Restricted random partitions for Bayesian curve fitting

Sonia Petrone

Bocconi University, Milano

joint work with Sara Wade (University of Cambridge) and Stephen Walker (University of Kent)

Paris, 14-15 May 2013

the general problem

When dealing with statistical models that involve a huge-dimensional space, it is convenient or necessary to

- reduce the dimension by setting at zero the probability of appropriate subspaces.
- At the same time, preserve desirable properties of the involved probability laws.

the general problem

When dealing with statistical models that involve a huge-dimensional space, it is convenient or necessary to

- reduce the dimension by setting at zero the probability of appropriate subspaces.
- At the same time, preserve desirable properties of the involved probability laws.

Here, the large-dimensional space is the space of all partitions of $\{1, 2, \ldots, n\}$. Random partitions appear in many contexts, combinatorics, genetics, clustering, nonparametric inference ...

We will illustrate the issue by using a simple problem of Bayesian nonparametric curve fitting through Dirichlet process mixtures.

Outline

1. general problem: introducing restrictions on random partitions

2. illustration: Bayesian curve fitting
   - Dirichlet process-based mixture models
   - Restricted DPM

3. Examples

Random partitions

Random partitions appear in many contexts, combinatorics, genetics, clustering, nonparametric inference, ...
Partition $\rho_n = (s_1, \ldots, s_n)$ of $\{1, 2, \ldots, n\}$
e.g., $n = 6$, $(s_1, \ldots, s_n) = (1, 1, 2, 1, 3, 2)$ defines the partition
$\rho_n = ((1, 2, 4), (3, 6), (5))$

In many problems we can assign $p(x_{1:n} \mid \rho_n)$.
Bayesian inference on ρ_n gives a prior $\pi(\rho_n)$, which combined with the likelihood by Bayes rule, gives the posterior

$$\pi(\rho_n \mid x_{1:n}) \propto p(x_{1:n} \mid \rho_n)\, \pi(\rho_n).$$

Yet, often the likelihood is intractable; and computing the normalizing constant involves a huge sum. Thus, explore the partition space $\mathcal{P}_n$ e.g. via MCMC.

Problems

- computations. Partition space $\mathcal{P}_n$ is huge, thus $\pi(\rho_n \mid x_{1:n})$ can be very spread out, and MCMC can visit only a subset of partitions, and each partition only once..

- methodological. For modeling reasons, one may want to incorporate information on what are desirable and undesirable partitions. But since $\mathcal{P}_n$ is huge, this info will be dramatically diluted in the prior and in the posterior.

example

In the example of curve fitting that we will discuss, one wants to introduce an ordering constraint.

- Number of ways to partition the n subjects:

$$B_n = \sum_{k=1}^{n} S_{n,k}, \quad \text{a Bell number}$$

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n \quad \text{a Stirling number of the second kind.}$$

- Under a ordering constraint, number of ways to partition the n subjects:

$$\sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1}.$$

- Example: if n=10, $B_{10} = 115,975$ and $2^{n-1} = 512$, 0.44% of the total partitions, and if $n = 100$, the percentage of partitions under this constraint is less than $10^{-83}\%$ of the total partitions.

our proposal

- Strictly restrict the partition space by putting the probability of undesirable partitions to zero.
- the problem is not so trivial as in restricting the partition space we may introduce undesirable bias in the prior on ρ_n. Thus, we want to restrict but at the same time preserve good properties of the prior.

our proposal

- Strictly restrict the partition space by putting the probability of undesirable partitions to zero.
- the problem is not so trivial as in restricting the partition space we may introduce undesirable bias in the prior on ρ_n. Thus, we want to restrict but at the same time preserve good properties of the prior.

We illustrate these issues in Bayesian curve fitting.

2. Bayesian curve fitting

Given

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where generally $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$, the problem is estimating the unknown curve $m(x)$.

In a Bayesian approach, this is solved by giving a prior on $m(\cdot)$; then the curve estimate at x_{n+1} w.r.t. quadratic loss corresponds to the point prediction at x_{n+1}

$$\hat{m}(x_{n+1}) = E(Y_{n+1} \mid x_{1:n}, y_{1:n}, x_{n+1}).$$

Popular priors are based on Dirichlet process mixtures. But we find that these models may have poor prediction properties and restrictions on the involved partition of subjects into mixture's components are necessary.

Dirichlet-based mixture models

The general form of a Dirichlet-based mixture model for regression is

$$Y|x, w, \mu, \sigma^2 \overset{indep}{\sim} \sum_{j=1}^{\infty} w_j(x)\ \mathrm{N}(\mu_j(x), \sigma_j^2(x))$$

This implies a flexible expression for $m(\cdot)$

$$m(x) = \mathrm{E}[Y|x, w, \mu, \sigma^2] = \sum_{j=1}^{\infty} w_j(x)\ \mu_j(x),$$

where $m(x)$ is modeled through

- a collection of curves (basis functions) $\mu_j(x)$, $j = 1, 2, \ldots$
- weights $w_j(x)$ that globally (if $w_j(x) \equiv w_j$) or locally (if $w_j(x)$ depends on x) select the curve μ_j "in use".

It is natural to ask for simple basis functions μ_j and local selection of μ_j (somehow such as for splines with random knots). Yet the most popular BNP models do not achieve this.

Dirichlet process-based mixture models

- Dirichlet process mixture models (DPM) assume linear 'basis functions' $\mu_j(x) = \beta_j^* x$ and constant weights w_j. Thus

$$m(x) = \mathrm{E}[Y|x, w, \mu, \sigma^2] = \sum_{j=1}^{\infty} w_j \, \beta_j^* x.$$

  This global selection of the μ_j may lead to poor prediction.

- joint DPM imply covariate dependent weights $w_j(x)$. This improves prediction, that is however still obtained by unnecessarily averaging on undesirable partitions.

We propose a restricted partition prior, that can improve computations and prediction.

Dirichlet process mixture models

A DPM is defined as

$$Y_i|x_i, \beta_i, \sigma_i^2 \overset{indep}{\sim} N(\beta_i' \underline{x}_i, \sigma_i^2),$$
$$\beta_i, \sigma_i^2|P \overset{i.i.d}{\sim} P,$$
$$\mathbf{P} \sim DP(\alpha P_0).$$

base measure P_0: conjugate Normal-Inverse Gamma prior(β_0, C^{-1}, a, b).

From well known properties of the DP, P is discrete with probability one, and integrating the parameters out

$$Y_i|x_i, w, \mu, \sigma^2 \overset{indep}{\sim} \sum_{j=1}^{\infty} w_j(x_i)\ \mathrm{N}(\beta_j^* x_i, \sigma_j^2(x_i)),$$

thus

$$m(x) = \mathrm{E}[Y|x, w, \mu, \sigma^2] = \sum_{j=1}^{\infty} w_j\ \beta_j^* x.$$

random partition

From the Pólya sequence scheme characterization of the DP, the model induces a random partition of the individual coefficients $(\beta_1, \sigma_i^*), i = 1, \ldots, n$, defined as $\rho_n = (s_1, \ldots, s_n)$ where $s_1 = 1$; $s_2 = 1$ if $\beta_2 = \beta_1 (= \beta_1^*)$ or $s_2 = 2$ if β_2 is a new value β_2^*, and so on.

The prior probability of a partition in k clusters of sizes $n_1, \ldots, n_k$ results

$$p(\rho_n) = \frac{\alpha^k}{\alpha^{[n]}} \prod_{j=1}^{k} (n_j - 1)!,$$

where $\alpha^{[n]} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$. It does not depend on $x_{1:n}$.

Given the partition, clusters are assumed independent and simple linear regression is used inside clusters:

$$Y_1, \ldots Y_n \mid x_{1:n}, \rho_n \sim \prod_{j=1}^{k} \prod_{i:s_i=j} \mathrm{N}(y_i \mid \beta_j^* x_i, \sigma_i^2);$$

Inference and prediction

- Posterior of ρ_n:

$$p(\rho_n|y,x) \propto \alpha^k \prod_{j=1}^{k}(n_j-1)! \prod_{j=1}^{k} p(y_i : i \in S_j | x_i : i \in S_j),$$

prior $\times$ the product of marginal likelihood of obs in cluster S_j.

- Given the partition, clusters are independent, so that conditional prediction of Y at x_{n+1} is

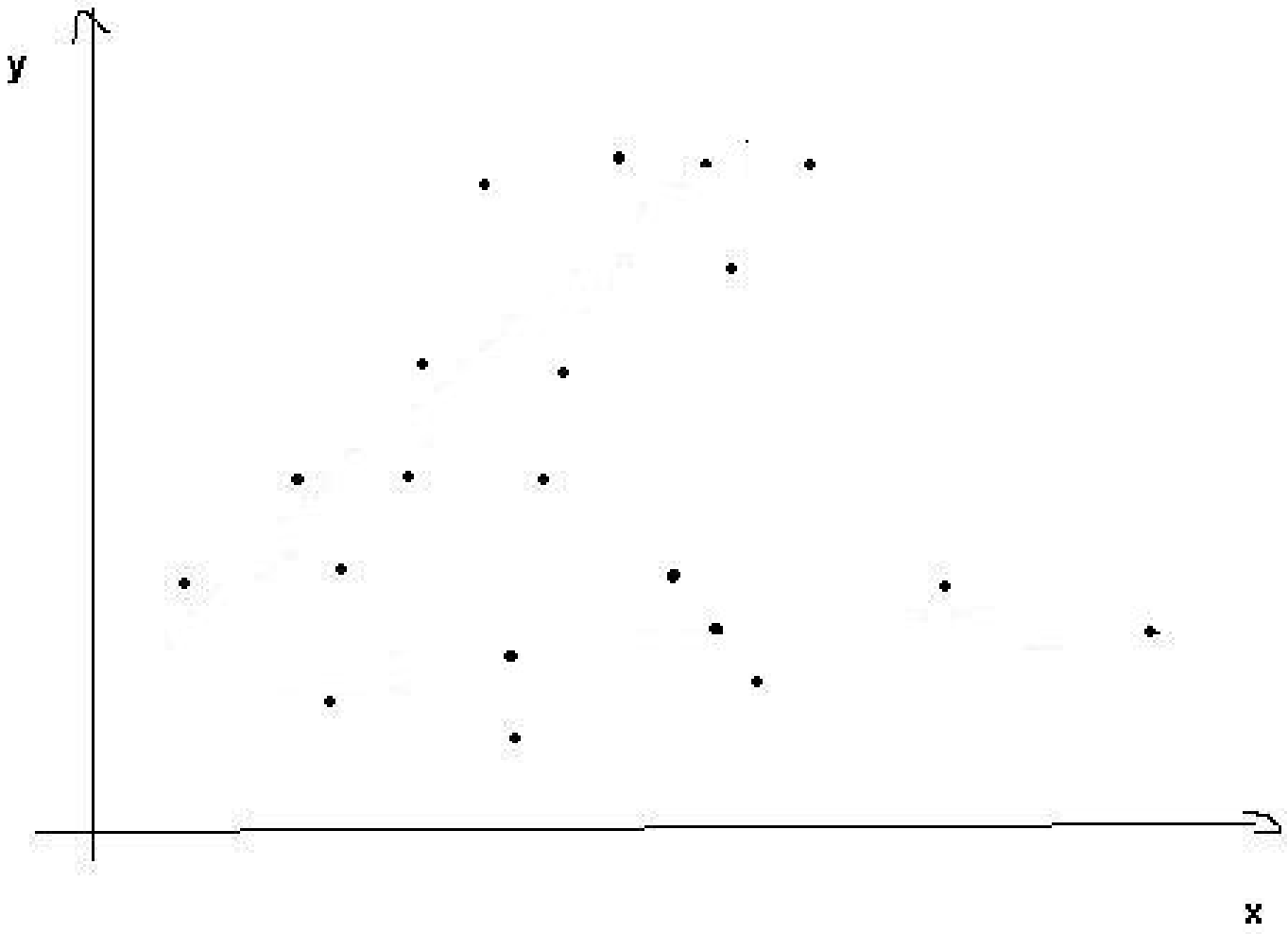
$$\mathrm{E}[Y|x_{n+1},y,x,\rho_n] = \sum_{s_{n+1}=1}^{k+1} E[Y_{n+1}|x_{n+1},y,x,\rho_n,s_{n+1}]p(s_{n+1} \mid x,y,x_{n+1},\rho_n$$

$$= \frac{\alpha}{\alpha+n}\beta_0'\underline{x}_{n+1} + \sum_{j=1}^{k}\frac{n_j}{\alpha+n}\widehat{\beta}_j'\underline{x}_{n+1},$$

where $\widehat{\beta}_j = E(\beta_j^* \mid x,y,\rho_n) = (C + X_j'X_j)^{-1}(C\beta_0 + X_j'\underline{y}_j)$. Note that $p(\rho_n|y,x)$ doesn't depend on x_{n+1}: the prediction (curve estimate) is linear.

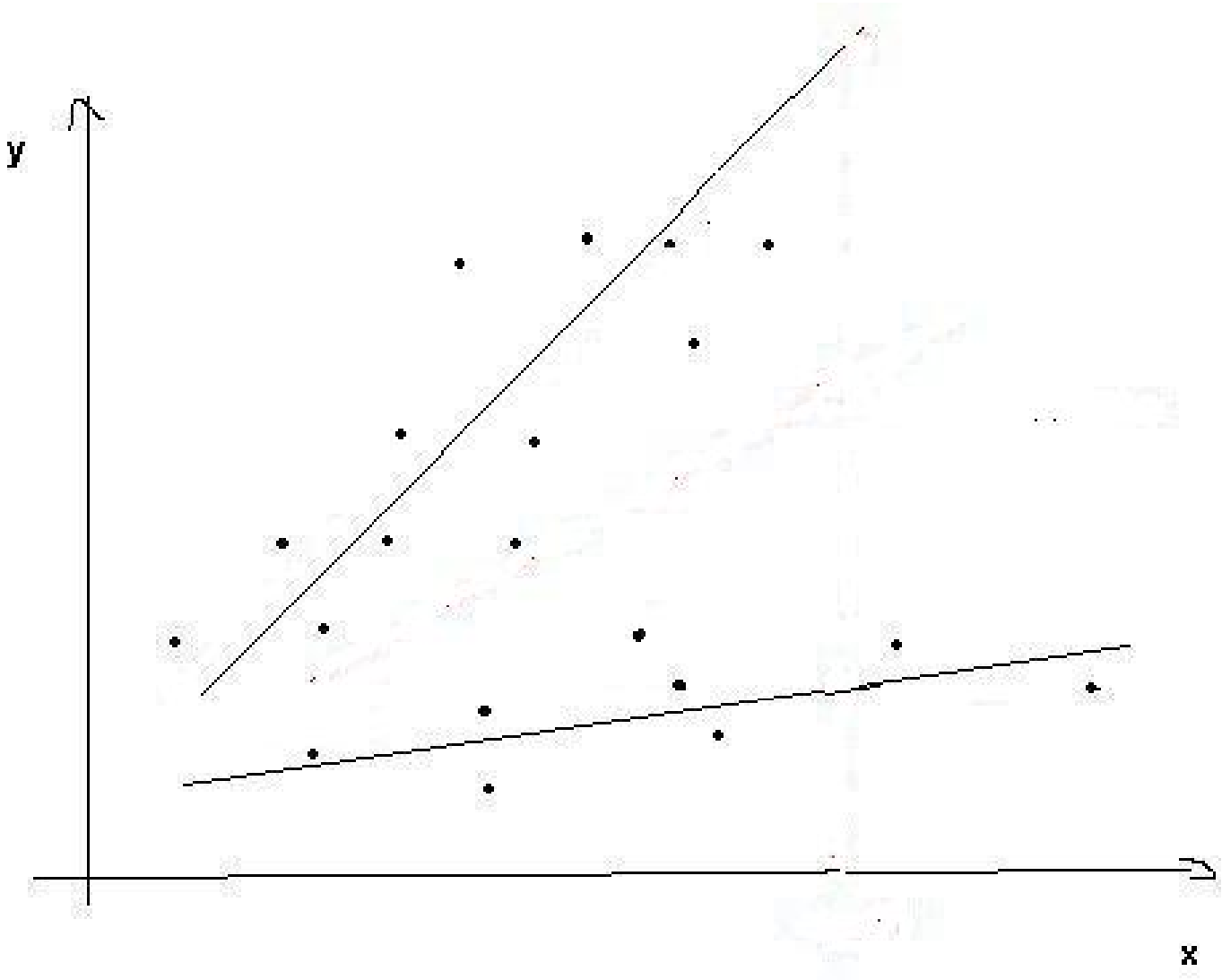
DP mixture of linear regression

more an exploratory tool



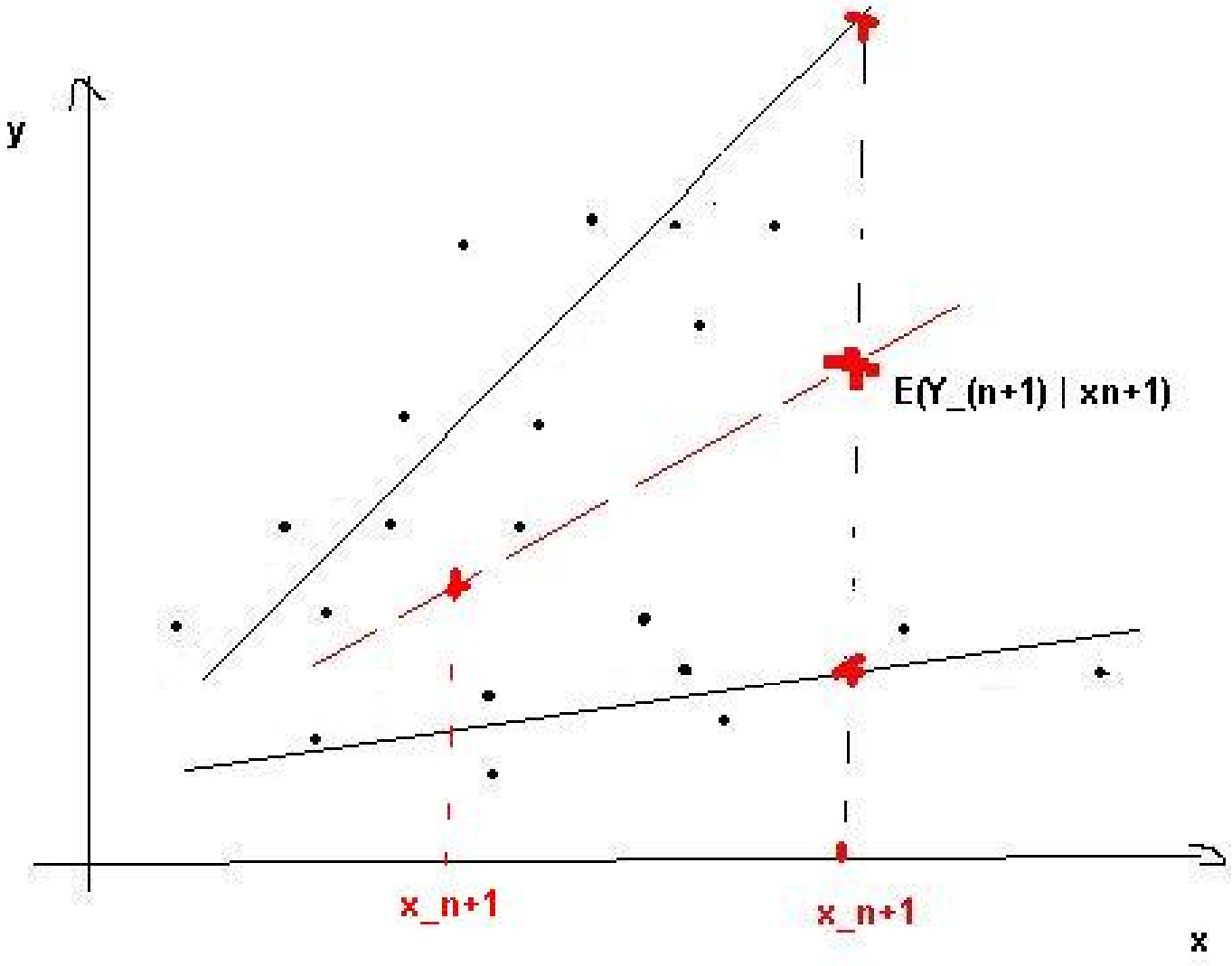
DP mixture of linear regression

more an exploratory tool

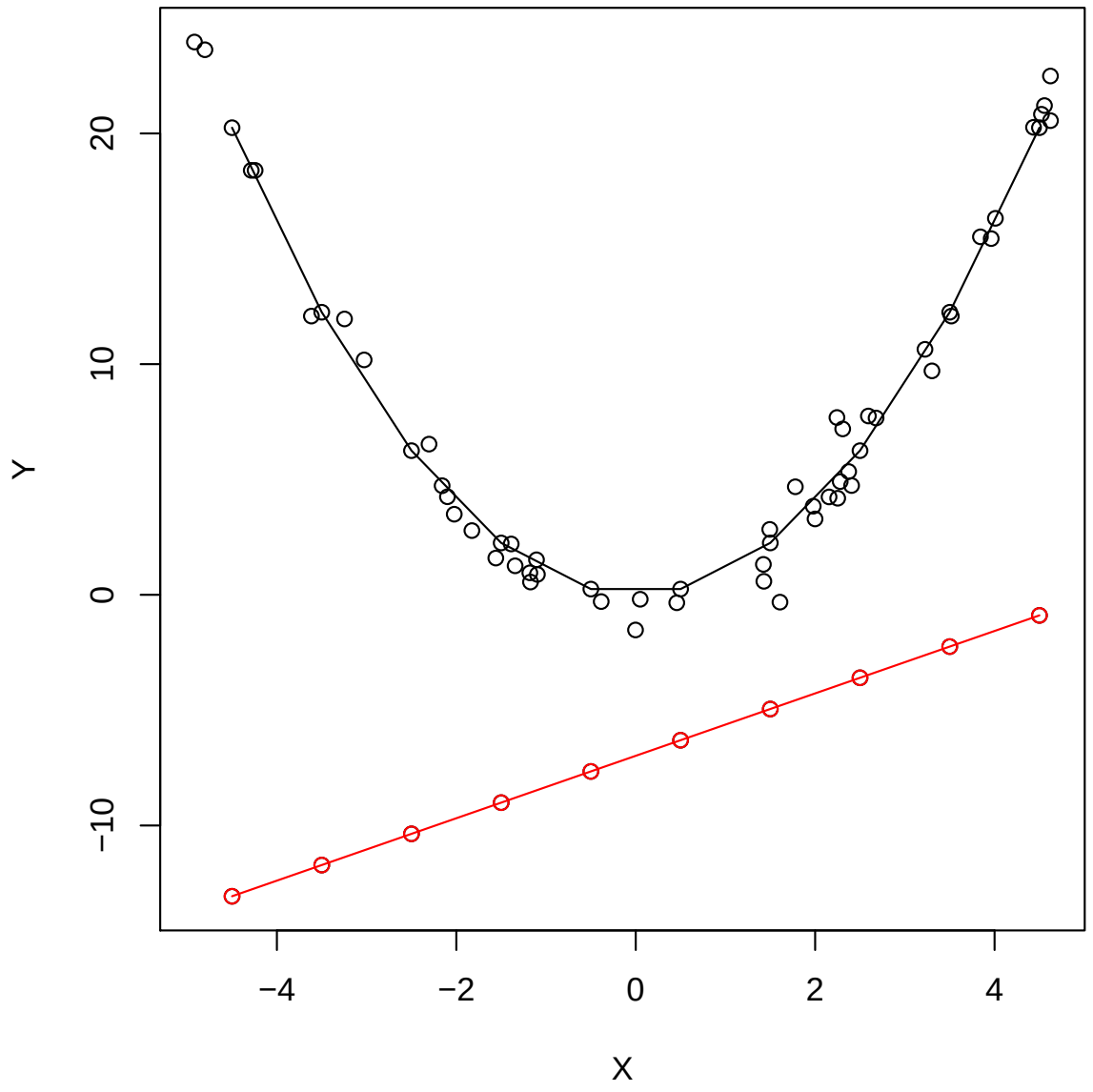


DP mixture of linear regression

but prediction generally uninformative

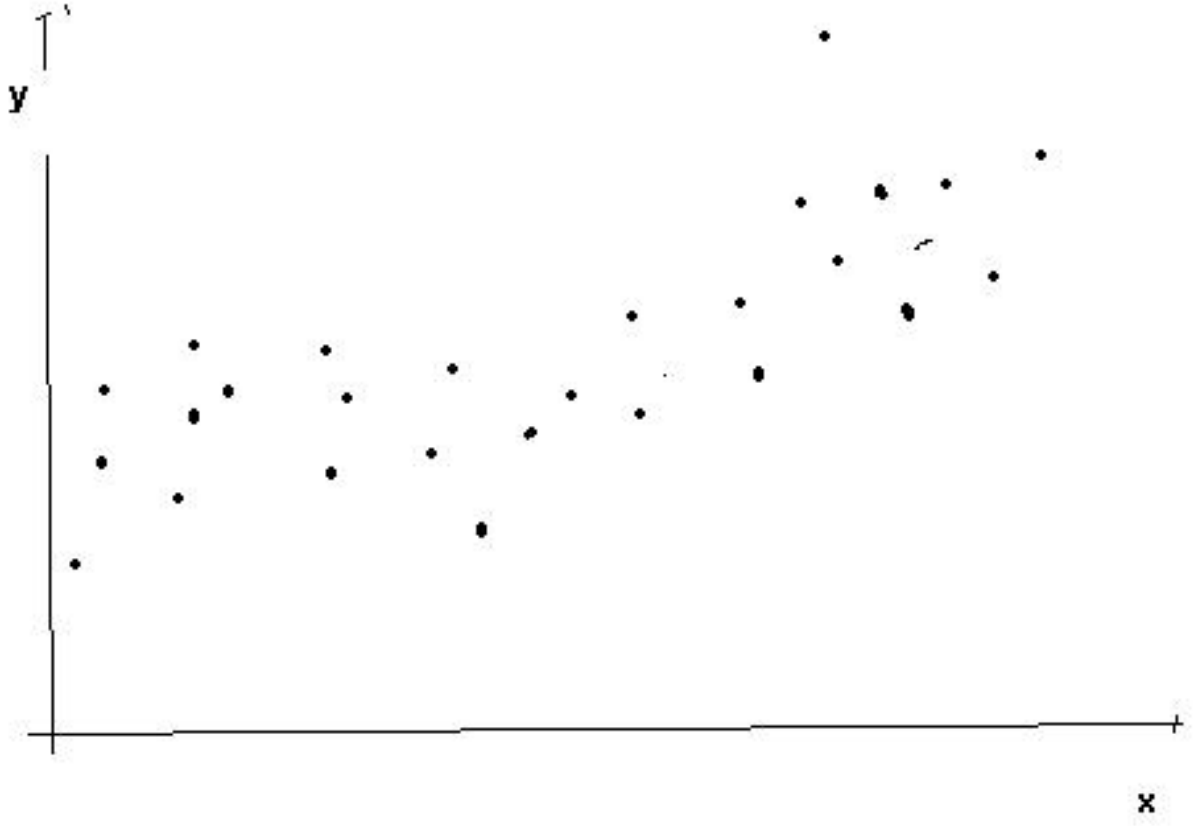


a more dramatic example



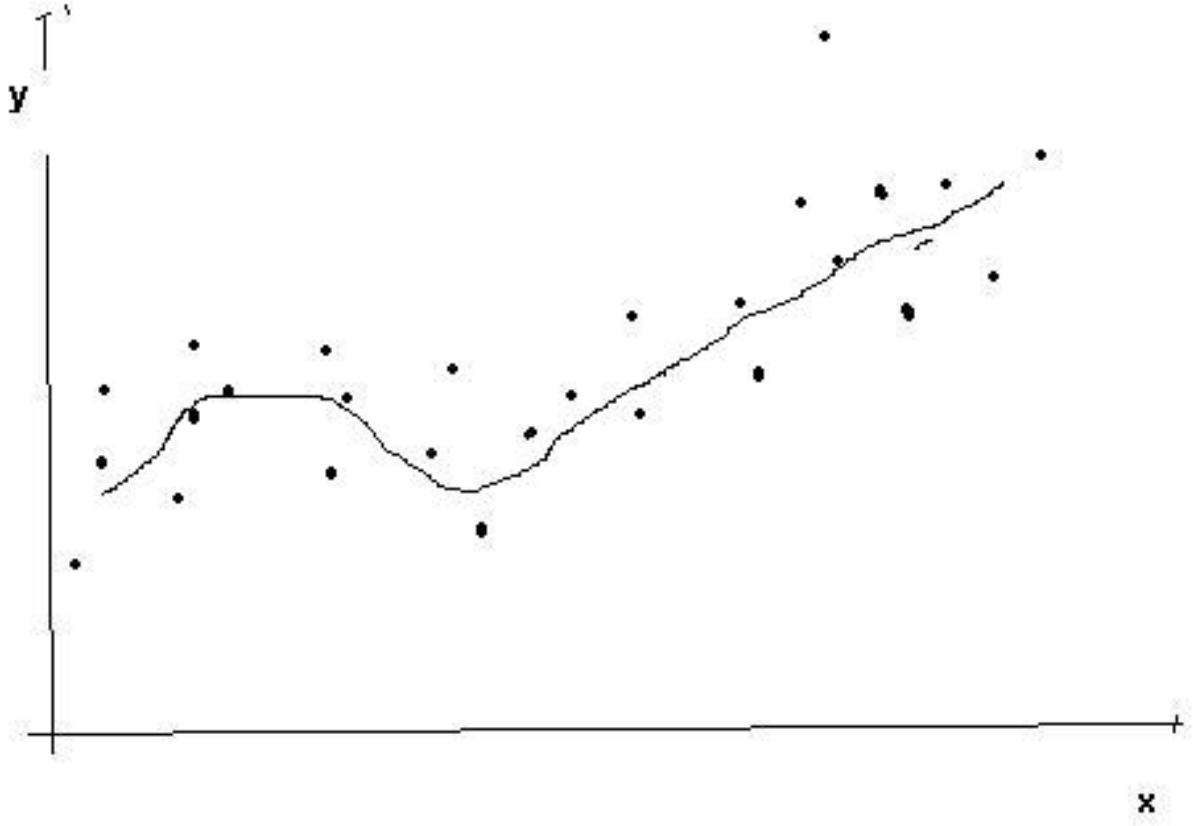
BNP regression

we exclude multiple behavior for a given x



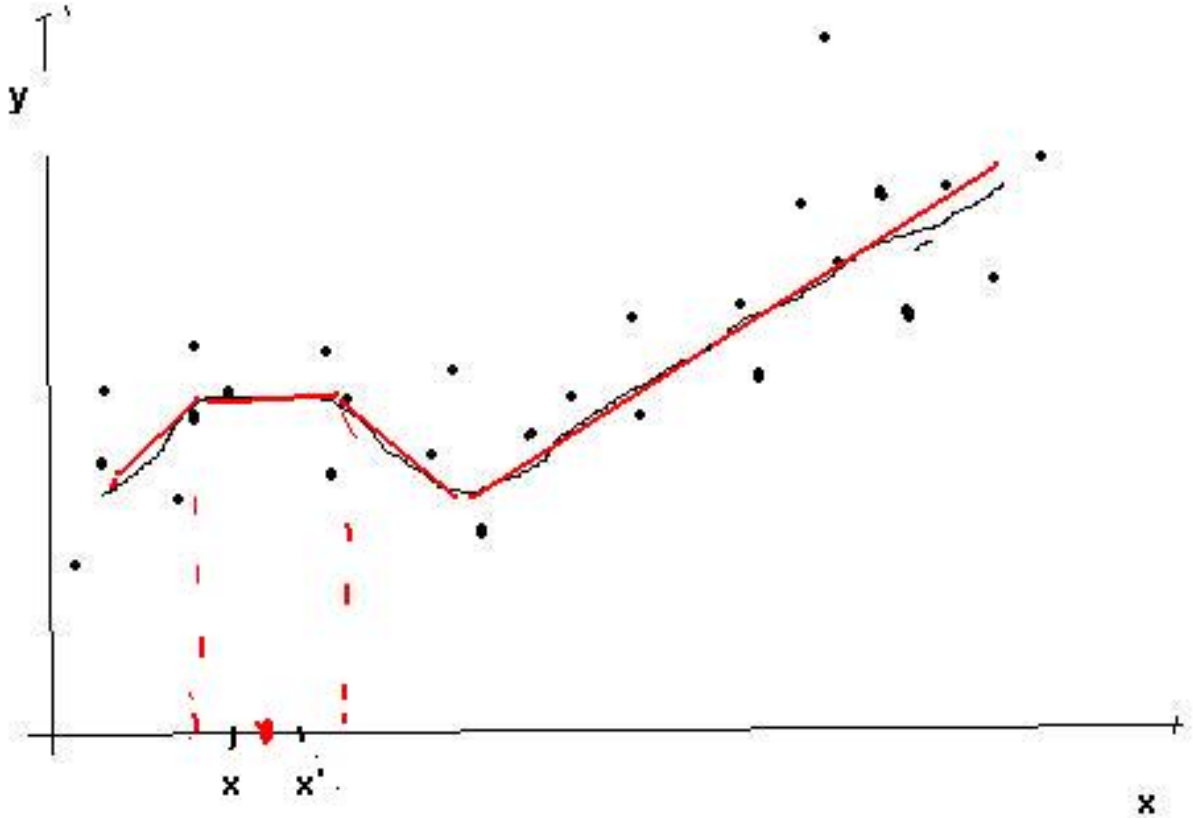
BNP regression

DP and 'clustering' is used for flexible regression



BNP regression

DP and 'clustering' is used for flexible regression



BNP regression

DP and 'clustering' is used for flexible regression:
locally select a curve (here, a line) from the collection of available curves (here, $\beta_i' \underline{x}_i$, $i = 1, 2, \ldots$)

then the partition should depend on x, and if $x \approx x'$, we expect that $E(Y \mid x) \approx E(Y \mid x')$, i.e., same cluster.

BNP regression

DP and 'clustering' is used for flexible regression:
locally select a curve (here, a line) from the collection of available curves (here, $\beta_i' \underline{x}_i$, $i = 1, 2, \ldots$)

then the partition should depend on x, and if $x \approx x'$, we expect that $E(Y \mid x) \approx E(Y \mid x')$, i.e., same cluster.

DDP and joint DP mixtures go in this direction: the random partition depends on covariates x, and the cluster allocation of Y_{n+1} depends on x_{n+1}.

However, we still unnecessarily give positive probability mass to 'bad' partitions, which still affect prediction.

Joint DP mixtures

- Model (Müller et al.(1996)):

$$
\begin{aligned}
Y_i | x_i, \beta_i, \sigma_i^2 &\overset{indep}{\sim} N(\beta_i' \underline{x}_i, \sigma_i^2), \\
X_i, | \mu_i, \Sigma_i &\overset{indep}{\sim} N(\mu_i, \Sigma_i), \\
\beta_i, \sigma_i^2, \mu_i, \Sigma_i | P &\overset{i.i.d}{\sim} P, \\
\mathbf{P} &\sim DP(\alpha P_{0Y} \times P_{0X}).
\end{aligned}
$$

$$
\Rightarrow \quad Y_i | x_i, w, \beta, \sigma^2 \overset{indep}{\sim} \sum_{j=1}^{\infty} w_j(x_i) N(\beta_j^* \underline{x}_i, \sigma_j^{2*}),
$$

where

$$
w_j(x) = \frac{w_j N(x_i; \mu_j^*, \Sigma_j^*)}{\sum_{j=1}^{\infty} w_j N(x_i; \mu_j^*, \Sigma_j^*)}.
$$

Joint DP mixtures: Random partition

- Prior for the random partition now depends on $x_{1:n}$:

$$p(\rho_n|x_{1:n}) \propto \alpha^k \prod_{j=1}^{k} (n_j - 1)!\; p_0(x_i \in S_j),$$

where $p_0(x_i \in S_j) = \int \prod_{i:s_i=j} N(x_i; \mu, \Sigma) dP_{0X}(\mu, \Sigma)$.

- Posterior of ρ_n given $x = x_{1:n}, y = y_{1:n}$:

$$p(\rho_n|y, x) \propto \alpha^k \prod_{j=1}^{k} (n_j - 1)!\; p(x_i \in S_j) p(y_i \in S_j|(x_i \in S_j)).$$

prior $\times$ product of clusters' marginal *joint* likelihoods of (x_i, y_i)

Joint DP mixtures: Prediction

- Prediction of y_{n+1}

$$\mathrm{E}[Y_{n+1}|x_{n+1}, y, x] = \tag{1}$$
$$\sum_{\rho_n \in \mathcal{P}_n} \left(\mathrm{E}[Y_{n+1}|x_{n+1}, y, x, \rho_n] \frac{p(x_{n+1}|\rho_n, y, x)}{p(x_{n+1}|y, x)} \right) p(\rho_n|y, x).$$

- Inner term of (1) is

$$\frac{\alpha}{c} p_0(x_{n+1}) \beta_0' \underline{x}_{n+1} + \sum_{j=1}^{k} \frac{n_j}{c} p_j(x_{n+1}|x_i \in S_j) \widehat{\beta}_j' \underline{x}_{n+1}$$

where $p_j(x_{n+1}|x_i \in S_j) = \int N(x_{n+1}; \mu, \Sigma) dP_{0X}(\mu, \Sigma|x_i \in S_j)$.

This favors allocation of Y_{n+1} in clusters for which the predictive density $p_j(x_{n+1} \mid x_i \in S_j)$ is higher. However, for prediction we still unnecessarily average on 'bad' partitions..

Restricted DP mixtures

- In regression settings, it is reasonable to assume that clusters should be based on proximity of x.
- However, because the total number of partitions is so large, placing higher prior mass on desirable partitions (joint DPM) is not enough to ensure:
  - prominence of these partitions in the posterior,
  - sufficiently small posterior mass for undesirable partitions.
- Only way to ensure this is to place zero mass on undesirable partitions in the prior.

Motivation

- Focus on a particular case: x is one-dimensional and continuous.
- Partitions should be based on the natural ordering of x.
- Number of ways to partition the n subjects:

$$B_n = \sum_{k=1}^{n} S_{n,k}, \quad \text{a Bell number}$$

$$S_{n,k} = \frac{1}{k!} \sum_{j=0}^{k} (-1)^j \binom{k}{j} (k-j)^n \quad \text{a Stirling number of the second kind.}$$

- Under a ordering constraint, number of ways to partition the n subjects:

$$\sum_{k=1}^{n} \binom{n-1}{k-1} = 2^{n-1}.$$

- Example: if n=10, $B_{10} = 115,975$ and $2^{n-1} = 512$, 0.44% of the total partitions, and if $n = 100$, the percentage of partitions under this constraint is less than $10^{-83}\%$ of the total partitions.

Restricted DP: Construction

- Let $x_{(1)} < \cdots < x_{(n)}$ denote the ordered values of x, and $s_{\pi_x(1)}, \ldots, s_{\pi_x(n)}$ the corresponding values of $s_1, \ldots, s_n$.
- Aim: remove partitions that violate the constraint that $s_{(1)} \leq \ldots \leq s_{(n)}$.
- Yet, simply multiply the prior for ρ_n by the indicator for this event does not work. We would remove no partitions for $k = 1$ or $k = n$ and many partitions for moderate k, inducing a strong bias in favour of $k = 1$ or $k = n$!
- Solution: define a covariate dependent random partition model that both removes undesirable partitions and retains certain properties of the random partition model induced by the DP.

Restricted DP: Construction

More specifically, we keep unchanged the probability law of the frequencies $(m_1, \ldots, m_n)$ corresponding to cluster sizes $(n_1, \ldots, n_k)$, where m_j is the number of $n_1, \ldots, n_k$ that are equal to j.
For the DP, $p(m_1, \ldots, m_n)$ is given by Ewens sampling formula.

By preserving the law of $(m_1, \ldots, m_n)$, we also preserve the probability law of the number of clusters k as for the DP.

Our proposed restricted covariate-dependent probability measure on the random partition is defined by

$$p^*(\rho_n|x) = \frac{\alpha^k}{\alpha^{[n]}} \frac{n!}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * I_{s_{\pi_x(1)} \leq \ldots \leq s_{\pi_x(n)}}$$

It satisfies the order constraint and has the same marginal for $(m_1, \ldots, m_n)$ and for k, as those induced by the Dirichlet process.

Restricted DP: Random partition

- Prior for the random partition:

$$p(\rho_n|x) \propto \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j} * I_{s_{(1)} \leq \ldots \leq s_{(n)}}.$$

- Posterior of ρ_n:

$$p(\rho_n|y,x) \propto \frac{\alpha^k}{k!} \prod_{j=1}^{k} \frac{1}{n_j} p(y_i \in S_j \mid x_i \in S_j) * I_{s_{(1)} \leq \ldots \leq s_{(n)}}(\rho_n).$$

- Computations for $p(\rho_n \mid y, x)$ use reversible jump MCMC (Fuentes-Garcia et al., 2010).
- Note: smaller parameter space $\rightarrow$ **faster computations and better mixing**!

Restricted DP mixtures: Prediction

- Prediction of y_{n+1}:

$$\mathrm{E}[Y_{n+1} \mid x_{n+1}, y, x] = \tag{2}$$

$$\sum_{\rho_n \in \mathcal{C}_n} \left(\sum_{s_{n+1} \in \mathcal{C}(x_{n+1}, \rho_n)} \mathrm{E}[Y_{n+1} \mid x_{n+1}, y, x, \rho_{n+1}] \frac{p(\rho_{n+1} \mid x, x_{n+1}) p(y \mid x)}{p(\rho_n \mid x) p(y \mid x, x_{n+1})} \right) p(\rho_n \mid y, x),$$

where $\mathcal{C}_n$ is the set of partitions under the constraint and $\mathcal{C}(x_{n+1}, \rho_n)$ is the set of s_{n+1} such that ρ_{n+1} restricted to n observed subjects is ρ_n.

Restricted DP mixtures: Prediction

- Inner term of (2) may have three forms depending on x_{n+1} and ρ_n:
  1. if $x_{n+1} < x_{(1)}$ (similarly for $x_{n+1} > x_{(n)}$):

$$\frac{\alpha}{(k+1)c}\beta_0' \underline{x}_{n+1} + \frac{n_1}{(n_1+1)c}\widehat{\beta}_1' \underline{x}_{n+1}.$$

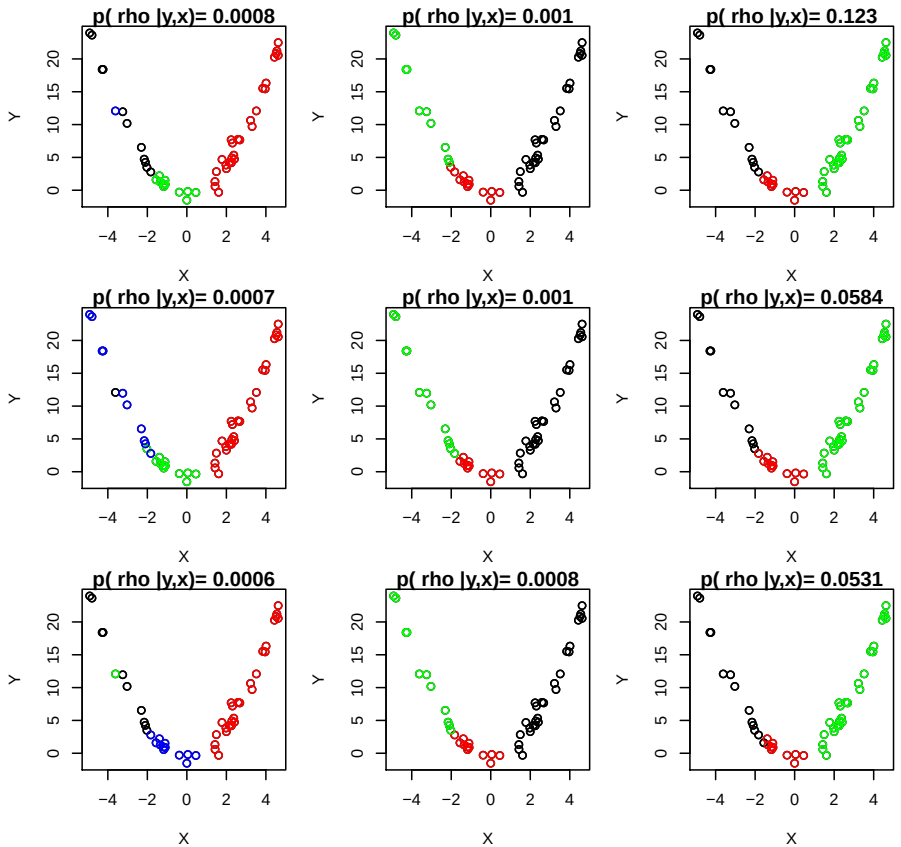
  2. if $x_{(i)} < x_{n+1} < x_{(i+1)}$ where $s_{(i)} = j$ and $s_{(i+1)} = j+1$:

$$\frac{\alpha}{(k+1)c}\beta_0' \underline{x}_{n+1} + \frac{n_j}{(n_j+1)c}\widehat{\beta}_j' \underline{x}_{n+1} + \frac{n_{j+1}}{(n_{j+1}+1)c}\widehat{\beta}_{j+1}' \underline{x}_{n+1}.$$

  3. if $x_{(i)} < x_{n+1} < x_{(i+1)}$ where $s_{(i)} = j$ and $s_{(i+1)} = j$:

$$\frac{n_j}{(n_j+1)c}\widehat{\beta}_j' \underline{x}_{n+1}.$$

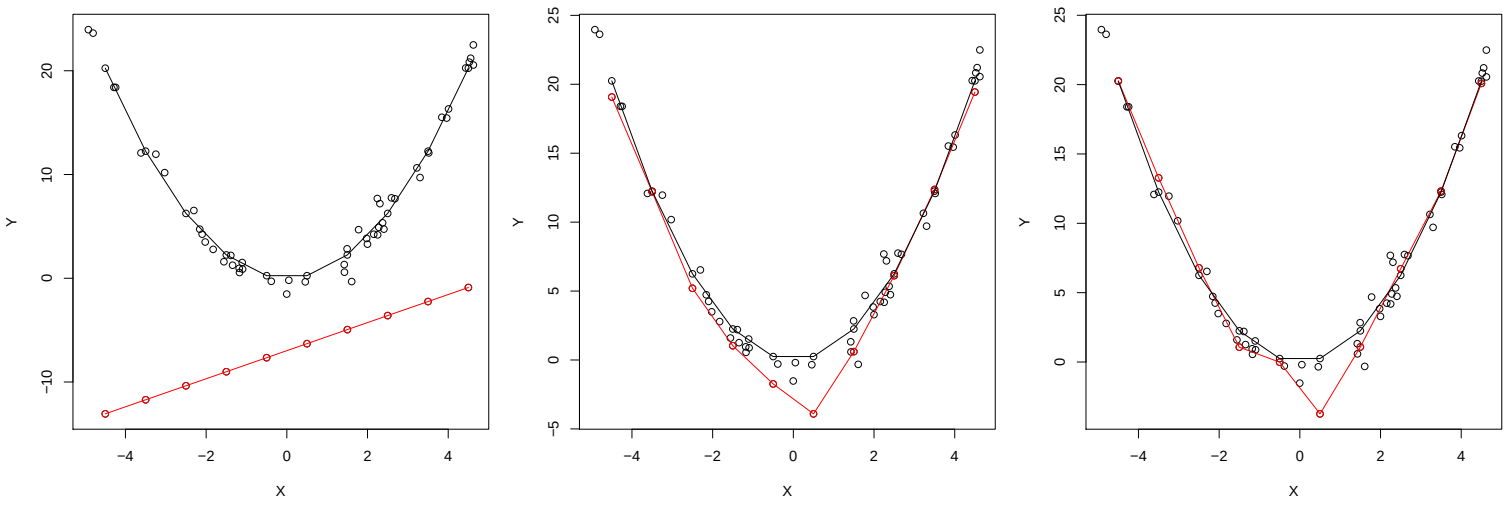
Simulated example: posterior on partition



(a) DPM (b) joint DPM (c) restricted DPM

Figure: Data are generated as $Y_i|x_i \overset{indep}{\sim} N(x_i^2, 1)$. Three partitions with the highest posterior probability. The very small values of the highest posterior probabilities show that the posterior for the DPM and jDPM is very spread out.

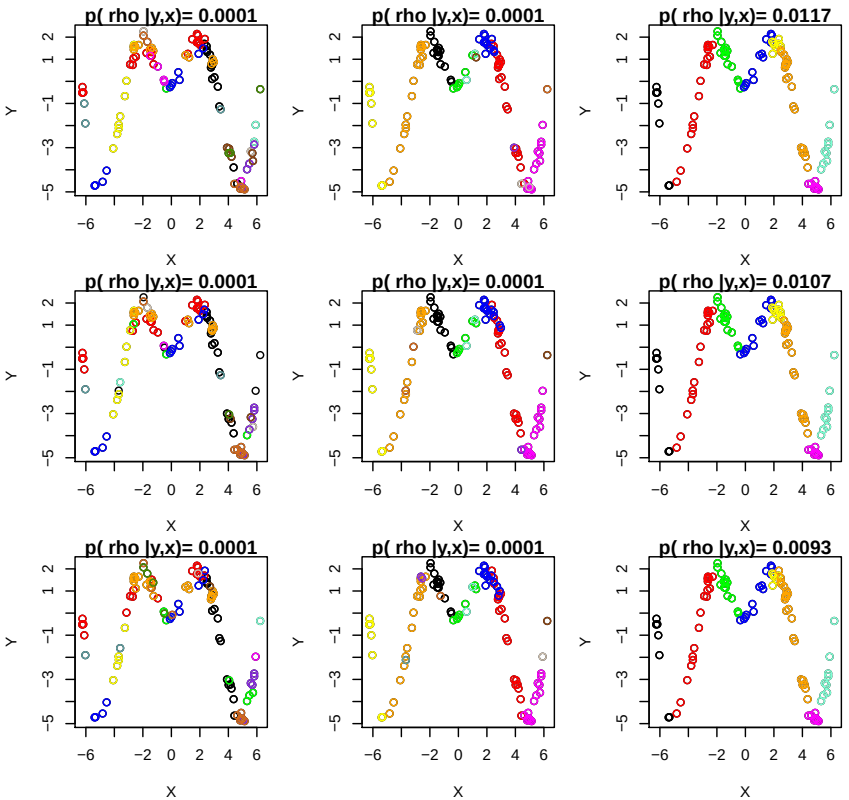
Simulated example: curve estimate (prediction)



(a) DPM (b) joint DPM (c) restricted DPM

Figure: Plot of the curve estimate in red at a grid of new x values, together with the true curve in black and observed data in black circles. The poor result for the DPM is due to the fact that the curve estimate is an average of the linear regressions from all clusters, independent of location of the new x value.

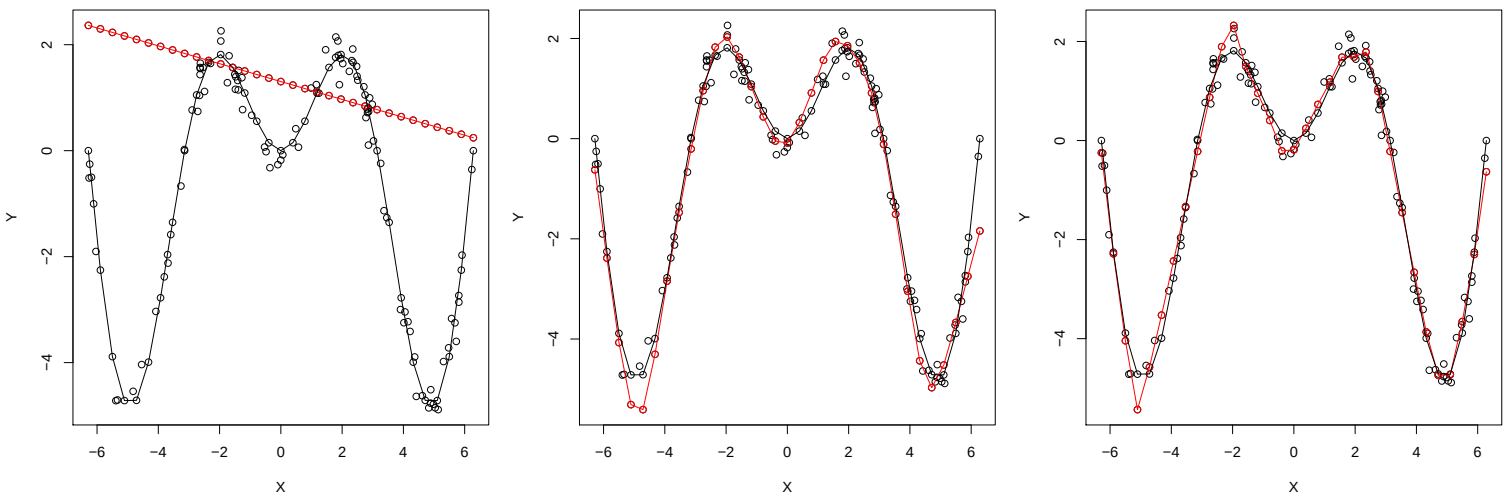
Simulated example 2: posterior on partition



(a) DPM (b) joint DPM (c) restricted DPM

Figure: Data are generated as $Y_i|x_i \overset{indep}{\sim} \mathrm{N}(x_i \sin x_i, 1/16)$. The posteriors on the partition for the DPM and jDPM are extremely spread out and the MCMC does not explore the partition space. Plots show three visited partitions. Column 3: the three partitions with the highest posterior prob. for the rDPM .

Simulated example 2: curve estimate



(a) DPM (b) joint DPM (c) restricted DPM

Figure: Curve estimate in red for a grid of new x values with the true curve in black and the observed data in black circles.

Restricted DP: Application

- Aim: Prediction of Alzheimer's Disease (AD) based on asymmetry of the hippocampus.
- Data: $n = 377$ of which 159 have been diagnosed with AD and 218 are cognitively normal, $y = 1$ indicate a healthy subject, and x represent the ratio of the volume of the left to right hippocampus.

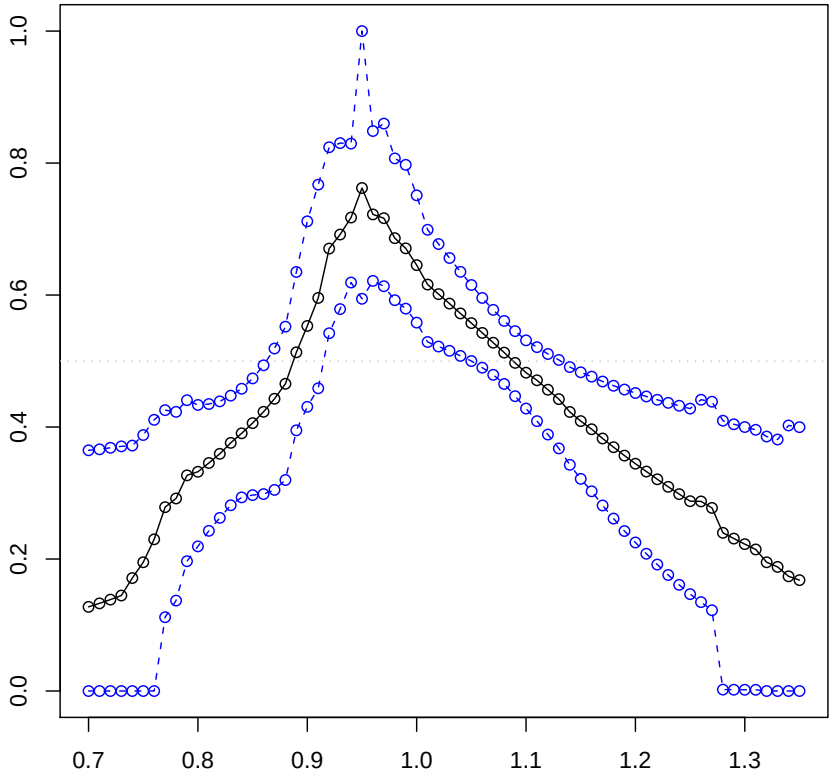


Figure: The estimated probability of being healthy for new subjects with left-to-right hippocampus ratios of 0.7 to 1.3 by 0.01 with 90% credible intervals.

summary

We underlined a problem with random partition in a simple regression mixture model, that needs to be addressed.

We discussed the need of setting at zero the prior on undesirable partitions, while preserving properties of the prior.

We obtain simpler computations, and better predictions.

For Bayesian statisticians, this is a contribution towards a more aware and thoughtful use of BNP methods.

For big-data people, hope this may give hints for appropriate restrictions in more general spaces..

summary

We underlined a problem with random partition in a simple regression mixture model, that needs to be addressed.

We discussed the need of setting at zero the prior on undesirable partitions, while preserving properties of the prior.

We obtain simpler computations, and better predictions.

For Bayesian statisticians, this is a contribution towards a more aware and thoughtful use of BNP methods.

For big-data people, hope this may give hints for appropriate restrictions in more general spaces..

thank you