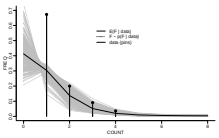
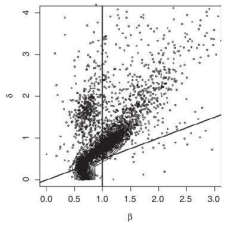
Slide 1

Nonparametric Bayesian Data Analysis Part 1 – Density Estimation

PETER MÜLLER, UT Austin





density estimation

www.math.utexas.edu/users/pmueller/ufscar

Slide 2

Outline

1. **Density estimation:** inference for a random distribution G
   - Dirichlet process (DP) & DP mixture
   - Polya tree (PT)
   - Normalized random measure (NRM)
2. **Random effects distributions:** density estimation for (latent) random effects
3. **Regression:** family of random dists $\{G_x,\ x \in X\}$
   - Dependent DP (DDP)
   - ANOVA DDP

1 Density Estimation

1.1 Example 1

Slide 3

Example 1: Time to resolution of air leaks
Xu et al. (2017 B.A.) model time to resolution of air leaks for a small size study of patients after pulmonary resection.

Data: T_i, time to resolution of air leaks after pulmonary resection;

Treatment: standard care ($Z_i = 0$) vs. liquid sealant "progel" ($Z_i = 1$)

Preferences: reduction by 1 day means more for early days than later!

- comparing means is inappropriate
- ... and impossible in this study (needs $n = 200$!)
- change in early resolution possible w/o change in mean, compare relative change in early resolution times.

Slide 4

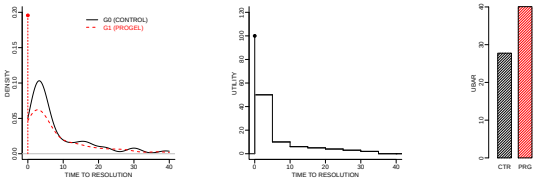
Model: details of the event time distribution $G_0(T_i)$ and $G_1(\cdot)$ matter → need a (flexible) prior $p(G_j)$;

- *Parametric Bayes:* assume $G \in \{G_\theta, \theta \in \Re^p\}$ → prior on G reduces to $p(\theta)$, $\theta \in \Re^p$. But restriction to, e.g., $G_{\mu,\sigma} = N(\mu, \sigma^2)$ is too strong!
- **BNP:** (nonparametric Bayes) prior on infinite dimensional space, like $p(G)$;

Bayes rule: preferences (utility) → Bayes rule for recommending S vs. P.

Slide 5

Ex 1: Time to resolution of air leaks



$G_0(T_i)$ and $G_1(T_i)$ $\times$ preferences u $=$ $\overline{V}_j$.

Bigger $\overline{V}$ (by at least $\epsilon = 18$) wins.

1.2 DP and DP mixture

Slide 6

Dirichlet process (DP)

The most commonly used BNP prior $p(F)$ for a random prob measure. (Ferguson, 1973 AnnStat).

$$F \sim \text{DP}(M, F^\star) \text{ with } F(y) = \sum_{h=1}^{\infty} p_h \delta_{m_h}(y).$$

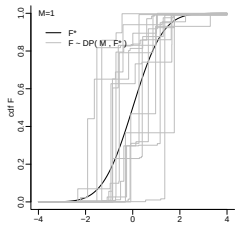
Locations: $m_h \sim F^\star$ i.i.d.

Weights: $p_h = v_h \prod_{\ell<h}(1-v_\ell)$ with $v_h \sim \text{Be}(1, M)$ (DN #1 "stick-breaking"),

Parameters: base measure, $F^\star(A) = E\{F(A)\}$ total mass M: $F(A) \sim \text{Be}(MF^\star(A), MF^\star(A))$

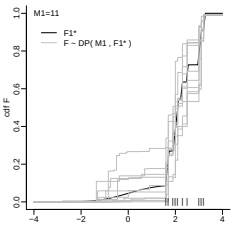
Posterior: easy – $p(y \mid F) = F$ and $G \sim \text{DP}(M, G_0) \rightarrow$ $p(G \mid y) = \text{DP}(M+1, G_1 \propto M \cdot G_0 + \delta_y)$

Slide 7



prior

(a) $F \sim \text{DP}(M, F^\star)$



posterior

(b) $F \mid \boldsymbol{x} \sim \text{DP}(M_1, F_1^\star)$
data = tick marks

Slide 8

DP Mixtures

DP random measure: discrete $F(\cdot)$ is awkward for many problems (e.g., gene expression example..)

DP mixture (DPM): convolution of discrete F with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

(DPM) continous $G(\cdot)$ (and hyperpar σ^2)[4pt] (Lo 1984 AnnStat; Escobar & West 1995 JASA);[4pt] Good review of DP(M) in Ghoshal (2010).

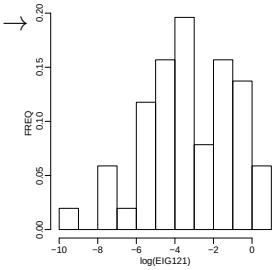
Slide 9

Example 2: Gene Expressions
Data from a biomarker trial to develop a screening test for high risk patients.

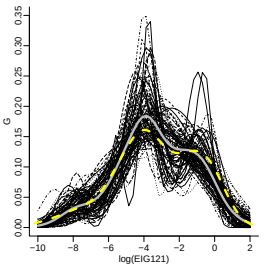
Data: y_i = EIG121 gene expression for $n = 51$ endometrial carcinoma patients.

Model: $y_i \sim G$ and $G \sim$ DPM

Posterior: simple finite DP implementation



(a) data y_i



(b) posterior draws $G \sim p(G \mid y)$

Slide 10

DPM as hierarchical model

Latent vars: write $y \sim \int \ldots dF(\theta)$ as hierarchical model

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n \\ \theta_i \mid F &\sim F \end{aligned}$$

Notation: discrete $F \Rightarrow K \leq n$ unique θ_i's $= \{\theta_1^\star, \ldots, \theta_K^\star\}$. Latent indicators $z_i = j$ iff $\theta_i = \theta_j^\star$ match θ_i with $\theta_j^\star$'s.

Slide 11

Polya Urn – Clustering

Clustering: unique $\theta_j^\star$ partition data into K clusters $S_j = \{i : \theta_i = \theta_j^\star\}$. Alternatively represent clusters by $(z_1, \ldots, z_n)$ with $z_i = j$ if $i \in S_j$

Marginal model: full prior $p(\boldsymbol{\theta}, F)$ allows marginal

$$p(\boldsymbol{\theta}|\cancel{F}) = p(\theta_1) \prod_{i=2}^{n} p(\theta_i \mid \theta_1, \ldots, \theta_{i-1}) \qquad (*)$$

Or equivalently as

$$p(\boldsymbol{z}) = \prod_{i=2}^{n} p(z_i \mid z_1, \ldots, z_{i-1})$$

– easy (**Polya urn**) → part II.

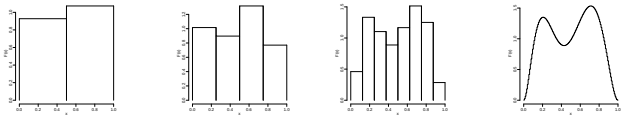
1.3 Polya Tree

Slide 12

1. Density Estimation (ctd)

Polya tree: random F as random histogram on bins given by a nested binary partitions, $\{B_0, B_1\}$, $\{B_{00}, B_{01}, B_{10}, B_{11}\}$ etc. Lavine (1992ab, AnnStat)

Sequence of random histograms ...



$F(B_{\epsilon_1})$ $F(B_{\epsilon_1\epsilon_2})$, $F(B_{\epsilon_1\epsilon_2\epsilon_3})$, $F(B_{\epsilon_1\epsilon_2\cdots\epsilon_m})$,

Slide 13

Polya Tree: $F \sim \text{PT}(\mathcal{A}, \Pi)$. Prior for $F(B_0), F(B_1), F(B_{00})$, etc.

- $F(B_{\epsilon 0} \mid B_\epsilon) \sim \text{Be}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$, independently across $\epsilon = \varepsilon_1 \cdots \varepsilon_m$. Arranged as tree (well, really "root" :-) Two pars: $\mathcal{A} = \{\alpha_\epsilon\}$, $\Pi = \{B_\epsilon\}$.
- prior centering at $E(F) = F^\star$
  1. B_ϵ = quantile sets under desired $F^\star$; *or*
  2. $\alpha_{\epsilon 0} = cF^\star(B_{\epsilon 0} \mid B_\epsilon)$ $\alpha_{\epsilon 1} = cF^\star(B_{\epsilon 1} \mid B_\epsilon)$

Continuous F: Choice of α_ϵ can guarantee continous F; important, e.g., for model validation (Berger & Guglielmi, 2001 JASA).

DP: DP = PT with $\alpha_\epsilon = \alpha_{\epsilon 0} + \alpha_{\epsilon 1}$ (DN #2 P.T.)

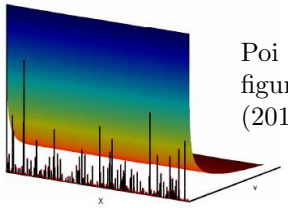
Mixture of PT: mixing w.r.t. α_ϵ and partition sequence (Hanson, 2006 JASA)

1.4 Normalized Random Measures (NRM)

Slide 14

Normalized Random Measures (NRM): random prob measure F on X: $F = \sum_h \eta_h \delta_{\mu_h}$ with $\eta_h = \frac{\tilde\eta_h}{\sum \tilde\eta_\ell}$ and

$\{(\mu_h, \tilde\eta_h),\ h = 1, \ldots\} \sim$ PoiP on $(X \times \Re^+)$ with Poi intensity $\nu(\tilde\eta, \mu)$



Poi intensity $\nu(\tilde\eta, \mu)$ figure from Favaro & Teh (2013, STS)

Regazzini et al. (2003, AnnStat) Lijoi et al. (2005 JASA; 2007 JRSSB).

Slide 15

- Elegant :-) Different Poi intensity function $\nu(\tilde\eta, \mu)$ gives different prior, but all are discrete.
- Still (relatively) easy posterior updating for mixture models (Barrios et al., 2013 STS; Favaro & Teh 2013 STS; Argiento et al. Comp Stat & Data Anal) All consider the normalized generalized gamma (NGG),

$$\nu(\tilde\eta, \mu) = \rho(\tilde\eta)\, \alpha H_0(\mu) \text{ with } \rho(\tilde\eta) = e^{-\kappa\tilde\eta}\tilde\eta^{-(1+\gamma)}/\Gamma(1-\gamma).$$

which includes as special cases the DP ($\kappa = 1, \gamma = 0$), normalized inverse Gaussian and others.

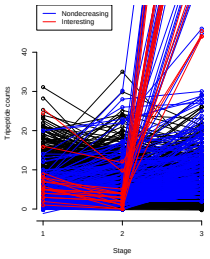
- DP is a special case - again :-), $\nu(\tilde\eta, \mu) = \ldots$ (DN #3 NRMI),

2 Random effects distributions

Slide 16

2. Random Effects Distributions

Example 3: Phage Display Data
León-Novelo & al., (2013 Bmcs).



Data: Counts (N_{i1}, N_{i2}, N_{i3}) for tripeptide/tissue pairs $i = 1, \ldots, n$

Aim: identify increasing counts

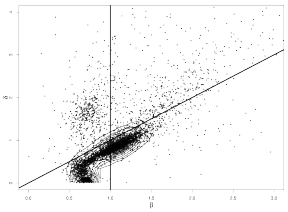
Sampling model: $\boldsymbol{N}_i \sim Poi(N_{i1}|\mu_i)\, Poi(N_{i2}|\mu_i\beta_i) Poi(N_{i3}|\mu_i\delta_i)$,

Random effects: Increments $(\beta_i, \delta_i) \sim G \Rightarrow$ Increasing $\Leftrightarrow$ $1 < \beta_i < \delta_i$

Hyper-Prior: $G \sim$ DPM

Slide 17

Posterior $E\{G(\beta, \gamma) \mid data\}$

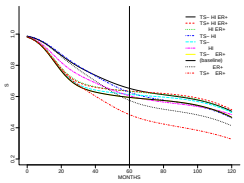


- Posterior prob's $p_i = p(1 < \beta_i < \delta_i \mid \boldsymbol{N})$ are adjusted for multiplicities (Scott & Berger, 2010 AnnStat);
- Use p_i for MCP; can show Bayes rule is "report if $p_i > p^\star$
- BNP prior on G is critical.

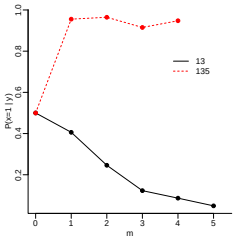
Slide 1

Nonparametric Bayesian Data Analysis Part 2 – Regression

Peter Müller, UT Austin



regression



classification

Slide 2

Outline

1. **Density estimation:** inference for a random distribution G
   - Dirichlet process (DP) & DP mixture
   - Polya tree (PT)
   - Normalized random measure (NRM)
2. **Random effects distributions:** density estimation for (latent) random effects
3. **Regression:** family of random dists $\{G_x,\ x \in X\}$
   - Dependent DP (DDP)
   - ANOVA DDP

3 Regression

Slide 3

2. Regression

Regression: $y_i \mid x_i = x \sim F_x(y_i)$.

1. **NP on residual:** $y_i = f_\theta(x_i) + \epsilon_i$, $\epsilon_i \sim G$ and $G \sim p(G)$. Semiparametric Bayes, density estimation for residuals ϵ_i, e.g., PT prior (Hanson & Johnson, 2002 JASA).
2. **Random regression mean function** : $y_i = f(x_i) + \epsilon_i$ and $f(\cdot) \sim p(f)$ GP prior, wavelet bases, neural networks, hierarchical mixture of experts, etc.
3. **Fully non-parametric regression:** $y_i \mid x_i \sim F_{x_i}$, with $\mathcal{F} = \{F_x,\ x \in X\} \sim p(\mathcal{F})$. For example, DDP model, dependent PT etc. Introduce the DDP next ...

3.1 Example

Slide 4

Example 4: BNP survival regression
de Iorio et al. (2009, Bmcs)

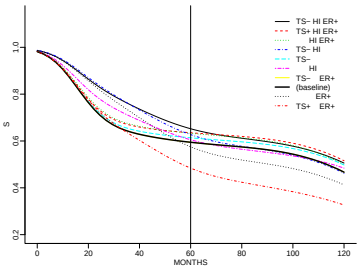
Event time y_i: event free survival for $n = 761$ women in a breast cancer study (Rosner, 2005 Bmcs)

Covariates $\boldsymbol{x}_i$: including indicators for high dose (HI), estrogen receptor (ER), tumor size (TS) and HI*ER interaction.

Model: $y_i \mid x_i \sim G_{x_i}$ and BNP prior $p(G_x,\ x \in X)$.

Slide 5

Posterior inference: $p(G_x \mid \boldsymbol{y})$. Interest is in the entire distribution – not just mean.



$E(S_x \mid \boldsymbol{y})$

(note: actually this figure is from another paper – but same data & study).

Prior: need a prior prob model $p(G_x,\ x \in X)$ for a family of related random prob measures.

3.2 Dependent Dirichlet Process (DDP)

Slide 6

Dependent Dirichlet Process (DDP)

MacEachern (1999, ASA proceedings). Prior $p(G_x,\ x \in X)$ with $G_x \sim$ DP marginally:

Marginal model for G_x: $G_x \sim DP(\cdot,\cdot)$, Stick breaking representation:

$$
\begin{aligned}
x = x_1: &\quad G_{x_1} = p_1\delta_{m_{11}} + p_2\delta_{m_{12}} + p_3\delta_{m_{13}} + \ldots \\
x = x_2: &\quad G_{x_2} = p_1\delta_{m_{21}} + p_2\delta_{m_{22}} + p_3\delta_{m_{23}} + \ldots \\
x = x_3: &\quad G_{x_3} = p_1\delta_{m_{31}} + p_2\delta_{m_{32}} + p_3\delta_{m_{33}} + \ldots
\end{aligned}
$$

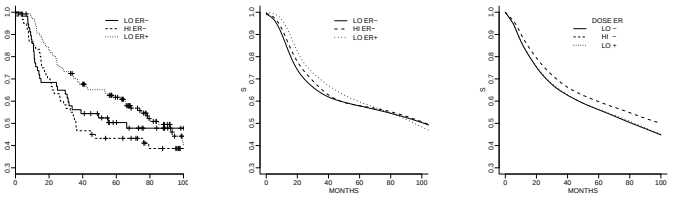
Dependent DP (DDP) : prior for dependent $\{G_x\}$

- Introduce dependence **across** x by assuming $m_1(x) = (m_{x1},\ x \in X)$, $m_2(x) = (m_{x2},\ x \in X)$ etc dependent **across** x.
- *Marginal DP:* Independence **across point masses** and stick breaking remains unchanged.

MCMC: easy – (almost) same as in DPM.

Slide 7

DDP prior for $\{G_x;\ x \in X\}$ in survival regression.



(a) data (KM) (b) ANOVA DDP (c) AFT

3.3 ANOVA DDP

Slide 8

ANOVA DDP

De Iorio *et al.* (2004, JASA)

- Special case ANOVA DDP: G_x for categorical $x \in X$
- Induce *dependence across x* by an ANOVA on m_{xh}. For example, assume $x = (v, w)$ with two categorical factors:

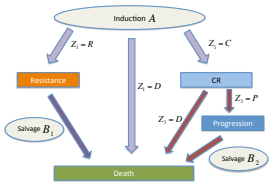
$$m_{xh} = \mu_h + \alpha_{h,v} + \beta_{h,w}.$$

Easy extension to include continuous covariates by ANCOVA.

- Used ANOVA DDP in the earlier survival regression example.

Slide 9

Example 5: Dynamic treatment regimen
Xu et al. (2016 JASA)



Motivating leukemia trial

Problem: Frontline therapy (A) is randomized, salvage therapy (B) is usually **not randomized**.
Adjust for the lack of randomization.

Aim: BNP approach to evaluate DTRs, using model-based inference to undo the lack of randomization.

- 4 **induction** trts: FAI, FAI+ATRA, FAI+GCSF, FAI+ATRA+GCSF.
- 2 **salvage** trts: HDAC or not.

Slide 10

BNP Model for Evaluating DTRs

Data:

Outcome: $Y^k = \log(T^k)$ = (log) k^{th} transition time (e.g., R → D)

Covariates: $\boldsymbol{x}^k$, incl. T^ℓ, $\ell < k$

Pars: $\mathcal{F} = \{F^k;\ k = 1, \ldots, K\}$, (unknown) distributions of 7 transition times

Likelihood:

$$\prod_{k=1}^{K} p(Y^k \mid \boldsymbol{x}^k, \mathcal{F}) = \prod_{k=1}^{K} F^k_{\boldsymbol{x}^k}(Y^k)$$

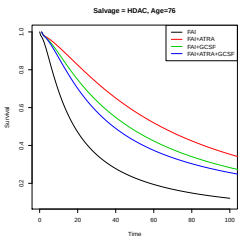
Prior: BNP prior for $\mathcal{F}$

$$\mathcal{F} = \{F^k_x;\ x \in X, \} \sim \text{DDP}, \quad k = 1, \ldots, K$$

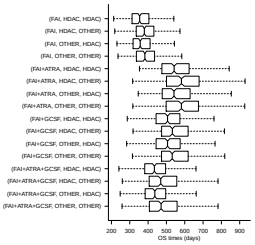
with $F^k_x = \sum_{h=0}^{\infty} w^k_h\, N(y;\ \theta_{hk}(\boldsymbol{x}^k), \sigma^k)$. GP prior on $\{\theta_{hk}(\boldsymbol{x})\}_{\boldsymbol{x}}$

Slide 11

Results: Survival regression and optimal policy



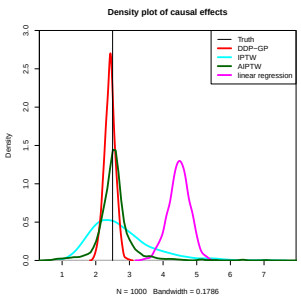
survival regr for T^{PD}



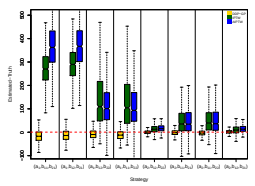
overall survival for alternative policies (A, B_1, B_2).

Slide 12

Comparison with double robust methods



single event time (correct model)



DTR, with both models wrong