Bayesian Nonparametric Estimation of Unimodal and Multimodal Distributions

George Kouvaras & George Kokolakis

National Technical University of Athens
Department of Applied Mathematics

First Athens - Pavia Meeting in Statistics
June 2008, Marathon - Athens

Outline

1. Introduction & Motivation
   - The Problem
   - General Background
2. Our Results/Contribution
   - Partial Convexification Procedure
   - Implementation

Our Goal

- Datasets are often derived from different groups and thus they are resulting to multimodal empirical distributions.
- A Bayesian nonparametric estimation procedure of unimodal and multimodal random distribution functions on a finite dimensional Euclidean space is introduced. As a result we get random probability measures that admit derivatives almost everywhere in $\mathbb{R}^d$.

Nonparametric Estimation Of CDF

- Let $X_1, \ldots, X_n$ be a random sample from an unknown c.d.f. F.
- If there is no assumption about the functional form of F, then the empirical distribution function is usually applied.

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{(-\infty,x]}(X_i), \text{ where } \delta_A(x) = \begin{cases} 1, & x \in A, \\ 0, & x \notin A. \end{cases}$$

- The problem of constructing Bayesian nonparametric estimators for F involves the constuction of a probability measure on the space $\mathcal{P}$ of c.d.f.'s endowed with a σ-field $\mathcal{S}$. The Dirichlet process can be used to define the prior information about F.

Dirichlet Process (Review)

Let $\mathcal{P}$ is the collection of all probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$, endowed with a σ-field $\mathcal{S}$.

Definition (Ferguson 1973,1974)

A random probability measure P is a Dirichlet Process (DP) on $(\mathcal{P}, \mathcal{S})$ with parameter α if for every $k = 1, \ldots$, and measurable partition $(B_1, \ldots, B_k)$ of $\mathcal{X}$,

$$(P(B_1), \ldots, P(B_k)) \sim D(\alpha(B_1), \ldots, \alpha(B_k)),$$

where $\alpha(.)$ is a non-null finite measure on $(\mathcal{X}, \mathcal{B})$.

Basic Properties of Dirichlet Process

If $P \sim DP(\alpha)$, then for any measurable sets A and B,

- $E(P(A)) = \frac{\alpha(A)}{\alpha(\mathcal{X})} \equiv \overline{\alpha}(A)$
- $Var(P(A)) = \frac{\alpha(A)\alpha(A^C)}{(\alpha(\mathcal{X}))^2(1+\alpha(\mathcal{X}))}$
- $Cov(P(A), P(B)) = \frac{(\alpha(\mathcal{X}))(\alpha(A\cap B)-\alpha(A)\alpha(B))}{(\alpha(\mathcal{X}))^2(1+\alpha(\mathcal{X}))}$
- $E(P(A) \mid X_1, \ldots, X_n) = \frac{\alpha(\mathcal{X})}{\alpha(\mathcal{X})+n}\overline{\alpha}(A) + \frac{n}{\alpha(\mathcal{X})+n}P_n(A),$

where P_n is the probability measure corresponding to the empirical c.d.f. F_n.

The normalized probability measure $\overline{\alpha}$ is called the *base distribution* and the total mass $\alpha(\mathcal{X})$ is usually called the *precision parameter*. The precision parameter controls the variability of any $P(A)$ around its prior mean. For instance, if someone expects the $N(0, 1)$ model to hold, but he is not quite confident about it, one way to solve the problem is to choose the normalized probability measure as $N(\mu_0, \sigma_0^2)$ and use a precision parameter reflecting the degree of confidence in this prior guess.

Some Properties of Dirichlet Process

- One of the most remarkable properties of DP is that the posterior distribution is again Dirichlet. Specifically we have the following:

  If $X_1, \dots, X_n \sim P$ and $P \sim DP(\alpha)$, then the posterior distribution $P \,|\, X_1, \dots, X_n \sim DP(\alpha^*)$, where $\alpha^* = \alpha + \sum_{i=1}^{n} \delta_{X_i}$.

- The major drawback of a Dirichlet process is that it selects discrete distributions with probability one.

Beyond Dirichlet Process

Several different classes of nonparametric priors, that all contain the Dirichlet process as a particular case, have been proposed. It seems worth mentioning, among others,

- The mixture of Dirichlet processes [Antoniak, 1974], which is a Dirichlet process where the base measure is random itself.
- The mixture of Dirichlet process prior [Lo, 1984], which is a convolution of a Dirichlet process with an appropriate kernel.
- Polya Trees [Lavine, 1992, 1994], Dirichlet Diffussion Trees [Neal, 2001],....

Nonparametric priors are extensively discussed in Walker et al., (1999), Hjort, (2003) and Ghosh, (2003).

Unimodality In $\mathbb{R}$

A univariate c.d.f. $F(x)$, $x \in \mathbb{R}$ is said to be unimodal with mode at zero, if F is convex on the negative real line and concave on the positive.

Some consequences of the above definition are:

- If F is unimodal about zero, then apart from a possible mass at zero, F is absolutely continuous.
- If F is unimodal about zero, then the left and right derivatives of F exist everywhere except possibly at zero.
- If F_1 and F_2 are both unimodal about zero, then $\lambda * F_1 + (1 - \lambda) * F_2$ is also unimodal about zero for every $\lambda \in [0, 1]$. The previous result clearly extends to mixtures involving more than two components.

Representation Results For Unimodality In $\mathbb{R}$

For univariate unimodal distributions there is a well known representation theorem due to Khinchin.

Theorem

A real valued random variable X is unimodal at zero if and only if it is a product of two independent random variables U and Y, with U uniformly distributed on $(0, 1)$ and Y having an arbitrary distribution.

Representation Results For Unimodality In $\mathbb{R}$

This can be expressed in the following equivalent form due to Shepp, (1962).

Theorem

The c.d.f. F is unimodal at zero, if and only if there exists a distribution function G on $\mathbb{R}$ such that F admits the representation:

$$F(x) = G(x) + xf(x),$$

for all x points of continuity of G.

Unimodality In $\mathbb{R}^d$

For multivariate distributions, however, there are several different ways that unimodality is defined. Among the main types of multivariate unimodality there are the following:

Definition (1)

The c.d.f. F is called *star unimodal* about **0** if it belongs to the closed convex hull of the set of all uniform distributions on sets S in $\mathbb{R}^d$ which are star-shaped about **0**.
(A set S in $\mathbb{R}^d$ is called star-shaped about **0**, if for any $\mathbf{x} \in S$ the line segment $[\mathbf{0}, \mathbf{x}]$ is contained in S).

Definition (2)

The c.d.f. F is called *linear unimodal* about **0** if the c.d.f. of any linear combination of the components of **X** is univariate unimodal about 0.

Unimodality In $\mathbb{R}^d$

Definition (3)

The $d-$variate c.d.f. F is called *block unimodal* (*Khinchin's classical unimodality*) about **0** if there exists a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with c.d.f. F, such that

$$(X_1, \ldots, X_d) = (Y_1 U_1, \ldots Y_d U_d),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{U} = (U_1, \ldots, U_d)$ are independent vectors, $\mathbf{U}$ is uniformly distributed on the rectangle $(0, 1) \times \ldots \times (0, 1)$ and $\mathbf{Y}$ having an arbitrary $d-$variate c.d.f. G.

Definition (4)

The c.d.f. F is called *beta unimodal* about **0** if it is generated by the Beta distribution, $Beta(\kappa, \nu)$, instead of Uniform distribution on the interval $(0, 1)$, and contains the block univariate unimodality as special case.

Unimodality In $\mathbb{R}^d$

An extended study of different types of unimodality and their useful consequences can be found in [Dharmadhikari and Kumar,1988] and [Bertin et al., 1997].
In what follows we focus our attention on *block unimodality*.

Partial Convexification - Multimodality

- Using the above procedure, i.e. by the component wise multiplication of two independent random vectors **Y** and **U**, where the latter is uniformly distributed on the unit d-cube, we always get a c.d.f. F with a single mode at zero, no matter what the distribution G, we start with, is.
- To overcome the limitation of getting always a c.d.f. F with a single mode at zero, we propose the following procedure, that we called **partial convexification** procedure that results to multimodal distributions.

Partial Convexification In $\mathbb{R}$

- Partial convexification of a c.d.f. G is based on using $U(\alpha, 1)$ distributions, with $0 < \alpha < 1$, instead of $U(0, 1)$. [Kokolakis and Kouvaras, 2007].
- The parameter α can be fixed, or random with a prior distribution $p(\alpha)$, on the interval $(0, 1)$.
- The expected number of modes of F increases from one, when $\alpha = 0$, to infinity, when $\alpha = 1$, having a finite number of modes when $0 < \alpha < 1$. This means that when $0 < \alpha < 1$, the c.d.f. $F(x)$ alternates between local concavities and local convexities, i.e a partial convexification of F is produced.

Partial convexification In $\mathbb{R}$

Using the partial convexification procedure, the representation theorem for unimodal distribution can be expressed [Kokolakis and Kouvaras, 2007] in the following form.

Theorem

If F is the partial convexification of G by $U(\alpha, 1)$ with $0 < \alpha < 1$ then the following result holds:

$$F(x) = xf(x) + \frac{1}{1-\alpha}[G(x) - \alpha G(\frac{x}{\alpha})],$$

for all x and x/α, points of continuity of G.

Partial convexification In $\mathbb{R}^2$

In the bivariate case, using Uniform distributions $U(\alpha_i, 1)$, $i = 1, 2$, with parameters α_i fixed in the interval $(0, 1)$, the following holds [Kouvaras and Kokolakis, 2008].

$$F(x_1, x_2) = x_1 F_{x_1}(x_1, x_2) + x_2 F_{x_2}(x_1, x_2) - x_1 x_2 f(x_1, x_2) + Q(x_1, x_2)$$

where

$$\begin{aligned} Q(x_1, x_2) = \frac{1}{(1-\alpha_1)(1-\alpha_2)} & \times \quad \{G(x_1, x_2) - \alpha_1 G(\frac{x_1}{\alpha_1}, x_2) \\ & - \quad \alpha_2 G(x_1, \frac{x_2}{\alpha_2}) + \alpha_1 \alpha_2 G(\frac{x_1}{\alpha_1}, \frac{x_2}{\alpha_2})\}, \end{aligned}$$

for all (x_1, x_2), $(x_1/\alpha_1, x_2)$, $(x_1, x_2/\alpha_2)$ and $(x_1/\alpha_1, x_2/\alpha_2)$ points of continuity of G.

Partial Convexification in $\mathbb{R}^d$

Definition

The $d-$variate c.d.f. F is called partially convexified if there exists a random vector $\mathbf{X} = (X_1, \dots, X_d)$ with c.d.f. F, such that

$$(X_1, \dots, X_d) = (Y_1 U_1, \dots Y_d U_d),$$

where $\mathbf{Y} = (Y_1, \dots, Y_d)$ and $\mathbf{U} = (U_1, \dots, U_d)$ are independent vectors, $\mathbf{U}$ is uniformly distributed on the rectangle $(\alpha_1, 1) \times \dots \times (\alpha_d, 1)$ and $\mathbf{Y}$ having an arbitrary $d-$variate c.d.f. G.

Partial convexification in $\mathbb{R}^d$

If F is the partial convexification of G by $U(\alpha_i, 1)$ with $0 < \alpha_i < 1$, $(i = 1, \cdots, d)$ then the following result holds:

$$\begin{aligned} F(\mathbf{x}) &= \sum_{i=1}^{d} x_i F_{x_i}(\mathbf{x}) - \sum\sum_{1\le i<j\le d} x_i x_j F_{x_i x_j}(\mathbf{x}) + \sum\sum\sum_{1\le i<j<k\le d} x_i x_j x_k F_{x_i x_j x_k}(\mathbf{x}) \\ &- \cdots + (-1)^{d-1} f(\mathbf{x}) \prod_{i=1}^{d} x_i + Q(\mathbf{x}), \end{aligned}$$

where,

$$\begin{aligned} Q(\mathbf{x}) &= \frac{1}{\prod_{i=1}^{d}(1-\alpha_i)} \times \{G(\mathbf{x}) - \sum_{i=1}^{d} \alpha_i G_i(\mathbf{x}) + \sum\sum_{1\le i<j\le d} \alpha_i \alpha_j G_{i,j}(\mathbf{x}) \\ &- \sum\sum\sum_{1\le i<j<k\le d} \alpha_i \alpha_j \alpha_k G_{i,j,k}(\mathbf{x}) + \cdots + (-1)^d G_{1,2,\cdots,d}(\mathbf{x}) \prod_{i=1}^{d} \alpha_i \} \end{aligned}$$

$$G_{i,j,k}(\mathbf{x}) = G(x_1, \cdots, \frac{x_i}{\alpha_i}, \cdots, \frac{x_j}{\alpha_j}, \cdots, \frac{x_k}{\alpha_k}, \cdots, x_d)$$

with all involved arguments, points of continuity of G.

Our Model

In our Bayesian nonparametric model specification we assume the following:

- $\mathbf{Y} = (Y_1, \cdots, Y_d)^T \sim G$, where G is a random c.d.f. produced by a DP in $\mathbb{R}^d$.
- $\mathbf{U} = (U_1, \cdots, U_d)^T$ is uniformly distributed on the d-cube $(\alpha_1, 1) \times \cdots \times (\alpha_d, 1)$ with α_i $(i = 1, \cdots, d)$ fixed on the interval $(0, 1)$.
- $\mathbf{Y}$ and $\mathbf{U}$ are independent.

Then $\mathbf{X} = \mathbf{U} \otimes \mathbf{Y} = (U_1 Y_1, \cdots, U_d Y_d) \sim F$ partially convexified.

Application

- We take a random sample with size $n = 300$ corresponding to the c.d.f. $F(x) = w_1 N(x \mid \mu_1, \sigma_1) + w_2 N(x \mid \mu_2, \sigma_2)$ with $w_1/w_2 = 3/2$ where $(\mu_1, \sigma_1) = (15, 3)$ and $(\mu_2, \sigma_2) = (30, 4)$ and obtain the empirical c.d.f. F_n.
- We deconvolute F_n by applying a variant of the rejection method. A randomly chosen Y is rejected unless $f_n(UY) \leq V$ with U, V independent, uniformly distributed in the intervals $(\alpha, 1)$ and $(0, M)$ respectively, where $M = sup(f_n(x))$.
- The empirical c.d.f. G_n of the accepted $Y's$ is used to update a $DP(\beta)$ with $\beta(\mathcal{X}) = 10$ and normalized parameter measure $\overline{\beta}(.)$ corresponding to the c.d.f. *Gamma* with shape parameter $\kappa = 25$ and scale parameter $\theta = 2$.
- We derive a random sample from a $DP[\beta^*(.)]$ and apply the partial convexification procedure, with $\alpha = 0.5$.

Results

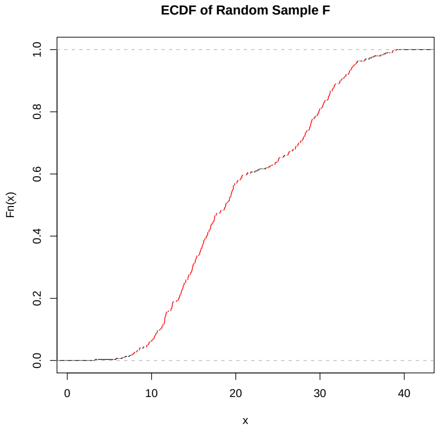

Figure: Empirical Distribution F_n

Results

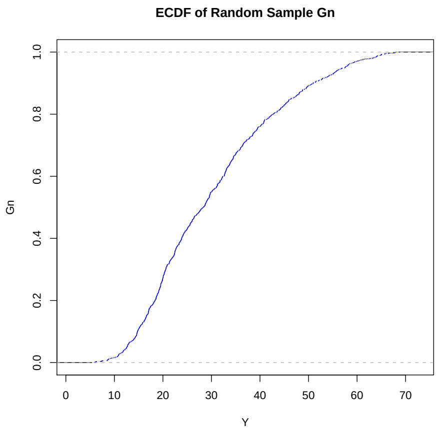

Figure: Empirical Deconvoluted G_n

Results

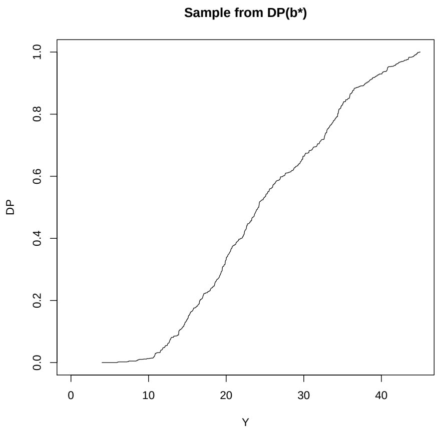

Figure: Empirical from $DP(\beta^*)$

Results

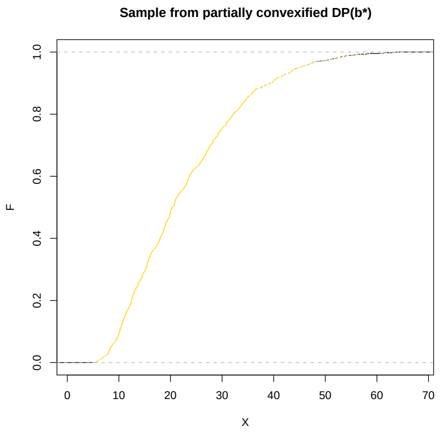

Figure: Empirical from partially convexified $DP(\beta^*)$

Results

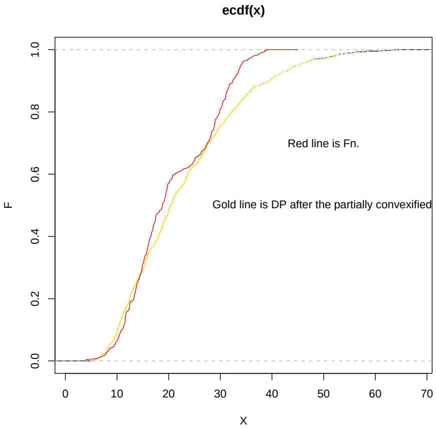

Figure: Comparison between the empirical F_n and the partially convexified $DP(\beta^*)$ with $\alpha = 0.5$

Results

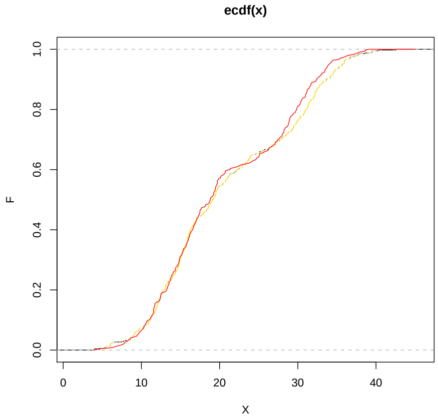

Figure: Comparison between the empirical F_n and the partially convexified $DP(\beta^*)$ with $\alpha = 0.9$

Results

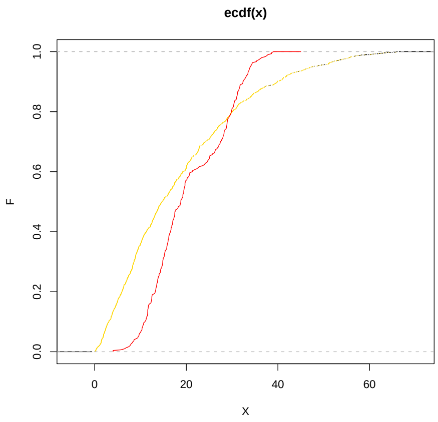

Figure: Comparison between the empirical F_n and the partially convexified $DP(\beta^*)$ with $\alpha = 0.01$

Summary

- The partial convexification procedure is a flexible, simple and efficient method.
- Our model pins down the number of modes of the random probability measure produced between the extremes *one* and *infinity*.
- Based on the results so far the predictive distributions have captured the characteristics of the simmulated data sets.

- Future Research
    - Put a prior on the parameter α.
    - Investigate hidden properties.
    - Work with other types of multimodality. (Star, Beta,...).
    - Work with other types of random processes.

References I

Ramamoorthi R. Ghosh, J.
Bayesian Nonparametrics.
New York: Springer-Verlag, Inc, 2003.

N. Hjort.
Topics in nonparametric Bayesian statistics.
In N. Hjort P. Green and S. Richardson, editors, *Highly structured stochastic systems*. Oxford: Oxford University Press, 2003.

Muller P. Sinha D. Dey, D.
Practical nonparametric and semiparametric Bayesian Statistics.
New York: Springer-Verlag, Inc, 1998.

S. Dharmadhikari and J-D. Kumar.
Unimodality, Convexity and Applications.
Academic Press, 1988.

W. Feller.
An Introduction to Probability Theory and its Applications, Volume 2.
Wiley, New York, 2nd edition, 1971.

Ch. Antoniak.
Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.
Ann. Statist., 2:1152–1174, 1974.

E.M-J. Bertin, I. Cuculescu, and R. Theodorescu.
Unimodality of Probability Measures.
Kluwer Academic Publishers, 1997.

References II

D. Blackwell.
Discreteness of Ferguson selections.
Annals of Statistics, 1:356–358, 1973.

L.J. Brunner.
Bayesian nonparametric methods for data from a unimodal density.
Statistics and Probability Letters, 14:195–199, 1992.

Th. S. Ferguson.
A Bayesian analysis of some nonparametric problems.
Annals of Statistics, 1:209–230, 1973.

G. Kokolakis and P. Dellaportas.
Hierarchical modelling for classifying binary data.
In *Bayesian Statistics 5*, pages 647–652, London, 1996. Oxford University Press.

G. Kokolakis and G. Kouvaras.
On the multimodality of random probability measures.
Bayesian Analysis, 2:213–220, 2007.

G. Kouvaras and G. Kokolakis.
Random Multivariate Multimodal Distributions.
Recent Advances in Stochastic Modelling and Data Analysis, World Scientific. (to appear).

References III

M. Lavine.
Some aspects of Polya tree distributions for statistical modelling.
Annals of Statistics, 20:1222–1235, 1992.

M. Lavine.
More aspects of Polya tree distributions for statistical modelling.
Annals of Statistics, 22:1161–1176, 1994.

A.Y. Lo.
On a class of nonparametric estimates: I. Density estimates.
Annals of Statistics, 12:351–357, 1984.

R. Neal.
Defining priors for distributions using Dirichlet diffussion trees.
Technical report, Department of Statistics, University of Toronto, March 2001.

L.A. Shepp.
Symmetric random walk.
Transactions of the American Mathematical Society, 104:144–153, 1962.

S. Walker, P. Damien, P.W. Laud, and A.F.M. Smith.
Bayesian nonparametric inference for random distributions and related functions.
J. Roy. Statist. Soc., Ser. B, 61:485–527, 1999.

Thank you for your attention!!