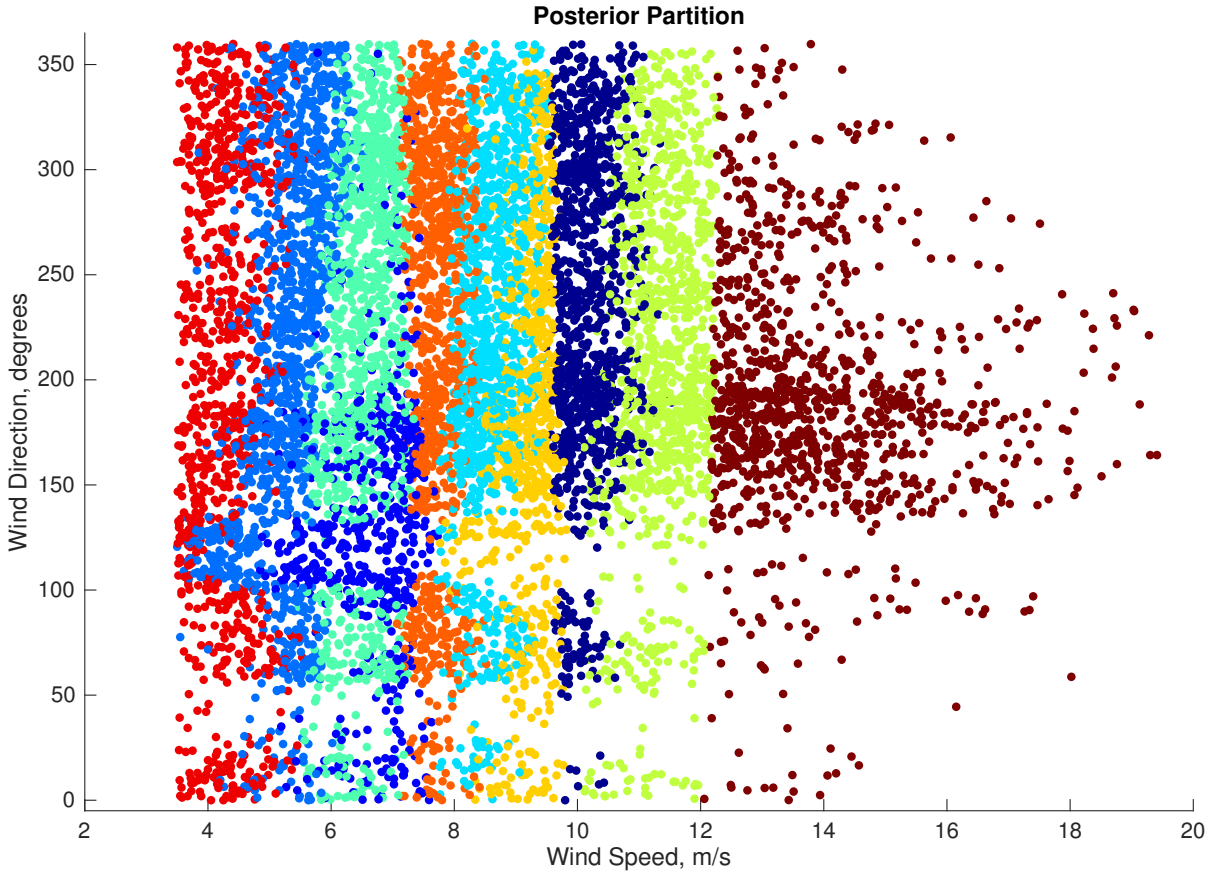


Normal Partition Model

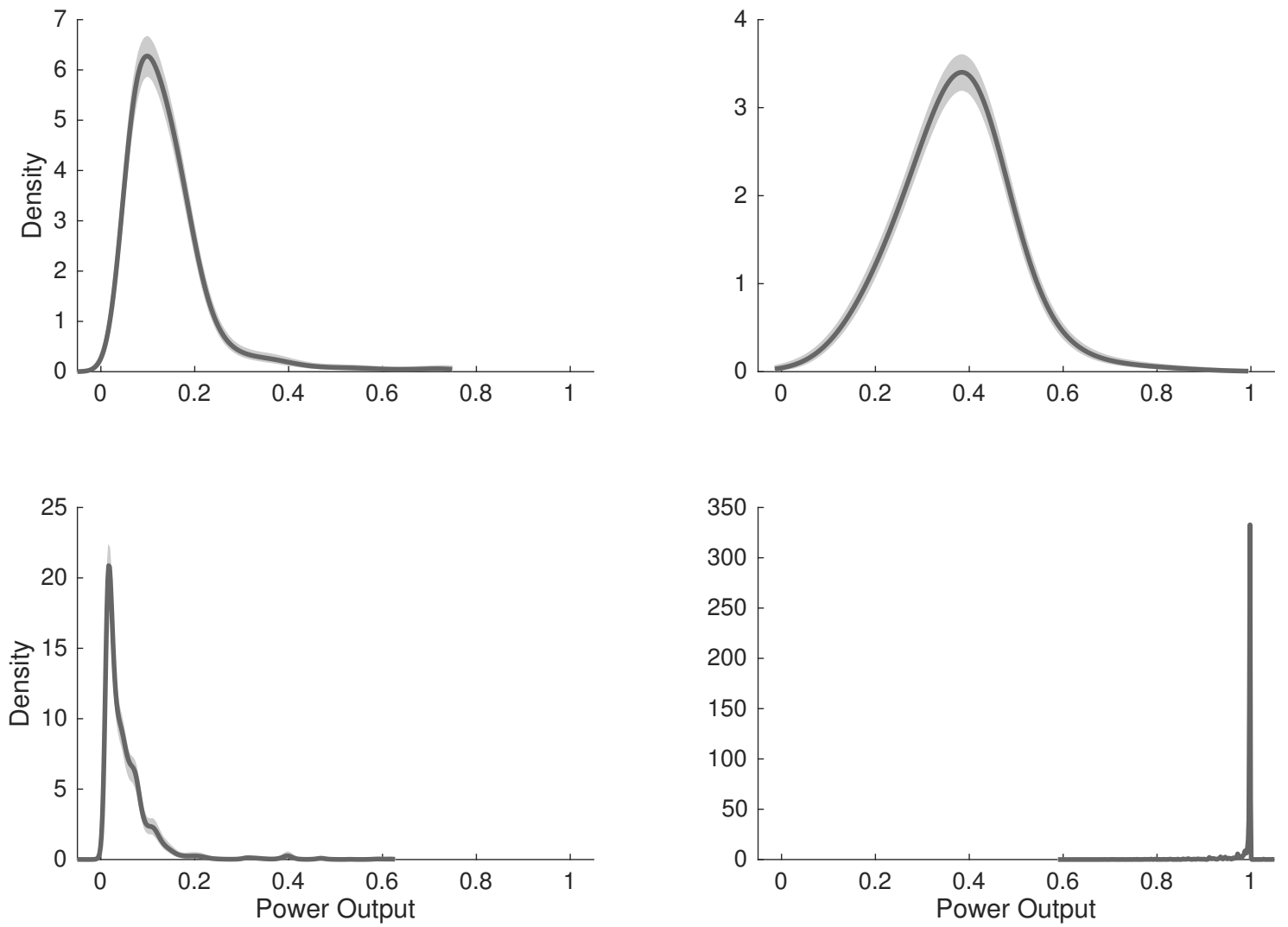


Windmill Data

- 10,000 observations
- Variables
  - wind speed
  - wind direction
  - air density
  - wind sheer
  - turbulence intensity



- Wind speed describes the majority of the changes power output
- The dramatic change in power output at about 120 degrees is believed to be caused by a wake effect from another turbine
- Successfully captures the wind speed at which the maximum power output is achieved at 12m/s


Density
Power Output
Power Output

Discussion

- Strengths
  - Provides interpretation
  - Flexible density estimation
  - Easy to tune MCMC

Discussion

- Strengths
  - Provides interpretation
  - Flexible density estimation
  - Easy to tune MCMC
- Limitations
  - Computation
  - Interpretation & MCMC become more difficult as dimension increases

Discussion

- Strengths
  - Provides interpretation
  - Flexible density estimation
  - Easy to tune MCMC
- Limitations
  - Computation
  - Interpretation & MCMC become more difficult as dimension increases
- Future Work / Work in Progress
  - Implement using a tree partition (simpler interpretation, better mixing)
  - Allow covariates to influence density within each partition element

Survival Analysis

A similar partition model can be developed for survival analysis.

Survival Analysis

A similar partition model can be developed for survival analysis.

- Hazard function is assumed to be piecewise constant
- Log-Hazard function if modeled by a Gaussian process
- Empirical Bayes is used to choose hyperparameters

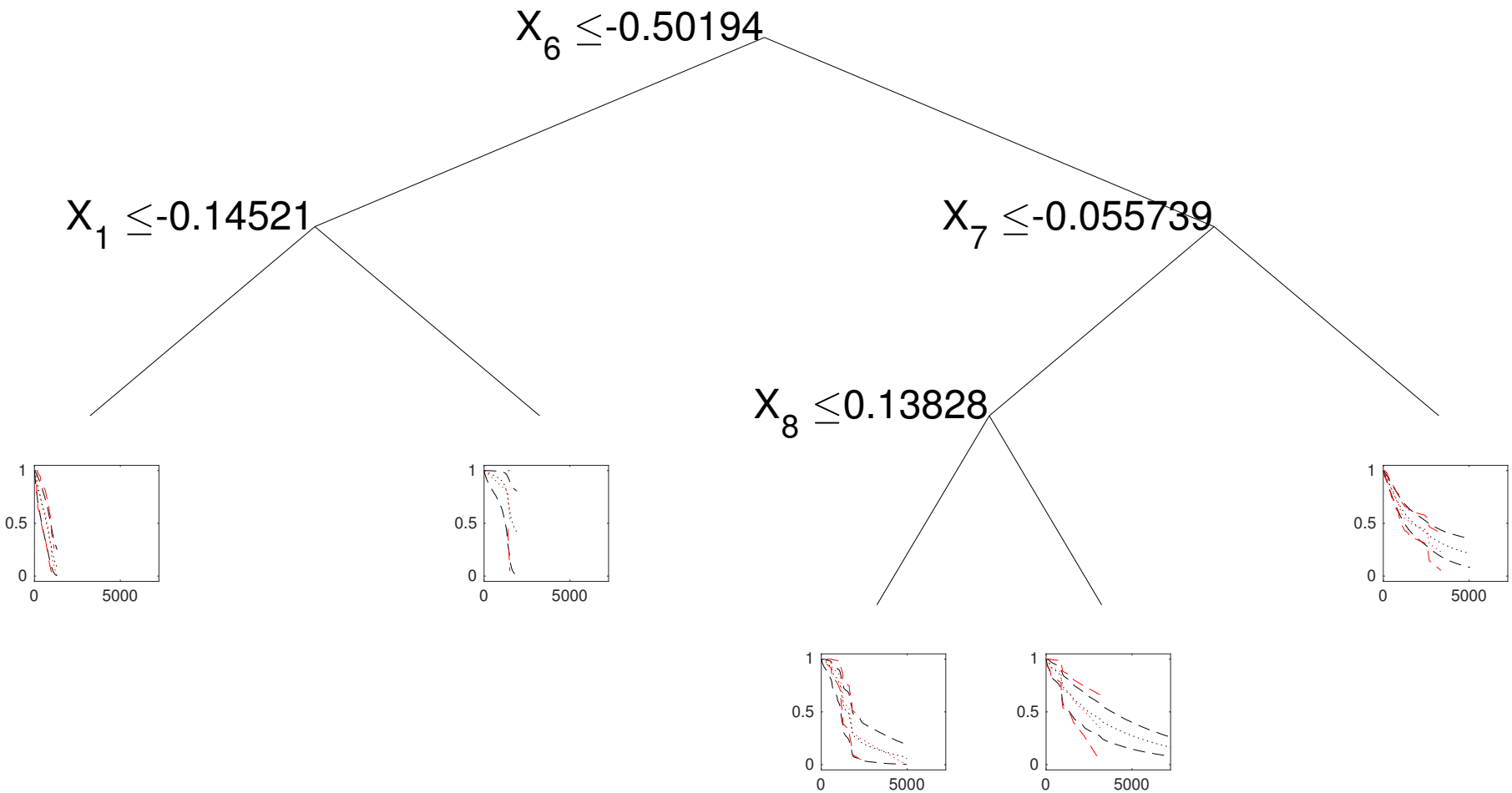
Survival Analysis

A similar partition model can be developed for survival analysis.

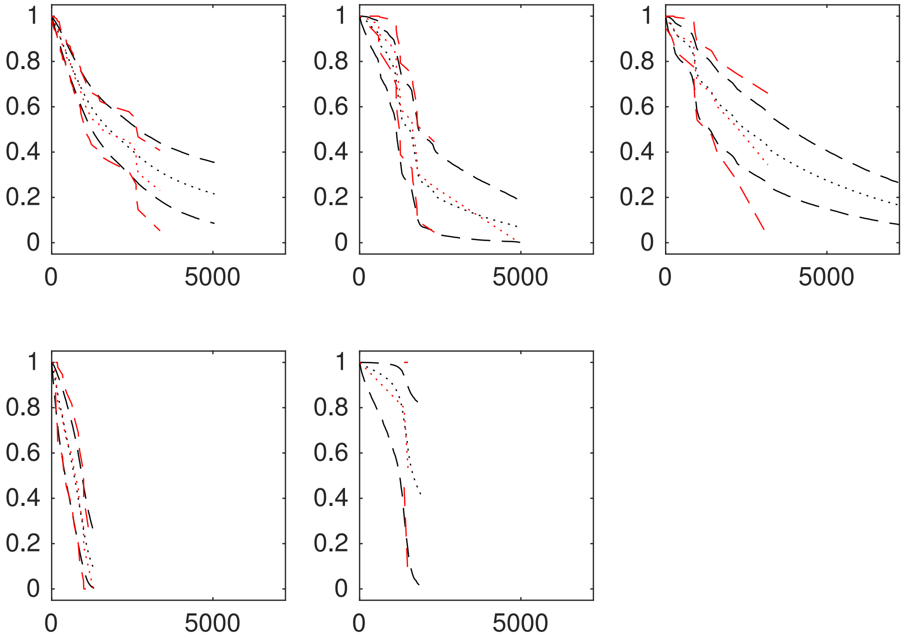
- Hazard function is assumed to be piecewise constant
- Log-Hazard function if modeled by a Gaussian process
- Empirical Bayes is used to choose hyperparameters
- A tree partitioning is used
- Laplace approximations are used to approximate $p(\mathbf{y}_i)$
- Reversible jump MCMC and parallel tempering

Lung Cancer Data

- Dependent variable is time until death (or censoring)
- Covariates are various protein expression levels
- $X_1 : CKIT, X_2 : GSK3ALPHABETA, X_3 : IGFBP2, X_4 : NOTCH1, X_5 : PI3KP110ALPHA, X_6 : RAB25, X_7 : TFRC, X_8 : MEK1, X_9 : pS217S221, X_{10} : NDRG1pT346$
- 357 patients
- Large number censoring


$X_6 \leq$-0.50194
$X_1 \leq$-0.14521
$X_7 \leq$-0.055739
$X_8 \leq$0.13828
1
0.5
0
0
5000

- Several different survival curves of various shapes
- Including groups with relatively short survival times and groups tend to survive longer
- For example tree suggests: subjects with Low expression levels of $X_1 : CKIT$ and $X_6 : RAB25$ have relatively short survival time
- Higher level of $X_6 : RAB25$ and $X_7 : TFRC$ generally lead to longer survival time


1
0.8
0.6
0.4
0.2
0
0
5000

HAPPY BIRTHDAY LYNN AND
MANY MORE TO COME!!!!

REFERENCES I

Barrientos, A. F., A. Jara, and F. A. Quintana (2017). Fully nonparametric regression for bounded data using dependent Bernstein polynomials. *Journal of the American Statistical Association 112*(518), 806–825.

Bhattacharya, A. and D. B. Dunson (2010). Nonparametric bayesian density estimation on manifolds with applications to planar shapes. *Biometrika 97*(4), 851–865.

Chung, Y. and D. B. Dunson (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association 104*(488), 1646–1660.

Denison, D. G., C. Holmes, B. Mallick, and A. Smith (2002). *Bayesian methods for nonlinear classification and regression*. Chichester, United Kingdom: John Wiley & Sons.

REFERENCES II

Dunson, D. B. and J.-H. Park (2008). Kernel stick-breaking processes. *Biometrika 95*(2), 307–323.

Dunson, D. B., N. Pillai, and J. Park (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 163–183.

Fan, J., Q. Yao, and H. Tong (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika 83*(1), 189–206.

Fu, G., F. Y. Shih, and H. Wang (2011). A kernel-based parametric method for conditional density estimation. *Pattern recognition 44*(2), 284–294.

REFERENCES III

Griffin, J. E. and M. J. Steel (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association 101*(473), 179–194.

Jacobs, R. A., M. I. Jordan, S. J. Nowlan, and G. E. Hinton (1991). Adaptive mixtures of local experts. *Neural Computation 3*(1), 79–87.

Jara, A. and T. E. Hanson (2011). A class of mixtures of dependent tail-free processes. *Biometrika 98*(3), 553–566.

Kooperberg, C. and C. J. Stone (1991). A study of logspline density estimation. *Computational Statistics & Data Analysis 12*(3), 327–347.

Kundu, S. and D. B. Dunson (2011). Single factor transformation priors for density regression. *arXiv preprint arXiv:1108.2720*.

REFERENCES IV

Müller, P., A. Erkanli, and M. West (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika 83*(1), 67–79.

Stone, C. J., M. H. Hansen, C. Kooperberg, Y. K. Truong, et al. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 Wald memorial lecture. *The Annals of Statistics 25*(4), 1371–1470.

Tokdar, S. T., Y. M. Zhu, J. K. Ghosh, et al. (2010). Bayesian density regression with logistic Gaussian process and subspace projection. *Bayesian analysis 5*(2), 319–344.