On Bayesian nonparametric regression

Sonia Petrone

Bocconi University, Milano, Italy

with Sara Wade (University of Warwick, UK)
and Michele Peruzzi (Bocconi University)

Building Bridges, Oslo, 22-14 May 2017

motivation

1. Regression is possibly the most important problem in Statistics!

motivation

1. Regression is possibly the most important problem in Statistics!
2. bridging parametric & nonparametric

Bayesian nonparametrics somehow extends parametric constructions to processes..

motivation

1. Regression is possibly the most important problem in Statistics!

2. bridging parametric & nonparametric

Bayesian nonparametrics somehow extends parametric constructions to processes..

3. explosion of methods for **Bayesian nonparametric regression**.

However:

- still, less theory than for density estimation
- rich but fragmented literature
- little “ready-to-go” software

→ aim: a brief overview, having these issues in mind

motivation - 1. regression

Regression is possibly the most important problem in Statistics

Classical approaches now ‘compete’ with machine learning methods, deep learning, and more....

more bridges?
frequentist/Bayesian; $\cdots$ parametric/nonparametric, $\cdots$
statistics/machine learning? ... probabilistic/non-probabilistic?

Anyway, the basic issue of quantifying the uncertainty remains crucial. The output of regression is the basis for risk evaluation and decision. A poor quantification of uncertainty may be a disaster...

motivation - 1. regression

Regression is possibly the most important problem in Statistics

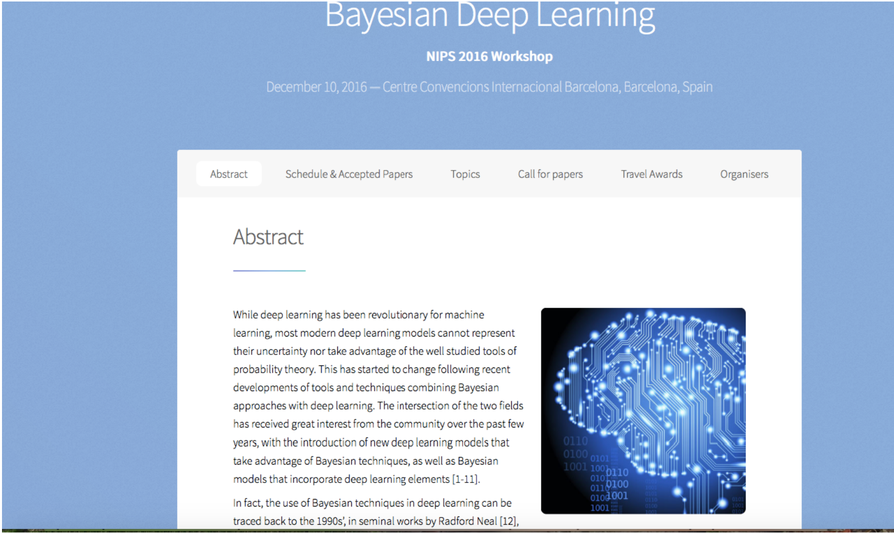
Classical approaches now ‘compete’ with machine learning methods, deep learning, and more....

more bridges?
frequentist/Bayesian; $\cdots$ parametric/nonparametric, $\cdots$
statistics/machine learning? ... probabilistic/non-probabilistic?

Anyway, the basic issue of quantifying the uncertainty remains crucial. The output of regression is the basis for risk evaluation and decision. A poor quantification of uncertainty may be a disaster...

..and the Bayesian approach is based on quantifying information and uncertainty!

deep learning, deep learning, deep learning....



motivation - 2. parametric & nonpar conjugate priors

Bayesian nonparametrics

$$X_i \mid F \overset{iid}{\sim} F, \quad F \sim \text{ prior distribution.}$$

The most popular *nonparametric* prior, the Dirichlet process, is the extension to a process of the Dirichlet conjugate prior for (*parametric*) multinomial sampling.

motivation - 2. parametric conjugate priors

$X \mid \xi \sim p(x \mid \xi)$ in the NEF,
$\xi \sim$ standard conjugate prior (Diaconis & Ylvisaker, 1979).

Bernoulli	Beta
Multinomial	Dirichlet
$\mathrm{N}(0, \Sigma)$	Inverse-Wishart
$\vdots$	$\vdots$

motivation - 2. parametric conjugate priors

$X \mid \xi \sim p(x \mid \xi)$ in the NEF,

$\xi \sim$ standard conjugate prior (Diaconis & Ylvisaker, 1979).

Bernoulli	Beta
Multinomial	Dirichlet
$\mathrm{N}(0, \Sigma)$	Inverse-Wishart
$\vdots$	$\vdots$

But, conjugate priors for a multivariate NEF are restrictive.

motivation - 2. parametric conjugate priors

$X \mid \xi \sim p(x \mid \xi)$ in the NEF,
$\xi \sim$ standard conjugate prior (Diaconis & Ylvisaker, 1979).

Bernoulli	Beta
Multinomial	Dirichlet
$\mathrm{N}(0, \Sigma)$	Inverse-Wishart
$\vdots$	$\vdots$

But, conjugate priors for a multivariate NEF are restrictive.

Generalized Dirichlet	Connor & Mosiman (1969)
Generalized Wishart	Brown, Le, Zidek (1994)
Enriched conjugate priors	Consonni and Veronese (2001)
$\vdots$	$\vdots$

If $f(x_1, \ldots, x_k; \xi)$ NEF, and can be decomposed in a product of NEF densities, each having its own parameters, the enriched conjugate prior on ξ is obtained by giving a standard conjugate prior for each of those densities.

examples

- $(X, Y) \mid \Sigma \sim \mathrm{N}_k(0, \Sigma)$.
  decompose $f(x, y; \Sigma) = f_x(x; \phi) f(y \mid x; \theta)$ (both Gaussian),
  and assign conjugate priors of ϕ and θ.
  $\rightarrow$ Generalized Wishart on Σ^{-1}.

- $(N_1, \ldots, N_k) \mid (p_1, \ldots, p_k) \sim \text{Multinomial}(N, p_1, \ldots, p_k)$.
  Re-write the multinomial as a product, by suitable reparametrization
  $\rightarrow$ Generalized Dirichlet distribution for $(p_1, \ldots, p_k)$.

motivation - 2. nonparametric conjugate priors

The DP extends to a process. Let $X_i \mid F \overset{iid}{\sim} F$.
A Dirichlet process prior for F, $F \sim DP(\alpha F_0)$, is such that, for any finite partition, the random vector of probabilities $(p_1, \ldots, p_k) = (F(A_1), \ldots, F(A_k)) \sim \mathrm{Dir}(\alpha F_0(A_1), \ldots, \alpha F_0(A_k))$ (Ferguson, 1973).

The DP inherits conjugacy from the conjugacy of the Dirichlet distribution for multinomial sampling; but also its lack of flexibility. This is the more true when F is on $\Re^k$.

...An enriched conjugate nonparametric prior?

Enriched conjugate prior distributions: non-parametric

Doksum's Neutral to the Right Process (Doksum, 1974) extends the enriched conjugate Dirichlet distribution to a process.

Enriched conjugate prior distributions: non-parametric

Doksum's Neutral to the Right Process (Doksum, 1974) extends the enriched conjugate Dirichlet distribution to a process.

However, NTR processes are limited to univariate random distributions.

The Dirichlet distribution implies that any permutation of $(p_1, \ldots, p_k)$ is completely neutral (that is, $p_1, p_2/(1-p_1), \ldots, p_k/(1-\sum_{j=1}^{k-1} p_j)$ are independent).
The Generalized Dirichlet only assumes that one ordered vector $(p_1, \ldots, p_k)$ is completely neutral. In $\Re$, there is a natural ordering. In more dimensions (e.g., contingency tables $p(x, y)$), there is no natural ordering.

Enriched Dirichlet Process

(Wade, Mongelluzzo, P., 2011).

Finite case: we define an Enriched Dirichlet distribution for $[p(x, y)]$ by choosing the ordering through $p(x, y) = p_{y|x}(y \mid x)p_x(x)$, and assuming that the vectors of the marginal and conditional probabilities have independent Dirichlet disributions.

Enriched Dirichlet Process

(Wade, Mongelluzzo, P., 2011).

Finite case: we define an Enriched Dirichlet distribution for $[p(x, y)]$ by choosing the ordering through $p(x, y) = p_{y|x}(y \mid x) p_x(x)$, and assuming that the vectors of the marginal and conditional probabilities have independent Dirichlet disributions.

Nonparametric case: $P(x, y)$ random distribution on $\Re^k$. Assume:

- $P_x \sim DP(\alpha_x P_{0,x})$
- for any x, $P_{y|x} \sim DP(\alpha_y(x)\, P_{0,y|x})$

all independent.

This well defines a probability law for the random $P(x, y)$, named Enriched Dirichlet Process (EDP)

motivation - 3. Bayesian nonparametric regression

Focus on mixture models for

1. conditional density estimation: estimate $f(y \mid x)$, and
2. density regression: how the density $f_x(y)$ of Y varies with x.

motivation - 3. Bayesian nonparametric regression

Focus on mixture models for
1. conditional density estimation: estimate $f(y \mid x)$, and
2. density regression: how the density $f_x(y)$ of Y varies with x.

1. Details of the 'nonparametric' prior matter – not only for asymptotics, but for finite sample prediction.
   We show the case of two different BNP priors, both consistent: but one of them is clearly *better* – better exploits information.
   Bayesian criteria to formalize such improvement?
2. Explosion of dependent Dirichlet processes (DDP) models.
   How can we compare? and offer a 'default choice' to practitioners, possibly in an **R**-package?

Outline

contents

Density regression

X predictor, p-dimensional; Y response.

$*$ Bayesian nonparametric (mean) regression:

$$: x \to m(x) = \mathrm{E}(Y \mid x)$$

flexible model/prior on $m(x)$ (basis expansions, Gaussian processes and splines (Wahba (1990), Denison, Holmes, Mallick, Smith (2002)), wavelets (Vidakovic, 2009), neural networks (Neal, 1996),..

$*$ Yet, the mean may be a too poor summary of the relationship between x and y.
$\Rightarrow$ median, quantile regression,.. density regression

$$: x \to f(y \mid x)$$

Limited literature on optimal estimators of $f(y \mid x)$. (Efromovich, Ann. Stat. 2007). How about Bayesian methods?

Random or fixed design

- regression: random design
  (X_i, Y_i), $i = 1, \ldots, n$ are a random sample from $f(x, y)$.

  Then, estimate the joint density $f(x, y)$ and from this the conditional $f(y \mid x) = \frac{f(x,y)}{f_x(x)}$.

- fixed design
  $x_1, \ldots, x_n$ is a deterministic sequence. If predictor is x, $Y \sim f(y \mid x)$. in fact, $f_x(y)$.

  - replicates of y values at a given x (x typically categorical or ordinal), (e.g. ANOVA)
  - no replicates (x typically continuous)
    still interest in $f_x(y)$: borrowing strength through smoothness conditions along x.

contents

Random design: DP mixtures for $f(x, y)$

$$(X_i, Y_i) \mid f \stackrel{iid}{\sim} f(x, y), \quad f \sim \text{ prior prob. law}$$

Random design: DP mixtures for $f(x, y)$

$$(X_i, Y_i) \mid f \overset{iid}{\sim} f(x, y), \quad f \sim \text{ prior prob. law}$$

$*$ Model f as a mixture of kernels:

$$\begin{aligned} (X_i, Y_i) \mid G &\overset{iid}{\sim} f_G(x, y) = \int K(x, y \mid \theta) dG(\theta) \\ G &\sim DP(\alpha G_0) \end{aligned}$$

Usually, $K(x, y \mid \theta) = N_{p+1}(x, y \mid \mu, \Sigma)$, with $\theta = (\mu, \Sigma)$, and $G_0(\theta)$ conjugate prior.

Random design: DP mixtures for $f(x, y)$

$$(X_i, Y_i) \mid f \overset{iid}{\sim} f(x, y), \quad f \sim \text{ prior prob. law}$$

$*$ Model f as a mixture of kernels:

$$\begin{aligned} (X_i, Y_i) \mid G &\overset{iid}{\sim} f_G(x, y) = \int K(x, y \mid \theta) dG(\theta) \\ G &\sim DP(\alpha G_0) \end{aligned}$$

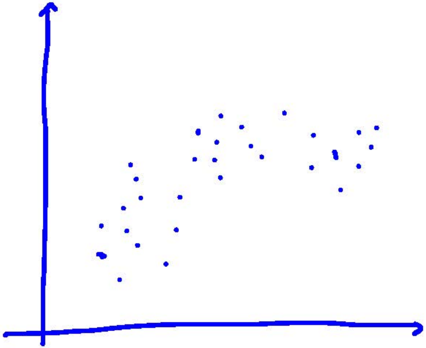
Usually, $K(x, y \mid \theta) = N_{p+1}(x, y \mid \mu, \Sigma)$, with $\theta = (\mu, \Sigma)$, and $G_0(\theta)$ conjugate prior.

Since, a.s., $G = \sum_{j=1}^{\infty} p_j \delta_{\theta_j^*}$, where $(p_j) \sim$ stick-breaking(α) independent of $\theta_j^* \overset{iid}{\sim} G_0$, the mixture above reduces to

$$f_G(x, y) = \sum_{j=1}^{\infty} w_j K(x, y \mid \theta_j^*).$$

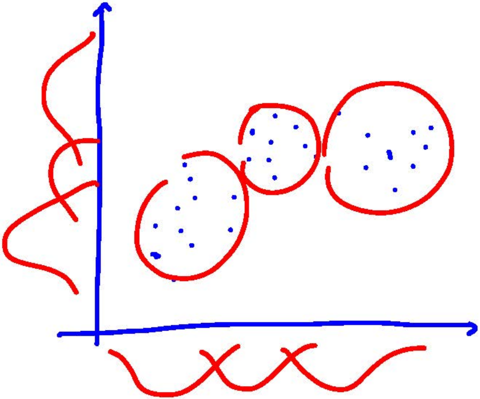
mixture of Gaussian kernels

in picture



mixture of Gaussian kernels

in picture



Latent variable representation

The DP-mixture model is equivalently expressed as

$$\begin{aligned} (X_i, Y_i) \mid \theta_i &\overset{ind}{\sim} K(x, y \mid \theta_i) \\ \theta_i \mid G &\overset{iid}{\sim} G \\ G &\sim DP(\alpha G_0) \end{aligned}$$

Integrating the θ_i out, one has back the countable mixture model

$$(X_i, Y_i) \mid G \overset{iid}{\sim} f_G(x, y) = \sum_{j=1}^{\infty} p_j \, K(x, y \mid \theta_j^*).$$

Latent variable representation

The DP-mixture model is equivalently expressed as

$$\begin{aligned}(X_i, Y_i) \mid \theta_i &\stackrel{ind}{\sim} K(x, y \mid \theta_i) \\ \theta_i \mid G &\stackrel{iid}{\sim} G \\ G &\sim DP(\alpha G_0)\end{aligned}$$

Integrating the θ_i out, one has back the countable mixture model

$$(X_i, Y_i) \mid G \stackrel{iid}{\sim} f_G(x, y) = \sum_{j=1}^{\infty} p_j \, K(x, y \mid \theta_j^*).$$

Then the conditional density $f(y \mid x)$ is obtained as

$$f_G(y \mid x) = \frac{\sum_j p_j \; K(x, y \mid \theta_j^*)}{\sum_j p_i K(x \mid \theta_j^*)} = \sum_j p_j(x) K(y \mid x, \theta_j^*)$$

where $p_j(x) = p_j \; K(x \mid \theta_j^*)/(\sum_{j'} p_{j'} K(x \mid \theta_{j'}^*)$.

Random partition

Since G is a.s. discrete, ties in a sample $(\theta_1, \ldots, \theta_n)$ from G have positive probability, so that

$$(\theta_1, \ldots, \theta_n) \text{ described by } (\rho_n; \theta_1^*, \ldots, \theta_{k(\rho_n)}^*)$$

- a random partition $\rho_n = (s_1, \ldots, s_n)$
- the cluster-specific parameters θ_j^*

Ex: for $n = 5$, $\rho_n = (1, 1, 2, 2, 1)$ gives $(\theta_1, \ldots, \theta_n) = (\theta_1^*, \theta_1^*, \theta_2^*, \theta_2^*, \theta_1^*)$, $k_n = 2$ two clusters of size $n_1 = 3, n_2 = 2$ resp., with cluster-specific parameters θ_1^*, θ_2^*.

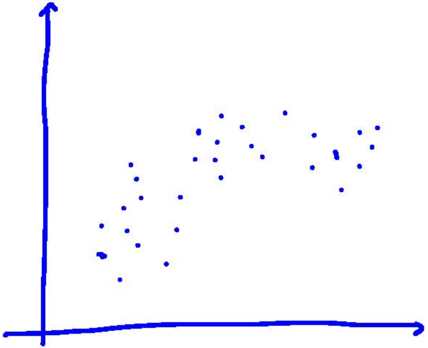
- The DP induces a probability law of the random partition

$$p(\rho_n) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^{k_n} \prod_{j=1}^{k_n} \Gamma(n_j)$$

- Given the partition ρ_n, the cluster specific parameters θ_j^* are i.i.d. $\sim G_0$.

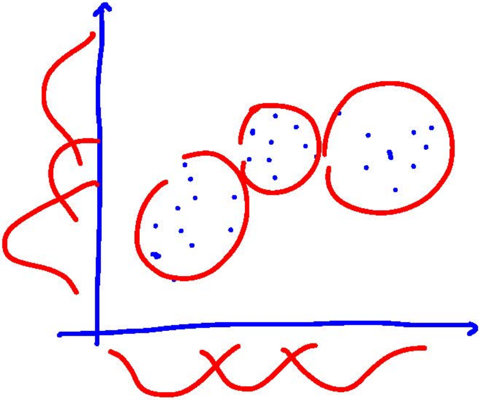
mixture of Gaussian kernels

in picture

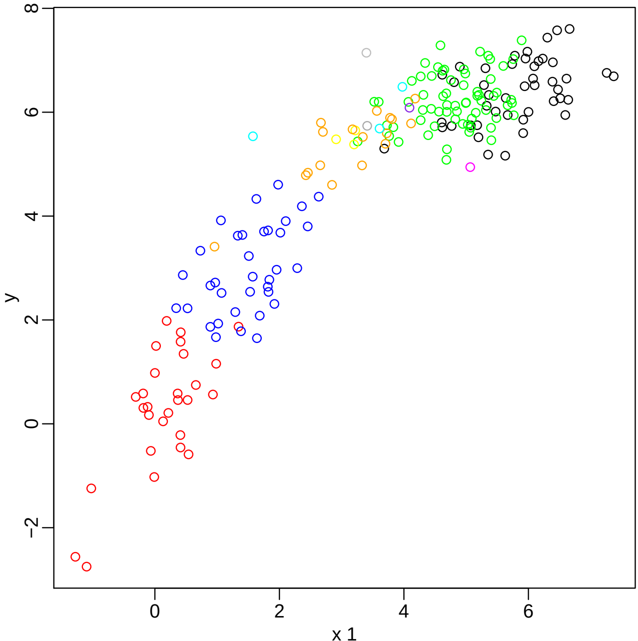


mixture of Gaussian kernels

in picture



mixture of Gaussian kernels



Inference

$*$ **posterior on ρ_n**

$$
\begin{aligned}
p(\rho_n \mid x_{1:n}, y_{1:n}) &\propto p(\rho_n) \prod_{j=1}^{k_n} \int \prod_{i:(x_i,y_i)\in S_j} K(x_i, y_i \mid \theta_j^*) dP_0(\theta_j^*) \\
&\propto p(\rho_n) \prod_{j=1}^{k_n} m(\{(x_i, y_i) \in C_j\} \mid \rho_n)
\end{aligned}
$$

prior $\times$ independent marginal likelihoods in each clusters.

Inference

$*$ **posterior on ρ_n**

$$\begin{aligned} p(\rho_n \mid x_{1:n}, y_{1:n}) &\propto p(\rho_n) \prod_{j=1}^{k_n} \int \prod_{i:(x_i,y_i)\in S_j} K(x_i, y_i \mid \theta_j^*) dP_0(\theta_j^*) \\ &\propto p(\rho_n) \prod_{j=1}^{k_n} m(\{(x_i, y_i) \in C_j\} \mid \rho_n) \end{aligned}$$

prior $\times$ independent marginal likelihoods in each clusters.

$*$ **posterior on (θ_j^*)** *Given the partition*, clusters are independent, and inference on θ_j^* is based only on obs. in group C_j

$$p(\theta_1^*, \dots, \theta_{k_n}^* \mid x_{1:n}, y_{1:n}, \rho_n) = \prod_{j=1}^{k_n} p(\theta_j^* \mid \rho_n, \{(x_i, y_i) \in C_j\})$$

Predictive distribution (density estimate)

With quadratic loss, the Bayesian density estimate is the predictive density

$$(X_{n+1}, Y_{n+1}) \mid x_{1:n}, y_{1:n} \sim \hat{f}(x, y) = \mathrm{E}(f(x, y) \mid x_{1:n}, y_{1:n}).$$

$*$ Recalling that $G \mid \theta_{1:n}, x_{1:n}, y_{1:n} \sim DP(\alpha G_0 + \sum_{i=1}^{n} \delta_{\theta_i})$,

$$\begin{aligned}
\hat{f}(x, y) &= \mathrm{E}(\mathrm{E}(f_G(x, y) \mid \theta_{1:n}) \mid x_{1:n}, y_{1:n}) \\
&= \frac{\alpha}{\alpha + n} f_{G_0}(x, y) + \frac{n}{\alpha + n} \mathrm{E}\left(\sum_{i=1}^{n} \frac{K(x, y \mid \theta_i)}{n} \mid x_{1:n}, y_{1:n} \right)
\end{aligned}$$

average of prior guess f_{G_0} and expectation of a kernel estimate with kernels centered at the θ_i.

conditional density estimate

$*$ Since $(\theta_1, \ldots, \theta_n) \leftrightarrow (\rho_n, \theta_1^*, \ldots, \theta_k^*)$, the joint density estimate is

$$\hat{f}(x,y) = \frac{\alpha}{\alpha+n} f_{G_0}(x,y)$$

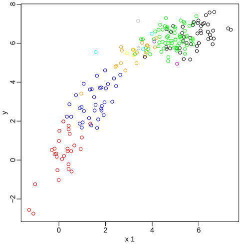
$$+ \sum_{\rho_n} \left(\sum_{j=1}^{k(\rho_n)} \frac{n_j(\rho_n)}{\alpha+n} \, f(x, y \mid (x_i, y_i) \in C_j(\rho_n)) \right) \, p(\rho_n \mid x_{1:n}, y_{1:n})$$

average of the prior guess f_{G_0}, and given the partition ρ_n, of the predictive densities in clusters $C_j(\rho_n)$.

From $\hat{f}(x,y)$, one can find an estimate of $f(y \mid x)$.

'clustering' and density estimate

The partition is not of main interest (mixture components just play the role of kernels), but $p(\rho_n \mid x_{1:n}, y_{1:n})$ plays a crucial role.



Such role of the prior and posterior distribution of the random partition is often overlooked.

How comparing nonparametric priors?

Frequentist properties.
Results on frequentist asymptotic properties for multivariate density estimation, and some results for regression and conditional density estimation (Wu & Ghosal (2008; 2010); Tokdar (2011), Norets & Pelenis (2012), Shen, Tokdar & Ghosal (2013), Canale & Dunson (2015), Bhattacharya, Pati, Dunson (2014)– anisotropic; Norets & Pati (2016+), ...

But, how about (Bayesian) finite sample properties and predictive performance?

How comparing nonparametric priors?

Frequentist properties.
Results on frequentist asymptotic properties for multivariate density estimation, and some results for regression and conditional density estimation (Wu & Ghosal (2008; 2010); Tokdar (2011), Norets & Pelenis (2012), Shen, Tokdar & Ghosal (2013), Canale & Dunson (2015), Bhattacharya, Pati, Dunson (2014)– anisotropic; Norets & Pati (2016+), ...

But, how about (Bayesian) finite sample properties and predictive performance?

Different nonparametric priors may be consistent, but have quite diverse predictive performance!
The implied distribution on the random partition plays a crucial role.

How comparing nonparametric priors?

Frequentist properties.
Results on frequentist asymptotic properties for multivariate density estimation, and some results for regression and conditional density estimation (Wu & Ghosal (2008; 2010); Tokdar (2011), Norets & Pelenis (2012), Shen, Tokdar & Ghosal (2013), Canale & Dunson (2015), Bhattacharya, Pati, Dunson (2014)– anisotropic; Norets & Pati (2016+), ...

But, how about (Bayesian) finite sample properties and predictive performance?

Different nonparametric priors may be consistent, but have quite diverse predictive performance!
The implied distribution on the random partition plays a crucial role.

Example 1. Wade, Walker & Petrone (*Scand. J. Statist.*, 2014) introduce a restricted partition model to overcome difficulties of DP mixtures in curve fitting. Their proposal clearly leads to improved prediction.

Problem: anisotropic case

X continuos predictors; Gaussian kernels.
The DP mixture of multivariate Gaussian distributions uses joint clusters to fit the density $f(x, y)$.

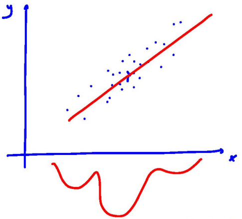
Problem: anisotropic case

X continuos predictors; Gaussian kernels.
The DP mixture of multivariate Gaussian distributions uses joint clusters to fit the density $f(x,y)$.
BUT, the conditional density $f_{y|x}$ and the marginal density f_x might have different smoothness; in regression, typically $f_{y|x}$ is smoother than f_x.

Problem: anisotropic case

X continuos predictors; Gaussian kernels.
The DP mixture of multivariate Gaussian distributions uses joint clusters to fit the density $f(x, y)$.
BUT, the conditional density $f_{y|x}$ and the marginal density f_x might have different smoothness; in regression, typically $f_{y|x}$ is smoother than f_x.

in picture



here, many small clusters (kernels) are needed to fit the f_x density, while much fewer kernels would suffice for $f_{y|x}$.
If the dimension of x is large, the likelihood is dominated by the x component and many small clusters are suggested by the posterior on ρ_n.
This impoverishes the performance of the model.

This undesirable behavior does not seem to vanish with increasing sample size.

If $f_x(x)$ requires many clusters, the unappealing behaviour of the random partition could be reflected in worse convergence rates. Efromovich [2007] shows that if the conditional density is smoother than the joint, it can be estimated at a faster rate.

Thus, improving inference on the random partition to take into account the different degree of smoothness of f_x and $f_{y|x}$ is crucial.

Improving prediction

Consider the Dirichlet mixture of Gaussian kernels

$$(X_i, Y_i) \mid G \sim \int \mathrm{N}_{p+1}(\mu, \Sigma) dG(\mu, \Sigma), \quad G \sim DP(\alpha G_0).$$

The base measure of the DP, $G_0(\mu, \Sigma)$, is usually Normal-Inv Wishart. BUT, this conjugate prior is restrictive if p is large

Improving prediction

- Write the kernels as

$$N_{p+1}(x, y \mid \mu, \Sigma) = N_p(x \mid \mu_x, \Sigma_x)\ N(y \mid x'\beta, \sigma^2_{y|x})$$

and use simple spherical x-kernels (Shahbaba and Neal, 2009, Hannah et al., 2011). Thus,

$$f_x(x \mid P) = \sum_{j=1}^{\infty} w_j N_p(x \mid \mu^*_{xj}, \begin{pmatrix} \sigma^{2*}_{x1,j} & 0 & \cdots & 0 \\ 0 & \sigma^{2*}_{x2,j} & 0 & 0 \\ 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \sigma^{2*}_{xp,j} \end{pmatrix})$$

$$f(y \mid x, P) = \sum_{j=1}^{\infty} w_j(x) N(y \mid x'\beta^*_j, \sigma^{2\,*}_{y|x,j}).$$

Improving prediction

- Write the kernels as

$$N_{p+1}(x, y \mid \mu, \Sigma) = N_p(x \mid \mu_x, \Sigma_x)\; N(y \mid x'\beta, \sigma^2_{y|x})$$

and use simple spherical x-kernels (Shahbaba and Neal, 2009, Hannah et al., 2011). Thus,

$$f_x(x \mid P) = \sum_{j=1}^{\infty} w_j N_p(x \mid \mu^*_{xj}, \begin{pmatrix} \sigma^{2*}_{x1,j} & 0 & \cdots & 0 \\ 0 & \sigma^{2*}_{x2,j} & 0 & 0 \\ 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \sigma^{2*}_{xp,j} \end{pmatrix})$$

$$f(y \mid x, P) = \sum_{j=1}^{\infty} w_j(x) N(y \mid x'\beta^*_j, \sigma^{2\,*}_{y|x,j}).$$

- Denote the x-parameters $\phi = ((\mu_{x1}, \ldots, \mu_{xp}), (\sigma^2_{x1}, \ldots, \sigma^2_{xp}))$ and the $y|x$-parameters $\theta = (\beta, \sigma^2_{y|x})$. Assign independent conjugate priors $G_{0,\phi}(\phi)$ and $G_{0,\theta}(\theta)$ (that leads to an enriched conjugate prior for (μ, Σ)

Improving prediction

- Write the kernels as

$$N_{p+1}(x, y \mid \mu, \Sigma) = N_p(x \mid \mu_x, \Sigma_x)\; N(y \mid x'\beta, \sigma^2_{y|x})$$

  and use simple spherical x-kernels (Shahbaba and Neal, 2009, Hannah et al., 2011). Thus,

$$f_x(x \mid P) = \sum_{j=1}^{\infty} w_j N_p(x \mid \mu^*_{xj}, \begin{pmatrix} \sigma^{2*}_{x1,j} & 0 & \cdots & 0 \\ 0 & \sigma^{2*}_{x2,j} & 0 & 0 \\ 0 & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \sigma^{2*}_{xp,j} \end{pmatrix})$$

$$f(y \mid x, P) = \sum_{j=1}^{\infty} w_j(x) N(y \mid x'\beta^*_j, \sigma^{2\,*}_{y|x,j}).$$

- Denote the x-parameters $\phi = ((\mu_{x1}, \ldots, \mu_{xp}), (\sigma^2_{x1}, \ldots, \sigma^2_{xp}))$ and the $y|x$-parameters $\theta = (\beta, \sigma^2_{y|x})$. Assign independent conjugate priors $G_{0,\phi}(\phi)$ and $G_{0,\theta}(\theta)$ (that leads to an enriched conjugate prior for (μ, Σ))
- In the DP mixture model, assume individual-specific (ϕ_i, θ_i) as a random sample from $G \sim DP(\alpha G_{0,\phi}\, G_{0,\theta})$

Joint DP mixture model

$$
\begin{aligned}
Y_i | x_i, \beta_i, \sigma^2_{y,i} &\overset{ind}{\sim} N(\beta_i' \underline{x}_i, \sigma^2_{y|x,i}), \quad \theta_i = (\beta_i, \sigma^2_{y|x,i}), \\
X_i | \mu_i, \sigma^2_{x,i} &\overset{ind}{\sim} \prod_{h=1}^{p} N(\mu_{x,h,i}, \sigma^2_{x,h,i}), \quad \phi_i = (\mu_{x,i}, \sigma^2_{x,i}) \\
(\theta_i, \phi_i) \mid G &\overset{i.i.d.}{\sim} G, \\
\mathbf{G} &\sim DP(\alpha G_{0\theta} \times G_{0\phi}).
\end{aligned}
$$

with $G_{0\theta}$ and $G_{0\phi}$ independent conjugate Normal-Inverse Gamma priors.

behavior for large p

The model is flexible, and MCMC computations are standard.

BUT, if p is large, many (independent) kernels will be typically needed to describe the (dependent) marginal f_x, while the relationship $Y \mid x$ can be smoother.

However, the DP only allows joint clusters of (ϕ_i, θ_i), $i = 1, \ldots, n$.

Given its crucial role, difficulties in the random partition have relevant consequences on prediction
We would like to use a prior on P that allows many ϕ_i clusters, to fit f_x, but fewer θ_j clusters, and it is still conjugate, so that computations remain simple.

behavior for large p

The model is flexible, and MCMC computations are standard.

BUT, if p is large, many (independent) kernels will be typically needed to describe the (dependent) marginal f_x, while the relationship $Y \mid x$ can be smoother.

However, the DP only allows joint clusters of (ϕ_i, θ_i), $i = 1, \ldots, n$.

Given its crucial role, difficulties in the random partition have relevant consequences on prediction
We would like to use a prior on P that allows many ϕ_i clusters, to fit f_x, but fewer θ_j clusters, and it is still conjugate, so that computations remain simple.

$\Rightarrow$ Enriched Dirichlet Process (Wade, Mongelluzzo, P., 2011)

Enriched Dirichlet Process

Extends the idea that leads from the Dirichlet distribution (conjugate to the multinomial) to the enriched (or generalized) Dirichlet dist. (Connor & Mosiman (1969);

and from the DP to Doksum (1974) neutral-to-the-right priors for random probability measures **on the real line**.

EDP: Assign a prior for the random prob. measure $G(\theta, \phi)$ by assuming:

- $P_\theta \sim DP(\alpha_\theta P_{0,\theta})$
- for any θ, $P_{\phi|\theta} \sim DP(\alpha_\phi(\theta)\, P_{0,\phi|\theta})$

all independent.

The EDP gives a nested random partition: $\rho_n = (\rho_{n,\theta}, \rho_{n,\phi})$, that allows many ϕ-clusters inside each θ-cluster.

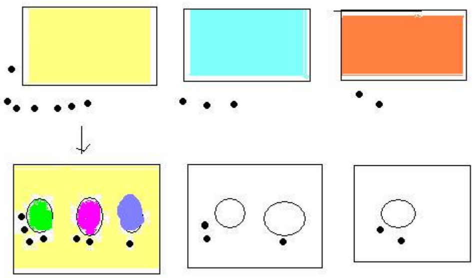
EDP: nested random partition

$\rho_n = (\rho_{n,\theta}, \rho_{n,\phi})$: many ϕ-clusters inside each θ-cluster.

EDP: nested random partition

$\rho_n = (\rho_{n,\theta}, \rho_{n,\phi})$: many ϕ-clusters inside each θ-cluster.

- $P_\theta \sim DP(\alpha P_{0\theta})$ gives a Chinese restaurant process: customers choose *restaurants*, and restaurants are colored with colors $\theta^*_h \overset{iid}{\sim} P_{0\theta}$ (nonatomic):
- $P_{\phi|\theta} \sim DP(\alpha_\phi(\theta)\, P_{0\phi|\theta})$ gives a *nested* CRP: within each restaurant, customers sits at tables as in the CRP. Tables in restaurant θ^*_h are colored with colors $\phi^*_{j|h} \overset{iid}{\sim} P_{0,\phi|\theta}(\phi \mid \theta)$.



nested Chinese restaurant

Joint EDP mixture model

- Model (replace DP with EDP):

$$
\begin{aligned}
Y_i | x_i, \beta_i, \sigma^2_{y,i} &\overset{ind}{\sim} N(\beta_i' \underline{x}_i, \sigma^2_{y|x,i}), \quad \theta_i = (\beta_i, \sigma^2_{y|x,i}), \\
X_i | \mu_i, \sigma^2_{x,i} &\overset{ind}{\sim} \prod_{h=1}^{p} N(\mu_{x,h,i}, \sigma^2_{x,i,h}), \quad \phi_i = (\mu_{x,i}, \sigma^2_{x,i}) \\
(\theta_i, \phi_i) \mid G &\overset{i.i.d.}{\sim} G, \\
\mathbf{G} &\sim EDP(\alpha_\theta, \alpha_\phi(\cdot), G_{0,\theta} \times G_{0,\phi|\theta}).
\end{aligned}
$$

- Computations remain simple, as the EDP is a conjugate prior;
- Inference on a cluster-specific $\theta^*_j = (\beta^*_j \sigma_{y|x^*})$, (thus, ultimately, on the conditional density $f(y \mid x, \theta)$), exploits the information from the observations in all the ϕ^*_h-clusters that share the same θ^*_j $\Rightarrow$ much improved inference and prediction

(Wade, Dunson, Petrone, Trippa, JMLR (2014))

Simulation study

Toy example to demonstrate two key advantages of the EDP model

- it can recover the true coarser θ-partition
- improved prediction and smaller credible intervals result.

Data: 200 obs (x_i, y_i) where
The covariates X_i are sampled from a p-variate normal,

$$X_i = (X_{i,1}, \ldots, X_{i,p})' \overset{iid}{\sim} \mathrm{N}(\mu, \Sigma),$$

centered at $\mu = (4, \ldots, 4)'$ with $\Sigma_{h,h} = 4$ for $h = 1, \ldots, p$, and covariances that model two groups of covariates with different correlation structure.

Simulation study

Toy example to demonstrate two key advantages of the EDP model

- it can recover the true coarser θ-partition
- improved prediction and smaller credible intervals result.

Data: 200 obs (x_i, y_i) where
The covariates X_i are sampled from a p-variate normal,

$$X_i = (X_{i,1}, \ldots, X_{i,p})' \stackrel{iid}{\sim} \mathrm{N}(\mu, \Sigma),$$

centered at $\mu = (4, \ldots, 4)'$ with $\Sigma_{h,h} = 4$ for $h = 1, \ldots, p$, and covariances that model two groups of covariates with different correlation structure.

The true regression only depends on the first covariate; it is a nonlinear regression obtained as a mixture

$$Y_i | x_i \stackrel{ind}{\sim} p(x_{i,1})\mathrm{N}(y_i|\beta_{1,0}+\beta_{1,1}x_{i,1}, \sigma_1^2)+(1-p(x_{i,1}))\mathrm{N}(y_i|\beta_{2,0}+\beta_{2,1}x_{i,1}, \sigma_2^2)$$

prediction error

Table: Prediction error for both models as p increases.

	$p=1$		$p=5$		$p=10$		$p=15$	
	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$	$\hat{l}_1$	$\hat{l}_2$
DP	**0.03**	**0.05**	0.16	0.2	0.25	0.34	0.26	0.34
EDP	0.04	**0.05**	**0.06**	**0.1**	**0.09**	**0.16**	**0.12**	**0.21**

Real data: Alzheimer disease diagnosis

Data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which is publicly accessible at UCLA's Laboratory of Neuroimaging.

covariates: summaries of $p = 15$ brain structures computed from structural MRI obtained at the first visit for 377 patients, of which 159 have been diagnosed with AD and 218 are cognitively normal (CN).

response: $Y_i = 1$ (cognitively normal subject); or $= 0$ (diagnosed with AD).

Aim: Prediction of AD status

Extension of the model to binary response

The model is extended to a local probit model:

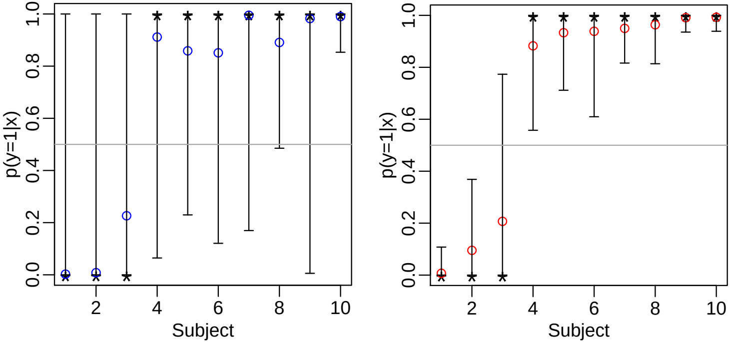
$$Y_i|x_i, \beta_i \overset{ind}{\sim} \mathrm{Bern}(\Phi(\underline{x}_i\beta_i)), \quad X_i|\mu_i, \sigma_i^2 \overset{ind}{\sim} \prod_{h=1}^{p} \mathrm{N}(\mu_{i,h}, \sigma_{i,h}^2),$$

$$(\beta_i, \mu_i, \sigma_i^2)|G \overset{iid}{\sim} G, \quad \mathbf{G} \sim Q.$$

First, DP prior for G: $G \sim DP(\alpha, G_{0\beta} \times G_{0\psi})$, with $G_{0\beta} = \mathrm{N}(0_p, C^{-1})$ and $G_{0\psi}$ product of p normal-inverse gamma. We let $\alpha \sim \mathrm{Gamma}(1,1)$.

EDP prior for G: correlation between the measurements of the brain structures and non-normal univariate histograms of the covariates suggest that many Gaussian kernels with local independence will be needed to approximate the density of X. The conditional density of the response, on the other hand, may not be so complicated. This motivates the choice of an EDP prior.

Prediction for 10 new subjects



(a) DP (b) EDP

Figure: Predicted probability of being healthy against subject index for 10 new subjects and represented with circles (DP in blue and EDP in red) with the true outcome as black stars. The bars about the prediction depict the 95% credible intervals.

contents

Fixed design: Conditional models

Sample $(x_i, Y_{i,\nu}), \nu = 1, \ldots, n_i)$, x fixed input.
Interest in $f_x(y)$ (no longer a conditional $f(y \mid x)$).

Fixed design: Conditional models

Sample $(x_i, Y_{i,\nu}), \nu = 1, \ldots, n_i)$, x fixed input.
Interest in $f_x(y)$ (no longer a conditional $f(y \mid x)$).
Joint DP mixture models are used in this context, too. Yet, they unnecessarily require to model the marginal density $f_x(x)$.

Fixed design: Conditional models

Sample $(x_i, Y_{i,\nu}), \nu = 1, \ldots, n_i)$, x fixed input.
Interest in $f_x(y)$ (no longer a conditional $f(y \mid x)$).
Joint DP mixture models are used in this context, too. Yet, they unnecessarily require to model the marginal density $f_x(x)$.

In a Bayesian approach, one wants to assign a prior on the set of random prob. measures $\{F_x(\cdot), x \in \mathcal{X}\}$.
The random measures F_x must be dependent, as we want some smoothness along x, for borrowing strength; and have a given marginal distribution, e.g. $F_x \sim DP$.

Early proposals: Cifarelli & Regazzini (1978), extending mixtures of DP (Antoniak, 1974).

Influential idea: dependent Dirichlet processes (DDP), McEachern (1999; 2000), based on dependent stick-breaking constructions.

DDP mixture models

Use a mixture model for $f_x(y)$

$$Y_{i,\nu} \mid x, G_x \overset{ind}{\sim} f_{G_x}(y) = \int K(y \mid x, \theta) dG_x(\theta)$$

where the mixing distribution G_x is indexed by x.

DDP mixture models

Use a mixture model for $f_x(y)$

$$Y_{i,\nu} \mid x, G_x \overset{ind}{\sim} f_{G_x}(y) = \int K(y \mid x, \theta) dG_x(\theta)$$

where the mixing distribution G_x is indexed by x.

A DDP prior on $\{G_x, x \in \mathcal{X}\}$ assumes that $G_x \sim DP(\alpha(x) G_{0,x})$, and dependence is introduced through the dependent stick-breaking constructions:

$$G_x(\cdot) = \sum_{j=1}^{\infty} p_j(x) \delta_{\theta_j^*(x)}(\cdot)$$

where, for each j:

- $(w_j(x), x \in \mathcal{X})$ is a stochastic process, with stick-breaking construction $w_1(x) = v_1(x); \quad w_2(x) = v_2(x)(1 - v_1(x)), \ldots$ where $(v_j(x), x \in \mathcal{X})$ is a stochastic process with marginals $v_j(x) \sim Beta(1, \alpha(x))$, and the $v_j(\cdot)$ are independent across x;
- $(\theta_j^*(x), x \in \mathcal{X})$ is a stochastic process with marginals $G_{0,x}$. The processes $\theta_j(\cdot)$ are independent across j, and indep. of the $v_j(\cdot)$.

Conditional approach: Single weights

The DDP allows both the mixing weights and the atoms to depend on x. But this is redundant, and one either considers 1) models with **single weights** and 2) models with **covariate dependent weights**.

Single weights: assume $w_j(x) = w_j$ with flexible $\theta_j(x)$:

$$f_{G_x}(y|x) = \sum_{j=1}^{\infty} w_j K(y|\theta_j(x)).$$

- Ex. $K(y|\theta_j(x)) = \mathrm{N}(y|\mu_j(x), \sigma_j^2)$ with $\mu_j \overset{iid}{\sim}$ GP.
- Popular because inference relies on established algorithms for BNP mixtures.

Conditional approach: Covariate dependent weights

Covariate dependent weights: flexible $w_j(x)$ with $\theta_j(x) = \theta_j$:

$$f_{P_x}(y|x) = \sum_{j=1}^{\infty} w_j(x) K(y|\theta_j, x),$$

for ex. $K(y|\theta_j(x)) = \mathrm{N}(y|\beta_j' x, \sigma_j^2)$.

- Most techniques to define $w_j(x)$ s.t $\sum_j w_j(x) = 1$ use a stick-breaking approach:

$$w_1(x) = v_1(x) \text{ and for } j > 1 \quad w_j(x) = v_j(x) \prod_{j'<j} (1 - v_{j'}(x)),$$

- Proposals for $v_j(x)$ include Griffin and Steel (2006), Dunson and Park (2008), normalized weights model (Antoniano, Wade, Walker; 2014); probit stick-breaking (Rodriguez and Dunson, 2011); logit stick-breaking (Ren, Du, Carin & Dunson (2011); Rigon & Durante, 2017+), ...

How can we compare?

Model comparison

The EDP example shows clear improvement in the predictive performance. We would like to formally express the gain of information that a model/prior can provide.

Model comparison

The EDP example shows clear improvement in the predictive performance. We would like to formally express the gain of information that a model/prior can provide.
So far

- Careful and detailed study of the information and assumption introduced through the prior/model; and the analytic implications on the predictive distribution;
- compare on simulated and real data.

Some findings (Peruzzi, Petrone & Wade, 2016+)

- The 'single weights' mixture model needs flexible atoms, for giving reasonable prediction. But this has drawbacks, including identifiability.
- Covariate-dependent weights allow local selection of the clustering, and generally better prediction.
  In particular, Peruzzi & Wade (2016+) compared the kernel stick-breaking model (Dunson & Park, 2008) and the normalized weights model (Antoniano, Wade, Walker, 2014). The latter appears to allow faster computations.

Computations

A substantial challenge for comparison is due to computations: one needs an algorithm that can give fast computations, and be fairly easily adapted to different models.

Peruzzi & Wade (2016+) suggest a clever algorithm, based on adaptive truncation and Metropolis-Hastings, moving from a proposal by Griffin (2016).

This is a necessary basis for providing R-packages for density regression and conditional density estimation.

contents

summary and discussion

I tried to give an overview of proposals for Bayesian density regression, based on mixture models.

Careful study of the finite-sample properties of these models, for comparison

Together with flexible, easily-exportable computational tools, these are necessary steps for providing R-packages for density regression.

references

- Wade, S., Mongelluzzo, S. & Petrone, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Analysis*, **3**, 359–386.
- Wade, S., Dunson, D., Petrone, S. & Trippa, L. (2014). Improving prediction from Dirichlet process mixtures via enrichment. *Journal of Machine Learning Research*, 1041–1071.
- Wade, S., Walker, S. & Petrone, S. (2014). A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting. *Scandin. J. Statist.*, **41**, 580–605.
- Wade, S., Peruzzi, M. & Petrone, S. (2016+) Review of Bayesian nonparametric mixture models for regression. (*manuscript*)