Applied Mathematical Sciences, Vol. 2, 2008, no. 33, 1619 - 1633

Bayesian Inference for Lancaster Probabilities

D. M. Cifarelli

Department of Decision Sciences, Universitá "L. Bocconi"
and
Universitá LUM "Jean Monnet"

R. Graziani

Department of Decision Sciences, Universitá "L. Bocconi"

E. Melilli

Department of Decision Sciences, Universitá "L. Bocconi"
eugenio.melilli@uni-bocconi.it

Abstract

Inference for bivariate distributions with fixed marginals is very important in applications. When a bayesian approach is followed, the problem of defining a (prior) distribution on a class of probabilities having given marginals arises. We consider the class of Lancaster distributions. It is a convex and compact set, so that any element may be represented as a mixture of extreme points. Therefore a prior distribution can be assigned to the Lancaster class by assuming the mixing measure as a random probability. We analyse in detail the Lancaster class with Gamma marginals. Choosing as mixing measure a Dirichlet process, the model turns out to be a Dirichlet process mixture model. Many quantities relevant for statistical purposes are linear functionals of the Dirichlet process. Posterior laws are determined; in order to approximate these laws a MCMC algorithm is suggested. Results of an example with simulated data are discussed.

Mathematics Subject Classification: 62F30, 62E10, 62G05

Keywords: Dirichlet Process, Distributions with given Marginals, Markov Chain Monte Carlo Methods, Mixture Models, Nonparametric Bayesian Inference

1 Introduction

Let us consider two quantitative characteristics whose joint distribution, in a statistical population, is unknown. In some cases it is reasonable to assume as known (at least approximately) the corresponding marginals. This usually happens when observations on each characteristic can be obtained for the whole population (or for large subsets of it), while only small sets of data are available on the joint distributions. For instance, [4] considers 1940 Census of Population, observing that "although there will be a complete count of certain characters for the individuals in the population, consideration of efficiency will limit to a sample many of the cross-tabulations (joint distributions) of these characters".
This problem was considered, within a classical approach, in the already cited paper by Deming and Stephan and discussed in some generality in [10] and [14]. According to a bayesian approach, the problem described above can be formalized as follows. Let $((X_n, Y_n))_{n\geq 1}$ be an exchangeable sequence of random vectors and denote by P the directing measure of the sequence; that is, given P, the elements of the sequence are independent and identically distributed according to P. P is then a random distribution whose marginals, denoted by P_1 and P_2 respectively, are assumed to be known. Therefore bayesian inference on P requires the assignment of a prior distribution on a class of probabilities having fixed marginals. In other words, a probability measure on the so-called Fréchet class generated by P_1 and P_2 is needed. This is in general a very hard problem, since probabilities in this class are subject to constraints difficult to deal with.
The problem could be addressed starting from the convexity of the Fréchet class. Such property makes it possible to represent any distribution P belonging to a given Fréchet class as a mixture over the extreme points of the class. In the case of marginal distributions having finite supports, this approach has been already followed in [18]. It is difficult to follow the same approach in the case of general (in particular continuous) marginals, since it is not available a useful characterization of the extreme points of the Fréchet class with non-finite support marginals; see for this problem [1]. Then, a possible way to address the problem is to restrict the attention to subclasses of the Fréchet class, which should be large enough to describe different kinds and degrees of association between the characteristics and, at the same time, should have good properties both from a mathematical and a computational point of view. Many classes of bivariate distributions with fixed marginals are known in literature; see, for a review, [9]. Being indexed by few parameters, such classes are generally not very large; a nonparametric approach appears more suitable. As observed in [12] and [16], a useful proposal in this direction is due to H.O.Lancaster (see [15]). Lancaster proposed a representation for bivariate distributions having

the so-called ϕ^2-boundedness property, i.e. distributions for which the Pearson mean square contingency index ϕ^2 is finite. In this representation, any such bivariate distribution is completely characterized by its marginal distributions and the (infinite) matrix of correlation coefficients between pairs of elements of complete sequences of orthonormal polynomials with respect to the marginals. The present paper considers the completion of the class of those ϕ^2-bounded bivariate distributions for which the above matrix is diagonal. This class is known as *Lancaster class* and it is parameterized by an infinite sequence of real numbers. Since, as proved in [13] by Koudou, Lancaster class is convex and compact, the approach described above can be followed and hence each element of the class may be represented as a mixture over its extreme points. Moreover, the extreme points of the Lancaster class can be, in some cases, identified and their explicit form can be obtained. Hence the problem of assigning a prior distribution (and then computing corresponding posterior quantities) on a Lancaster class can be solved by representing each element P of it as a mixture over the extreme points of the class and then by treating the mixing measure as a random probability. Choosing as mixing measure a Dirichlet process, the model turns out to be a Dirichlet process mixture model, as introduced in [5], discussed in [17] and deeply investigated in the last years. In this paper, the Lancaster class with Gamma marginals is analyzed in detail, its extreme points having a particular straight form. Interesting results are obtained: many quantities relevant for statistical purposes turn out to be linear functionals of the Dirichlet process. Many works in the recent literature address this topic.
As it is known, explicit forms for the distributions of these functionals are very complex and it is necessary to resort to simulation procedures in order to obtain suitable approximations. Following [11], a MCMC algorithm is suggested allowing to obtain the posterior estimates of interest. An example is proposed in which posterior estimates are obtained applying this algorithm to a simulated data set.
The set-up of the paper is as follows. Section 2 recalls some properties of Lancaster classes of probabilities. In Section 3 a family of prior distributions on the Lancaster class with Gamma marginals is introduced and discussed. Posterior computations are described in Section 4. Finally, in the last section, a numerical example is proposed.

2 Preliminary Notes

Let G and H be (univariate) distribution functions; in the paper, the same symbol will be used to denote a probability measure (on $\Re$ or $\Re^2$) and the corresponding distribution function.
Let us suppose that G and H have all moments and are determined by the

moments; sufficient conditions for this to hold are well known in the literature. We consider the class $L(G,H)$ of bivariate distribution functions with marginals G and H defined as follows:

$$L(G,H)=\left\{F: E_F\left(Y^n|X=x\right)=\sum_{i=0}^{n}a_i x^i, E_F\left(X^n|Y=y\right)=\sum_{i=0}^{n}b_i y^i, n\geq 1\right\},$$

(X,Y) being a random vector distributed according to F and E_F the expected value with respect to F. The elements of $L(G,H)$ are called *Lancaster probabilities*. $L(G,H)$ is, then, the class of all bivariate distributions, with marginals G and H, whose conditional moments of any order n are polynomials with degree less than or equal to n. In particular, the regression functions of Lancaster probabilities are linear. The class $L(G,H)$ was introduced in 1958 by H. O. Lancaster and studied by several authors; see, for instance, [15],[19] and [7]. $L(G,H)$ has many interesting properties from both a probabilistic and a statistical point of view; for a review of such properties see [13]. The following characterization of $L(G,H)$ is of particular interest: F is an element of $L(G,H)$ if and only if there exists a sequence of real numbers $\rho=(\rho_n)_{n\geq 0}$ such that

$$E_F\left(\xi_n(X)\eta_m(Y)\right)=\begin{cases}\rho_n & \text{if } n=m\\ 0 & \text{if } n\neq m.\end{cases} \tag{1}$$

$\xi=(\xi_n)_{n\geq 0}$ and $\eta=(\eta_n)_{n\geq 0}$ (with $\xi_0=\eta_0=1$) are orthogonal polynomials complete sequences on the Hilbert spaces $\mathcal{L}_2(G)$ and $\mathcal{L}_2(H)$, respectively. The sequence $\rho=(\rho_n)_{n\geq 0}$ characterizes F; that is, if F and F' are elements of $L(G,H)$ such that, for every $n\geq 0$, $E_F\left(\xi_n(X)\eta_n(Y)\right)=E_{F'}\left(\xi_n(X)\eta_n(Y)\right)=\rho_n$, then $F=F'$. $\rho=(\rho_n)_{n\geq 0}$ is called *Lancaster sequence* corresponding to F. Hence there is a one-to-one correspondence between the class $S(G,H)$ of all Lancaster sequences and $L(G,H)$. Observe that ρ_1 is, up to the sign, the Pearson correlation coefficient of the random vector (X,Y) (distributed according to the Lancaster probability F corresponding to the sequence ρ). $L(G,H)$ and the corresponding class $S(G,H)$ are convex and, under suitable topologies, compact sets; see [13].

Moreover the Lancaster probabilities are absolutely continuous, under weak conditions on the corresponding Lancaster sequences. More precisely, if G and H are absolutely continuous with respect to the measures $\mathcal{G}$ and $\mathcal{H}$, respectively, with densities g and h, and $F\in L(G,H)$ is such that its corresponding Lancaster sequence $\rho=(\rho_n)_{n\geq 0}$ satisfies

$$\sum_{n=0}^{+\infty}\rho_n^2<+\infty, \tag{2}$$

then F is absolutely continuous with respect to the product measure $\mathcal{G}\times\mathcal{H}$ and a version of the corresponding density has the following $\mathcal{L}_2(\mathcal{G}\times\mathcal{H})$

representation

$$f(x,y) = g(x)h(y)\sum_{n=0}^{+\infty}\rho_n\xi_n(x)\eta_n(y); \tag{3}$$

for a proof of this result, see [15].
The quantity $\sum_{n=0}^{+\infty}\rho_n^2$, when finite, coincides with the Pearson mean square contingency index ϕ^2.

3 The model

In this section, we propose a family of prior distributions on the Lancaster class whose marginals are Gamma distributions.
Let G and H be the distribution functions corresponding to the densities (with respect to Lebesgue measure on $(0,+\infty)$) $g(x) = (\Gamma(\alpha))^{-1}x^{\alpha-1}\exp(-x)$ and $h(x) = (\Gamma(\beta))^{-1}x^{\beta-1}\exp(-x)$ respectively, with $0 < \alpha \leq \beta$. In this case, the orthonormal basis for the spaces $\mathcal{L}_2(G)$ and $\mathcal{L}_2(H)$ are the sequences of standardized Laguerre polynomials $(L_n^\alpha)_{n\geq 0}$ and $(L_n^\beta)_{n\geq 0}$, where

$$L_n^\alpha(x) = (\Gamma(\alpha+n)\Gamma(\alpha)n!)^{1/2}\sum_{m=0}^{n}\frac{(-1)^m}{m!(n-m)!\Gamma(\alpha+m)}x^m.$$

Our aim is to define a family of probability measures on $L(G,H)$ to be used as priors for bayesian inference, as described in the Introduction. Since, as already observed, $L(G,H)$ is a convex and compact set, we can follow the general approach towards the construction of prior distributions on convex sets of probabilities suggested by P. Hoff in [8].
Hoff considers a convex and compact class $\mathcal{P}$ of probabilities and the subset $\mathcal{E}$ of the extreme points of $\mathcal{P}$. An application of the Choquet theorem makes it possible to express each element P of $\mathcal{P}$ as a mixture on $\mathcal{E}$. More precisely, for each P in $\mathcal{P}$, there exists a probability measure μ on $\mathcal{E}$ such that

$$P = \int_{\mathcal{E}} q\mu(dq). \tag{4}$$

If μ is a random probability measure, P turns out to be a random probability measure too. The arising model is then a mixture model. Representation (4) makes it possible to transform the (general very difficult) problem of assigning a distribution on a class of constrained probability measures into a non-constrained problem.
Let us go back to the class $L(G,H)$, with G and H being gamma distributions as defined at the beginning of this section. In order to follow Hoff approach, the extreme points of such class are needed. Their characterization is given in the following theorem, whose proof can be found in [7]; for $a > 0$, $(a)_n$ is Pochammer symbol, i.e. $(a)_n = \Gamma(a+n)/\Gamma(a)$.

Theorem 3.1 *A distribution function $F \in L(G,H)$ is an extreme point of this class if and only if its corresponding Lancaster sequence is $\rho = (\rho_n)_{n\geq 0}$, with*

$$\rho_n = \frac{\sqrt{(\alpha)_n}}{\sqrt{(\beta)_n}} t^n \tag{5}$$

for some $t \in [0,1]$.

Then, denoting by F_t^* the distribution function in $L(G,H)$ corresponding to the (extreme) Lancaster sequence $(\frac{\sqrt{(\alpha)_n}}{\sqrt{(\beta)_n}} t^n)_{n\geq 0}$, any $F \in L(G,H)$ can be represented as follows:

$$F = \int_{[0,1]} F_t^* \mu(dt). \tag{6}$$

Theorem 3.1 along with the condition for absolute continuity of Lancaster probabilities shows that, if $t < 1$, all extreme points F_t^* are absolutely continuous distributions. Indeed

$$\sum_{n=0}^{+\infty} \rho_n^2 = \sum_{n=0}^{+\infty} \left\{ \frac{\sqrt{(\alpha)_n}}{\sqrt{(\beta)_n}} t^n \right\}^2 \leq \sum_{n=0}^{+\infty} t^{2n} < +\infty.$$

The remaining extreme distribution F_1^* is singular with respect to the Lebesgue measure on $(0,+\infty)^2$.
The density of the extreme point F_t^*, for $t < 1$, has the following series expansion:

$$f_t^*(x,y) = \mathrm{e}^{-(x+y)} \frac{x^{\alpha-1} y^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \sum_{m=0}^{+\infty} \frac{\sqrt{(\alpha)_m}}{\sqrt{(\beta)_m}} t^m L_m^{\alpha}(x) L_m^{\beta}(y)(x,y). \tag{7}$$

Griffiths in [8] shows that the series in (7) converges pointwise for all positive x, y.
It is interesting to establish when the mixture F in (6) has density, i. e. when the Lancaster sequence ρ corresponding to F satisfies condition (2). The following theorem, due to Griffiths (see [8]), gives a sufficient condition on the mixing measure μ for F to be absolutely continuous with square integrable density.

Theorem 3.2 *If there exists $\varepsilon > 0$ such that*

$$\lim_{t\to 0} \frac{\mu\left((1-t,1]\right)}{t^{(\alpha-\beta+1)/2+\varepsilon}} = 0 \tag{8}$$

then $F = \int_{[0,1]} F_t^* \mu(dt)$ is absolutely continuous and its density has a (convergent) series representation as in (7).

Roughly, Theorem 3.2 states that if the mass given by the mixing measure μ near 1 is negligible than F is absolute continuous. Of course, the condition is satisfied when 1 does not belong to the support of μ.
In order to follow Hoff approach, μ will be considered as a random probability measure. We choose a Dirichlet process for μ for several reasons. First, Dirichlet processes have well known good properties and have been widely studied, so that a large number of both theoretical and computational results are available on them. Second, as proved by Dalal in [3], all prior distributions can be approximated resorting to mixtures of distributions of Dirichlet processes, so that the choice of the Dirichlet process can be considered in some sense a general one.
Let μ be a Dirichlet process on $[0,1]$ with base measure $c\mu_0$, where c is a positive number and μ_0 a probability measure on $[0,1]$. Then, the random probability measure F in (6) turns out to be a mixture of known distribution functions having a Dirichlet process as mixing measure. These particular mixture models are widely studied in literature; see [5], [6] and [17].
Since $\rho_n = \int_{[0,1]} \frac{\sqrt{(\alpha)_n}}{\sqrt{(\beta)_n}} t^n \mu(dt)$, $\rho = (\rho_n)_{n\geq 0}$ is, up to constants, the sequence of all moments of the Dirichlet process μ. In particular the Pearson correlation coefficient of F is, up to a multiplicative constant, a random Dirichlet mean. Many papers have been published on this topic; see, for instance, [2].

4 Main Results

Let us consider the model described in the previous section, that is

$$F = \int_{[0,1]} F_t^* \mu(dt),$$

where F_t^* is the distribution function defined after Theorem 3.1 and μ is a Dirichlet process with base measure $c\mu_0$; we will denote by $\tau_{c\mu_0}$ the distribution of μ. Next proposition shows that if μ_0 satisfies (8), then (8) holds (almost surely) for μ too .

Theorem 4.1 *If μ_0 is a measure on $[0,1]$ such that there exists $\varepsilon > 0$ for which*

$$\lim_{t\to 0} \frac{\mu_0\left((1-t,1]\right)}{t^{(\alpha-\beta+1)/2+\varepsilon}} = 0 \tag{9}$$

and c is any positive number, then μ satisfies with probability 1 condition (8).

Proof

Fix $\varepsilon > 0$; it is enough to show that $\lim_{n\to+\infty} \frac{\mu((1-t_n,1])}{t_n^{(\alpha-\beta+1)/2+\varepsilon}} = 0$ for any arbitrary sequence $(t_n)_{n\geq 1}$ such that $t_n \to 0$ for $n \to +\infty$; we show such result

for $t_n = n^{-1}$, the proof being the same for any other sequence converging to zero for $n \to +\infty$.
Let $r_n = \mu_0((1-1/n, 1])$ and let us consider, for arbitrary positive δ, the event

$$A_n = \left\{ \mu((1-1/n, 1]) > \frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right\}.$$

We will show that $\sum_{n=1}^{+\infty} P(A_n) < +\infty$, so that, by Borel-Cantelli Lemma, $P(A_n, i.o.) = 0$ and the thesis follows.
Since, by the hypothesis and the properties of the Dirichlet process, the random variable $\mu((1-1/n, 1])$ has distribution beta with parameters cr_n and $c - cr_n$, then

$$P(A_n^c) = \int_0^{\frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}}} \frac{\Gamma(c)}{\Gamma(cr_n)\Gamma(c-cr_n)} t^{cr_n-1}(1-t)^{c-cr_n-1} dt$$

$$= \frac{\Gamma(c)}{\Gamma(cr_n)\Gamma(c-cr_n)} \left(\frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right)^{cr_n} \int_0^1 y^{cr_n-1} \left(1 - \frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} y \right)^{c-cr_n-1} dy$$

$$= \frac{\Gamma(c)}{cr_n\Gamma(cr_n)\Gamma(c-cr_n)} \left(\frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right)^{cr_n} {}_2F_1 \left(cr_n - c + 1, cr_n, cr_n + 1, \frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right).$$

By hypothesis, ${}_2F_1 \left(cr_n - c + 1, cr_n, cr_n + 1, \frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right)$ and $\left(\frac{\delta}{n^{(\alpha-\beta+1)/2+\varepsilon}} \right)^{cr_n}$ are asymptotically equivalent ($\sim$) to 1 , so that

$$P(A_n^c) \sim \frac{\Gamma(c)}{cr_n\Gamma(cr_n)\Gamma(c-cr_n)}$$

or, equivalently,

$$P(A_n) \sim 1 - \frac{1}{cr_n\Gamma(cr_n)}$$

$$= 1 - \frac{1}{cr_n} \frac{\sin(\pi cr_n)}{\pi} = \frac{1}{6}\pi^2 c^2 r_n^2 + o(r_n^2)$$

and the proof is complete since $\sum_n r_n^2$ converges by hypothesis.

Hence, if μ_0 satisfies (8), the mixture F has, with probability one, a density f. Then (6) is equivalent to

$$f = \int_{[0,1)} f_t^* \mu(dt), \tag{10}$$

where f_t^* is the density of F_t^*. Observe that, for fixed (x, y) in $(0, +\infty)^2$, $F(x, y)$ and $f(x, y)$ are linear functionals of the Dirichlet process μ.

In the remainder of the paper, posterior and predictive distributions and estimates are derived. μ_0 is assumed to satisfy condition (8).

By introducing latent random variables $T_1, T_2, \ldots, T_n$, with values in $[0,1]$ (6) is equivalent to the following hierarchical model; see [17] and [8] for a general discussion of this equivalence:

$$\mathcal{L}((X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n)|T_1=t_1,\cdots,T_n=t_n)=\prod_{i=1}^{n}F^*_{t_i}; \tag{11}$$

$$\mathcal{L}(T_1,\cdots,T_n|\mu)=\mu^n; \qquad \mathcal{L}(\mu)=\tau_{c\mu_0}; \tag{12}$$

$((X_1,Y_1),(X_2,Y_2),\cdots,(X_n,Y_n))$ and μ are independent given $(T_1,\cdots,T_n)$.

Now consider the following measures on the class of Borel sets of $[0,1]^n$:

$$\lambda(A_1\times A_2\times\ldots\times A_n)=\int_{\bigotimes_{i=1}^n A_i}\prod_{i=1}^{n}\frac{c\mu_0+\sum_{j=1}^{i-1}\delta_{t_j}}{c+i-1}(dt_i) \tag{13}$$

and, for each n-uple $[x,y]=[(x_1,y_1),\ldots,(x_n,y_n)]\in((0,+\infty)^2)^n$,

$$\gamma_{[x,y]}(A_1\times A_2\times\ldots\times A_n)=\frac{\int_{\bigotimes_{i=1}^n A_i}\prod_{i=1}^{n}f^*_{t_i}(x_i,y_i)\lambda(dt_1,dt_2,\ldots,dt_n)}{\int_{[0,1]^n}\prod_{i=1}^{n}f^*_{t_i}(x_i,y_i)\lambda(dt_1,dt_2,\ldots,dt_n)}$$

$$=\frac{\int_{\bigotimes_{i=1}^n A_i}\prod_{i=1}^{n}\left\{\sum_{m=0}^{+\infty}\frac{\sqrt{(\alpha)_m}}{\sqrt{(\beta)_m}}t_i^m L_m^\alpha(x_i)L_m^\beta(y_i)\right\}\lambda(dt_1,dt_2,\ldots,dt_n)}{\int_{[0,1]^n}\prod_{i=1}^{n}\left\{\sum_{m=0}^{+\infty}\frac{\sqrt{(\alpha)_m}}{\sqrt{(\beta)_m}}t_i^m L_m^\alpha(x_i)L_m^\beta(y_i)\right\}\lambda(dt_1,dt_2,\ldots,dt_n)}.$$

We prove the following theorem.

Theorem 4.2 *The posterior distribution of the random probability measure μ, given a sample $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$ from F, is as follows:*

$$\mathcal{L}(\mu|(X_1,Y_1)=(x_1,y_1),(X_2,Y_2)=(x_2,y_2),\ldots,(X_n,Y_n)=(x_n,y_n))=$$

$$\int_{[0,1]^n}\tau_{c\mu_0+\sum_{i=1}^n\delta_{t_i}}\gamma_{[x,y]}(dt_1,dt_2,\ldots,dt_n),$$

$\tau_{c\mu_0+\sum_{i=1}^n\delta_{t_i}}$ *being the law of a Dirichlet process with base measure* $c\mu_0+\sum_{i=1}^n\delta_{t_i}$.

Proof

Consider, as above, a sample $T_1,\ldots,T_n$ from μ. Then, for Borel sets in $[0,1]$ $A_1,\ldots,A_n$, denoting by $P_{T_k}(\cdot|T_1=t_1,\ldots,T_{k-1}=t_{k-1})$ the conditional distribution of T_k given $T_1,\ldots,T_{k-1}$, for $k\in\{1,2,\ldots,n\}$, we have

$$P(T_1\in A_1,T_2\in A_2,\ldots,T_n\in A_n)=$$

$$=\int_{[0,1]^n}P_{T_n}(A_n|T_i=t_i,i=1,\ldots,n-1)P_{T_{n-1}}(dt_{n-1}|T_i=t_i,i=1,\ldots,n-2)\cdots P_{T_1}(dt_1)$$

$$= \lambda(A_1 \times A_2 \times \ldots \times A_n),$$

where the last equality follows from a well known property of Dirichlet processes. Since $(X_1, Y_1), \ldots, (X_n, Y_n)$ and μ are conditionally independent given $T_1, T_2, \ldots, T_n$, then, for a Borel set C in the class $\mathcal{P}$ of all probability measures on $[0, 1]$,

$$P(\mu \in C | (X_1, Y_1) = (x_1, y_1), \ldots, (X_n, Y_n) = (x_n, y_n))$$

$$= \int_{[0,1]^n} P(\mu \in C | T_1 = t_1, \ldots, T_n = t_n) P_{T_1,\ldots,T_n}(dt_1, \ldots, dt_n | (X_i, Y_i) = (x_i, y_i), i = 1 \ldots n)$$

$$= \int_{[0,1]^n} \tau_{c\mu_0 + \sum_{i=1}^n \delta_{t_i}}(C) P_{T_1,\ldots,T_n}(dt_1, \ldots, dt_n | (X_i, Y_i) = (x_i, y_i), i = 1 \ldots n)$$

where $P_{T_1,\ldots,T_n}(\cdot | (X_1, Y_1) = (x_1, y_1), (X_2, Y_2) = (x_2, y_2), \ldots, (X_n, Y_n) = (x_n, y_n))$ denotes the conditional law of $T_1, \ldots, T_n$ given the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. An application of Bayes theorem gives

$$P(T_1 \in A_1, T_2 \in A_2, \ldots, T_n \in A_n | (X_i, Y_i) = (x_i, y_i), i = 1 \ldots n) =$$

$$= \frac{\int_{\bigotimes_{i=1}^n A_i} \prod_{i=1}^n \left\{ \sum_{m=0}^{+\infty} \frac{\sqrt{(\alpha)_m}}{\sqrt{(\beta)_m}} t_i^m L_m^\alpha(x_i) L_m^\beta(y_i) \right\} \lambda(dt_1, dt_2, \ldots, dt_n)}{\int_{[0,1]^n} \prod_{i=1}^n \left\{ \sum_{m=0}^{+\infty} \frac{\sqrt{(\alpha)_m}}{\sqrt{(\beta)_m}} t_i^m L_m^\alpha(x_i) L_m^\beta(y_i) \right\} \lambda(dt_1, dt_2, \ldots, dt_n)}$$

$$= \gamma_{[x,y]}(A_1, A_2, \ldots, A_n)$$

that, together with (9), gives the thesis.

Remark 4.1 *Theorem* 4.2 *shows that the posterior distribution of μ, given a sample from F, is a mixture of Dirichlet processes, as defined and discussed by Antoniak* (1974).

For a bounded real function σ on $[0, 1]$, let $\varphi(\mu) = \int_{[0,1]} \sigma(t)\mu(dt)$ be a linear functional of μ.
The following proposition gives the posterior bayes estimate (with respect to quadratic loss) of $\varphi(\mu)$.

Theorem 4.3

$$E(\varphi(\mu) | (X_1, Y_1) = (x_1, y_1), \ldots, (X_n, Y_n) = (x_n, y_n)) =$$

$$\frac{c}{c+n}\varphi(\mu_0) + \frac{1}{c+n} \sum_{i=1}^n \int_{[0,1]^n} \sigma(t_i) \gamma_{[x,y]} dt_1, dt_2, \ldots, dt_n). \qquad (14)$$

Proof

Since the posterior law of μ is a mixture of Dirichlet processes, with mixing measure $\gamma_{[x,y]}$, then the posterior expectation of every linear functional of μ is equal to the same mixture of the posterior expectations of its components. By well known properties of the Dirichlet processes, the expected value of $\varphi(\mu) = \int_{[0,1]} \sigma(t)\mu(dt)$ with respect to $\tau_{c\mu_0+\sum_{i=1}^n \delta_{t_i}}$ is

$$\int_{[0,1]} \sigma(s) \left\{ \frac{c\mu_0 + \sum_{i=1}^n \delta_{t_i}}{c+n} \right\} (ds) = \frac{c}{c+n} \int_{[0,1]} \sigma(s)\mu_0(ds) + \frac{1}{c+n} \sum_{i=1}^n \sigma(t_i),$$

so that (14) is true.

Remark 4.2 *When $\sigma(t) = t^k$, $\varphi(\mu)$ is the (random) k-th moment of the Dirichlet process μ; that is, up to a constant, the k-th term ρ_k of the Lancaster sequence corresponding to the mixture.*

Remark 4.3 *Pearson mean square contingency*

$$\phi^2 = \sum_{n=0}^{+\infty} \rho_n^2 = \sum_{n=0}^{+\infty} \frac{(\alpha)_n}{(\beta)_n} \left\{ \int t^n \mu(dt) \right\}^2$$

can be estimated, when finite, by the truncated series of the estimates of the random moments of μ. Pommeret (2004) *considers the same problem in a classical approach.*

Remark 4.4 *For (u, v) in $(0, +\infty)^2$, when $\sigma(t) = f_t^*(u, v)$, $\varphi(\mu)$ is the density $f(u, v)$. Then, by (14), we have*

$$E(f(u,v)|(X_1,Y_1) = (x_1,y_1), \ldots, (X_n,Y_n) = (x_n,y_n)) = \tag{15}$$

$$= \frac{c}{c+n} \int_{[0,1]} f_t^*(u,v)\mu_0(dt) + \frac{1}{c+n} \sum_{i=1}^n \int_{[0,1]^n} f_{t_i}^*(u,v)\gamma_{[x,y]}(dt_1, \ldots, dt_n)$$

Of course, (15) gives the predictive density of (X_{n+1}, Y_{n+1}), given $(X_1, Y_1) = (x_1, y_1), (X_2, Y_2) = (x_2, y_2), \ldots, (X_n, Y_n) = (x_n, y_n)$.

The expressions in Theorems 4.2 and 4.3 are difficult to use. For this reason, we resort to a MCMC algorithm to approximate the posterior distribution of the process μ along with the distributions of its linear functionals. The approach suggested in [11] is followed.

Let μ_0 satisfy condition (8) and have density h_0 with respect to the Lebesgue measure and let c be a positive constant. Starting from [20], the Dirichlet process with base measure $c\mu_0$ can be approximated resorting to the random probability measure $P_N = \sum_{j=1}^N p_j \delta_{Z_j}$, where N is a positive integer, $(p_1, \cdots, p_N)$

has Dirichlet distribution with parameter $(c/N, \cdots, c/N)$, the Z_j's are independent random variables identically distributed according to μ_0 and $(p_1, \cdots, p_N)$ is independent on $(Z_1, \cdots, Z_N)$.
Then the posterior distribution of P_N can be approximated resorting to a Gibbs sampler with full conditional distributions:

$$\mathcal{L}((p_1, \ldots, p_N), (Z_1, \cdots, Z_N) | (K_1, \cdots, K_n), (X_1, Y_1), \cdots, (X_n, Y_n)), \tag{16}$$

$$\mathcal{L}((K_1, \ldots, K_n) | (p_1, \ldots, p_N), (Z_1, \ldots, Z_N), (X_1, Y_1), \ldots, (X_n, Y_n)). \tag{17}$$

$K_1, \ldots, K_n$ are latent classifications variables defined by $K_i = j$ if and only if $T_i = Z_j$, $T_1, T_2, \ldots, T_n$ being the random variables introduced in (11).
Due to assumptions, (16) reduces to the following product:

$$\mathcal{L}(p_1, \ldots, p_N | K_1, \ldots, K_n) \cdot \prod_{j=1}^{m} h_0(z_{K_j^*}) \cdot \prod_{i:K_i = K_j^*} f^*_{z_{K_i}}(x_i, y_i) \cdot \prod_{Z_j \in Z_K} h_0(z_j),$$

where $\mathcal{L}(p_1, \ldots, p_N | K_1, \ldots, K_n)$ is a Dirichlet distribution with parameter $(c/N + m_1, \ldots, c/N + m_N)$, m_i being the number of K_j's equal to i, for $i = 1, \ldots, n$ and $K_1^*, \ldots, K_m^*$ are the distinct values of $(K_1, \ldots, K_n)$.
By standard computations, for the second full conditional (17), we have

$$\mathcal{L}(K_1, \ldots, K_n | (p_1, \ldots, p_N), (Z_1, \ldots, Z_N), (X_1, Y_1), \ldots, (X_n, Y_n)) = (\sum_{j=1}^{N} p^*_{j,i} \delta_j)^n,$$

where $p^*_{j,i}$ is, up to a constant, equal to $f^*_{Z_j}(x_i, y_i)$.

5 Numerical results

The model described in the previous sections is applied to a simulated data set of 50 pairs drawn from a bivariate distribution F belonging to the Fréchet class with both marginals equal to the exponential distribution with mean 1. The Pearson correlation coefficient of the sample is 0.8877.
Using the modified Bessel function I_0 of order 0, the density (7) of the extreme point can be written as

$$f^*_t(x, y) = \frac{1}{1-t} e^{-\frac{x+y}{1-t}} I_0\left(2\sqrt{xy}\frac{\sqrt{t}}{1-t}\right).$$

We set, in the approximating random measure P_N, $N = 30$. Posterior estimates are obtained with different choices of c and μ_0.
After a burn-in of 10000 iterations, 10000 trajectories of the process μ are drawn from its posterior distribution.
Since the marginals are equal, the Pearson correlation coefficient ρ is the mean

of the random probability measure μ.
Tables 1-2 show the bayesian estimates (with respect to quadratic loss function) of ρ corresponding to different values of c and different choices of μ_0; a uniform distribution on $(0,1)$ ($U(0,1)$) and a Beta distribution with parameters $700, 300$ ($Beta(700,300)$) are considered.

Table 1: Bayesian estimates of the correlation coefficient, $\mu_0 = U(0,1)$

c	estimate
0.1	0.8626
1	0.8343
50	0.6119
1000	0.5631

Table 2: Bayesian estimates of the correlation coefficient, $\mu_0 = Beta(700,300)$

c	estimate
0.01	0.7053
1	0.7027
100	0.7002

Figure 1 shows the histogram of the sample drawn from the posterior distribution of ρ for c equal to 1 and $\mu_0 = U(0,1)$.

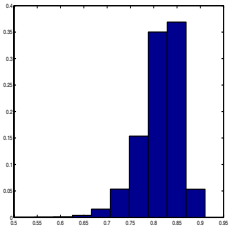

Figure 1: Histogram of ρ, $c = 1$, $\mu_0 = U(0,1)$

An estimate of the Pearson mean square contingency, ϕ^2, can be obtained by truncating $\sum_{n=0}^{+\infty} \rho_n^2$ at a suitable value and by plugging in the sum the estimates of the ρ_i's.
Table 3 displays the bayesian estimates of ϕ^2 obtained truncating the series at $n = 100$, for different values of c and $\mu_0 = U(0,1)$.

Table 3: Bayesian estimates of ϕ^2

c	estimates
0.1	3.0595
1	2.9250
50	1.0358
1000	0.8094

References

[1] Beneš, V. and Š. J., Extremal solutions in the marginal problem, in *Advances in probability distributions with given marginals (Rome, 1990)*, Kluwer Acad. Publ. (1991), **67** 189–206.

[2] Cifarelli, D.M. and Regazzini, E., Distribution functions of means of a Dirichlet process, *The Annals of Statistics*, **18** (1990), 429–442.

[3] Dalal, S. R. and Hall, Jr. G. J., On approximating parametric Bayes models by nonparametric Bayes models, *The Annals of Statistics*, **8** (1980), 664–672.

[4] Deming, W. E. and Stephan, F. F., On a least squares adjustment of a sampled frequency table when the expected marginal totals are known, *The Annals of Mathematical Statistics*, **11** (1940), 427–444.

[5] Ferguson, T.S, Bayesian density extimation by mixtures of normal distributions in *Recent Adavances in Statistics,* Academic Press, New York, 1983.

[6] Ghosh, J. K. and Ramamoorthi, R. V., *Bayesian nonparametrics,* Springer-Verlag, New York, 2003.

[7] Griffiths, R. C, The canonical correlation coefficients of bivariate gamma distributions, *The Annals of Mathematical Statistics*, **40** (1969), 1401–1408.

[8] Hoff, P. D., Nonparametric estimation of convex models via mixtures, *The Annals of Statistics*, **31** (2003), 174–200.

[9] Hutchinson, T. P. and Lai, C. D., *Continuous bivariate distributions, emphasising applications,* Rumsby Scientific Publishing, Adelaide, 1990.

[10] Ireland, C. T. and Kullback, S., Contingency tables with given marginals, *Biometrika*, **55** (1968), 179–188.

[11] Ishwaran, H. and Zarepour, M., Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models, *Biometrika*, **87** (2000), 371–390.

[12] Kotz,S., Multivariate distributions at a cross road in *A Modern Course on Distributions in Scientific Work,* Patil, G.P. and Kotz, S. and Ord, J.K., eds., , volume I- Models and Structures, Reidel, Dordrecht, 1975.

[13] Koudou, A. E., Probabilités de Lancaster, *Expositiones Mathematicae. International Journal*, **14** (1996), 247–275.

[14] Kullback, S., Probability densities with given marginals, *The Annals of Mathematical Statistics*, **39** (1968), 1236–1243.

[15] Lancaster, H. O., The structure of bivariate distributions, *The Annals of Mathematical Statistics*, **29** (1958), 719–736.

[16] Letac, G., The Lancaster's probabilities on $\mathbf{R}^2$ and their extreme points in *Distributions with given marginals and moment problems (Prague, 1996)*, Kluwer Acad. Publ., Dordrecht (1997), 179–190.

[17] Lo, A. Y., On a class of Bayesian nonparametric estimates. I. Density estimates, *The Annals of Statistics*, **12** (1984), 351–357.

[18] Melilli, E. and Petris, G., Bayesian Inference for Contingency Tables with Given Marginals, *Journal of the Italian Statistical Society*, **2** (1995), 215-233.

[19] Sarmanov, O.V., Generalized normal correlation and two-dimensional Fréchet classes, *Soviet Mathematics*, **7** (1966), 596–599.

[20] Sethuraman, J., A constructive definition of Dirichlet priors, *Statistica Sinica*, **4** (1994), 639-650.

Received: November 30, 2007