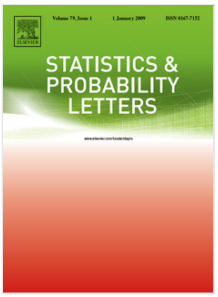
Bayesian nonparametric binary regression via random tessellations

Lorenzo Trippa, Pietro Muliere

Accepted Manuscript



Bayesian nonparametric binary regression via random tessellations

Lorenzo Trippa, Pietro Muliere

PII: S0167-7152(09)00289-2
DOI: 10.1016/j.spl.2009.07.026
Reference: STAPRO 5487

To appear in: *Statistics and Probability Letters*

Received date: 5 October 2008
Revised date: 4 June 2009
Accepted date: 28 July 2009

Please cite this article as: Trippa, L., Muliere, P., Bayesian nonparametric binary regression via random tessellations. *Statistics and Probability Letters* (2009), doi:10.1016/j.spl.2009.07.026

Bayesian Nonparametric Binary Regression via Random Tessellations.

Lorenzo Trippa[1,2,*], Pietro Muliere[1]

[1] *Department of Decision Sciences, Università Bocconi*

[2] *Department of Biostatistics, M.D.Anderson Cancer Center*

Abstract

A Bayesian nonparametric model for binary random variables is introduced. The characterization of the probability model is based on the Dirichlet process and on the Poisson hyperplane tessellation model. These two stochastic models are combined in order to adapt, under the hypothesis of partial exchangeability, the reinforcement mechanism of the Pólya urn scheme. A Gibbs sampling algorithm for implementing predictive inference is illustrated and an application of the inferential procedure is discussed.

Key words: Binary regression, Random tessellations, Dirichlet process

1 Introduction

Consider an heterogeneous population of subjects with covariates. A dichotomous random variable (r.v.) Y_i is associated with each subject, with response probability depending on the covariates:

$$P(Y_i = 1|X_i) = f(X_i),$$

i indexes the individual and X_i represents its profile. We propose a Bayesian nonparametric model for the regression function f.

* E-mail: lorenzo.trippa@unibocconi.it

The response probability function f is modeled by means of a probability measure $\mathcal{D}$ on a space of mosaics having cells associated with the values 0 or 1. We will refer to a mosaic with values assigned to each cell as a colored tessellation. Throughout the article we use colored tessellations of the covariates space. The response probability $f(X)$, for a generic point X, will be defined as the probability of sampling from $\mathcal{D}$ a tessellation having the cell containing X associated with the value 1.

The most widely used parametric models for binary regression are defined by means of a parametric function l_θ, which maps the covariates space on the real line, and a cumulative density function (cdf) H; such models assume that $P(Y_i = 1 \mid X_i, \theta) = H[l_\theta(X_i)]$. In the probit model, for example, l_θ is linear and H is the standard Gaussian distribution function. This framework is exploited in Newton et al. (1996) to define a semiparametric Bayesian binary regression model, therein l_θ is linear and a Normal prior distribution on the coefficients is combined with a Dirichlet-Ferguson prior (Ferguson , 1973) for the cdf H. Wood and Kohn (1998) proposed an alternative Bayesian semiparametric model, the cdf H is fixed and a flexible prior for the function l, which is assumed to be a sum of functions of single covariates, is defined.

More recently Choudhuria et al. (2007) studied a nonparametric prior distribution on the response probability functions which is not entirely concentrated on the monotone functions nor on the additive ones; the typical assumptions of the semiparametric models are removed. Therein the link function H is the standard Normal cdf and l is modeled as a Gaussian process.

In this article a Multivariate-beta prior, whose characterization is based on the distribution of a random tessellation, is proposed to model binary response variables. The prior is specified through an instrumental Dirichlet process on a space of colored tessellations. The process is centered on the probability law of a random tessellation with colored cells. The regression model is structured as if the response variables $\{Y_i\}_{i\geq 1}$ were functions of a latent sequence of random tessellations drawn from an unknown distribution $\mathcal{D}$. If the cell of the i-th tessellation where is located X_i is associated with the value 1 then Y_i is equal to 1 and 0 otherwise. We use the Dirichlet-Ferguson prior for the unknown distribution $\mathcal{D}$. The regression model is characterized by an easily interpretable dependency structure. If the random probability measure $\mathcal{D}$ is expected to

concentrate on few tessellations, intuitively, Y_1 is strongly predictive of the response variable Y_2, especially if X_1 is near to X_2.

The article is organized in 4 sections. In Section 2 the the Poisson hyperplane tessellation is briefly described. We consider this random tessellation model to specify the parameter of the random probability measure $\mathcal{D}$ on the tessellations space. The Poisson hyperplane tessellation model has been studied by several authors; important properties are discussed in Miles (1964), Miles (1972), Møller (1989) and Hug et al. (2004). More generally, random tessellations have received much attention during the last decades and find applications in many fields as geostatistics and stereology. For a comprehensive overview on random tessellation models we refer to Okabe et al. (1992).

In Section 3 we introduce the random probability measure $\mathcal{D}$ in order to characterize a class of dependent random distributions on the real line $\{\mathcal{D}_X\}_{X\in\mathbb{R}^d}$ indexed by covariates. The adopted approach allows us to center the random probability measures $\{\mathcal{D}_X\}_{X\in\mathbb{R}^d}$ on any arbitrarily chosen distribution F. In particular if F concentrates on 0 and 1, i.e. $F(\{0,1\}) = 1$, it allows us to specify the random response probability function $f(X) \equiv \mathcal{D}_X(\{1\})$. We conclude Section 3 describing a simulation algorithm for posterior inference on the response probability function f and illustrating an application of the proposed binary regression model. Some final remarks are given in the last section.

2 Poisson hyperplane tessellations

Random tessellation models have been extensively studied in stochastic geometry and are applied in a wide range of fields ranging from geography to molecular biology. A tessellation t of the Euclidean space $\mathbb{R}^d$ can be described as a mosaic of d-dimensional polytopes which entirely cover $\mathbb{R}^d$. In this section we give an intuitive description of the Poisson hyperplane tessellation model and provide the definition of the stochastic model.

A hyperplane tessellation of $\mathbb{R}^d$ or of a subset, say $[0,1]^d$, is defined by a set of hyperplanes. Figure 1 displays an example; the class of represented polytopes which are not crossed by any of the plotted lines is an hyperplane tessellation of $[0,1]^2$. We can similarly define hyperplane tessellations of an ar-

bitrarily chosen Euclidean set of interest. A tessellation of $[0,1]^d$, for example, can be specified simply substituting the lines in Figure 1 by a set Q of hyperplanes. If Q is a possibly empty random set of hyperplanes crossing $[0,1]^d$ we obtain a random hyperplane tessellation of $[0,1]^d$. The outlined construction can be easily extended for specifying a random tessellation of $\mathbb{R}^d$. Throughout the article Q is constituted by the atoms of a Poisson process on a space of hyperplanes and defines a random tessellation.

A random tessellation of $\mathbb{R}^d$ is a stochastic counting measure on the measurable space $(\mathcal{P}, \mathfrak{B})$, where $\mathcal{P}$ denotes the class of polytopes in $\mathbb{R}^d$ and $\mathfrak{B}$ is the Borel σ-field with respect to the Hausdorff metric. Consider the class of counting measures T such that for every $t \in T$ the following hold:

(i) for every polytope $p \in \mathcal{P}$ $\;t(p) \in \{0,1\}$,
(ii) $\bigcup_{t(p)=1} p = \mathbb{R}^d$,
(iii) for every pair of polytopes (p_1, p_2) such that $t(p_1) = t(p_2) = 1$ the interior of p_1 and p_2 are disjoint and
(iv) given a bounded set $c \subset \mathbb{R}^d$ the class $\{p \in \mathcal{P} : t(p) = 1,\, p \cap c \neq \emptyset\}$ has finite cardinality.

Each counting measure $t \in T$ is a tessellation of $\mathbb{R}^d$. Let $\mathcal{T}$ denotes the smallest σ-field that guarantees, for every set $B \in \mathfrak{B}$, the measurability of the function $t \to t(B)$. The measurable space $(T, \mathcal{T})$ endowed with a probability measure P defines a random tessellation.

As we previously mentioned, a Poisson hyperplane tessellation of $\mathbb{R}^d$ is characterized by a Poisson process Q on a space of hyperplanes. In what follows $S^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the unit sphere, given $(s,r) \in S^{d-1} \times [0,\infty)$, $H_{s,r}$ indicates the hyperplane $\{x \in \mathbb{R}^d : \langle x, s\rangle = r\}$, and γ denotes the product measure of the uniform distribution on S^{d-1} and the Lebesgue measures on the real line. Q is an homogeneous Poisson process on the space of hyperplanes $\mathbb{H}^d \equiv \{H_{s,r};\, (s,r) \in S^{d-1} \times [0,\infty)\}$ with intensity λ if for every measurable set $A \subset \left(S^{d-1} \times [0,\infty)\right)$, $Q(H_{s,r} \in \mathbb{H}^d : (s,r) \in A)$ is a Poisson r.v. with parameter $\lambda\gamma(A)$, and for every class $(A_1, \ldots, A_J)$ of disjoint sets the variables $\{Q(H_{s,r} \in \mathbb{H}^d : (s,r) \in A_j)\}_{j=1}^J$ are independently distributed.

The atoms of the Poisson process Q specify a Poisson hyperplane tessellation t of $\mathbb{R}^d$. The polytopes that constitute the mosaic, i.e. $\{p \in \mathcal{P} : t(p) = 1\}$,

can be represented as nonempty intersections of closed half-spaces. The single polytope p is a cell of the random tessellation, i.e. $t(p) = 1$, if and only if there exist a set of hyperplanes $\{H_{s_j,r_j}\}_{j=1}^m$ and a binary vector $(I_1, \ldots, I_m) \in \{0,1\}^m$ such that $Q(H_{s_j,r_j}) = 1$ for every $j \in \{1, \ldots, m\}$,

$$p = \left\{ \bigcap_{I_j=0} (x \in \mathbb{R}^d : \langle x, s_j \rangle \leq r_j) \bigcap_{I_j=1} (x \in \mathbb{R}^d : \langle x, s_j \rangle \geq r_j) \right\} \quad \text{and}$$

$Q(H \in \mathbb{H}^d : H \cap int(p) \neq \emptyset) = 0$, where $int(p)$ is the interior of p.

[$Figure 1$]

The described random tessellation model has received much attention in literature. For an overview of the properties of this stochastic model we refer to Stoyan et al. (1995). In the reminder of the article it will be relevant that the Poisson process Q and the introduced random tessellation are isotropic and stationary (cf. Miles (1964)): the laws of both processes are invariant with respect to translations and rotations. We will also use the following proposition.

Proposition 1 *Given a Poisson hyperplane tessellation of* $\mathbb{R}^d$ *with intensity* λ*, the probability that a segment* $\overline{ab}$ *is entirely contained in a single cell is*

$$P\left(Q(H \in \mathbb{H}^d : H \cap \overline{ab} \neq \emptyset) = 0\right) = \exp\left(-\lambda \frac{\|\overline{ab}\|}{2} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right). \qquad (1)$$

Proof. Recall that the tessellation process is isotropic. We can therefore assume, without loss of generality, that $a = (0, \ldots, 0)$ and $b = (b_1, 0, \ldots, 0)$, with $b_1 > 0$. Let $(u_1, \ldots u_d)$ be a random vector uniformly distributed on the unit sphere. It is known (see for example Eaton (1981)) that $z \equiv u_1^2$ has a beta distribution with parameters $(\frac{1}{2}, \frac{d-1}{2})$. It follows that

$$\log\Big(\, P\left(Q(H \in \mathbb{H}^d : H \cap \overline{ab} \neq \emptyset) = 0\right)\Big) =$$

$$= -\lambda\gamma\Big(\, (s,r) \in S^{d-1} \times [0,\infty) : H_{s,r} \cap \overline{ab} \neq \emptyset\Big) =$$

$$= -\lambda\gamma\Big(\, ([s_1, \ldots, s_d], r) \in S^{d-1} \times [0,\infty) : 0 \leq r \leq b_1 s_1\Big) =$$

$$= -\frac{\lambda}{2}\gamma\Big(\, ([s_1, \ldots, s_d], r) \in S^{d-1} \times [0,\infty) : 0 \leq r \leq b_1\sqrt{s_1^2}\Big) =$$

$$= -\frac{\lambda}{2}\int b_1 z^{\frac{1}{2}} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d-1}{2})} z^{\frac{1}{2}-1}(1-z)^{\frac{d-1}{2}-1} dz = -\lambda \frac{b_1}{2} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}$$

■

3 Partial exchangeability via random tessellations

We propose a Bayesian nonparametric model for partially exchangeable random variables $\{Y_i\}_{i=1}^n$ with covariates $\{X_i \in \mathbb{R}^d\}_{i=1}^n$. We characterize the joint distribution of $\{Y_i\}_{i=1}^n$ exploiting an exchangeable sequence of colored tessellations $\{t_i\}_{i\geq 1}$. The intuitive idea is that Y_i is the color of the cell containing X_i of the i-th tessellation. The law of the random sequence $\{t_i\}_{i\geq 1}$, which is specified introducing a random probability measure $\mathcal{D}$ on a space of colored tessellations, characterizes the response variables $\{Y_i\}_{i=1}^n$.

Let $(T \otimes \mathbb{R}^{\mathbb{N}}, \mathcal{T} \otimes \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ be the measurable space of colored tessellations; each tessellation t is constituted by the cells $\{p_1, p_2, \ldots\}$ ordered with respect to an arbitrary criterion and the i-th real coordinate of the product space indicates the color of the i-th cell. In what follows, p_X^t denotes the cell with the minimum index containing X and $g(X;\ t)$ is the cell color. The selection of the minimum is required for covariate values that fall on a boundary and are contained in more than one cell. $\mathcal{D}$ is a Dirichlet process on the measurable space $(T \otimes \mathbb{R}^{\mathbb{N}}, \mathcal{T} \otimes \mathcal{B}(\mathbb{R}^{\mathbb{N}}))$ with parameter $M[P_\lambda \otimes F^{\mathbb{N}}]$, where M is a positive constant, P_λ denotes the distribution of a Poisson hyperplane tessellation and F is a distribution on the possible cells colors. The cells colors are real values and F is a probability measure on the real line. The Bayesian model for the partially exchangeable variables $\{Y_i\}_{i=1}^n$ is:

$$Y_i \mid t_i = g(X_i;\ t_i) \quad i = 1, \ldots, n\ , \tag{2}$$

$$\{t_i\}_{i\geq 1} \mid \mathcal{D} \overset{i.i.d.}{\sim} \mathcal{D} \quad \text{and} \quad \mathcal{D} \sim Dirichlet(M[P_\lambda \otimes F^{\mathbb{N}}])\ .$$

The elements of the partially exchangeable sequence $\{Y_i\}_{i\geq 1}$ are, conditionally on $\mathcal{D}$, independently distributed with

$$P(Y_i \in B|\mathcal{D}) = \mathcal{D}(t \in T \otimes \mathbb{R}^{\mathbb{N}} : g(X_i;\ t) \in B) \qquad \forall B \in \mathcal{B}(\mathbb{R})\,. \tag{3}$$

We will write $\mathcal{D}_X$ to indicate the random distribution

$$\{\mathcal{D}_X(B) \equiv \mathcal{D}(t \in T \otimes \mathbb{R}^{\mathbb{N}} : g(X;\ t) \in B)\}_{B\in\mathcal{B}(\mathbb{R})}\ .$$

The definition of the model implies that the random probability measure $\mathcal{D}_X$ associated with a single covariate point X is Dirichlet distributed with concen-

tration parameter M and centering distribution F. This fact follows directly from the definition of the Dirichlet process (Ferguson, 1973).

The equality (1) allows us to evaluate the degree of dependency between the random probability measures $\mathcal{D}_{X_1}$ and $\mathcal{D}_{X_2}$ for any pair of covariates (X_1, X_2). The following proposition illustrates that, for every measurable set B, the correlation coefficient $Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B))$ has a simple closed form.

Proposition 2 *For every measurable set $B \in \mathcal{B}(\mathbb{R})$, such that $0 < F(B) < 1$, and every pair of covariates $(X_1, X_2) \in \mathbb{R}^d \times \mathbb{R}^d$*

$$Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B)) = \exp\left(-\lambda \|X_1 - X_2\| \frac{\Gamma(\frac{d}{2})}{2\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right). \qquad (4)$$

Proof. Expressions (3) and (1) respectively imply the equalities

$P(Y_1 \in B, Y_2 \in B | \mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B)) = \mathcal{D}_{X_1}(B)\mathcal{D}_{X_2}(B)$ and

$P(Y_1 \in B, Y_2 \in B \mid t_1 = t_2) = F(B)^2 + (F(B) - F(B)^2) \exp\left(-\lambda \frac{\|\overline{ab}\|}{2} \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right)$.

It follows that

$E(\mathcal{D}_{X_1}(B)\mathcal{D}_{X_2}(B)) = P(Y_1 \in B, Y_2 \in B) =$

$= P(Y_1 \in B, Y_2 \in B \mid t_1 \neq t_2)P(t_1 \neq t_2) + P(Y_1 \in B, Y_2 \in B \mid t_1 = t_2)P(t_1 = t_2) =$

$= F^2(B) + (F(B) - F^2(B)) \frac{1}{M+1} \exp\left(-\lambda \|X_1 - X_2\| \frac{\Gamma(\frac{d}{2})}{2\Gamma(\frac{d+1}{2})\Gamma(\frac{1}{2})}\right)$.

The fact that $\mathcal{D}_{X_1}(B)$ and $\mathcal{D}_{X_2}(B)$ are identically beta distributed with mean $F(B)$ and variance $\frac{F(B) - F^2(B)}{M+1}$ completes the proof. ■

If the variables $\{Y_i\}_{i=1}^n$ are dichotomous, i.e. $F(\{0, 1\}) = 1$, the parameters of the defined model have a clear interpretation. The response probability $f(X) \equiv \mathcal{D}_X(\{1\})$ is a priori beta distributed with mean $F(\{1\})$ and variance $\frac{F(\{1\}) - F^2(\{1\})}{M+1}$, while the correlation between $f(X_1)$ and $f(X_2)$ depends on the intensity parameter λ and on the distance between X_1 and X_2.

Remark 1. The correlation function (4) depends exclusively on the random tessellation distribution P_λ that parameterizes the random probability measure $\mathcal{D}$. It can be easily verified that this peculiarity is preserved if the Poisson hyperplane tessellation model is substituted, in the construction of the prior, with alternative random tessellation models. Consider, for example, the dead leaves tessellation model (Matheron (1968)), with spherical leaves having random radius R with $P(R > r) = \exp(-\frac{r^2}{2/a^2})$. An elementary application of the

results in Bordenave et al. (2006), allows us to obtain the correlation function of this slightly modified version of the proposed model: for every $B \in \mathcal{B}(\mathbb{R})$

$$Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B)) = \frac{\Phi(-a\|X_1 - X_2\|)}{\Phi(a\|X_1 - X_2\|)}, \tag{5}$$

where Φ denotes the standard Gaussian cumulative distribution function.

Remark 2. An alternative characterization of dependent Dirichlet processes, whose dependency relationships allow a representation similar to expression (4), is discussed in Walker and Muliere (2003) and Muliere et. al. (2005). A relevant difference between such a characterization and the introduced prior consists in the fact that, with the proposed model, the correlation coefficients $Corr(\mathcal{D}_{X_1}(B), \mathcal{D}_{X_2}(B))$ gradually decrease as the distances $\|X_1 - X_2\|$ increase.

3.1 Predictive inference.

In the following paragraphs we propose a Gibbs sampling algorithm which, conditionally on $\{Y_1, \ldots, Y_n\}$, allows us to perform predictive inference on a future response variable Y_{n+1}. In the reminder of the article the centering distribution F is assumed to be discrete. In the next subsections we will use the algorithm to estimate unknown response probability functions f. The algorithm approximately samples from the conditional law of the latent tessellations $\{t_i\}_{i=1}^n$, given $\{Y_1, \ldots, Y_n\}$. We emphasize that sampling $\{t_1, \ldots, t_n\}$ conditionally on the data allows us to generate Y_{n+1} from the predictive distribution and to make inference on the unknown distributions $\{\mathcal{D}_X\}_{X \in \mathbb{R}^d}$. Both these objectives can be achieved exploiting the Monte Carlo method and the conjugacy property of the Dirichlet process. The Gibbs sampling strategy consists of iteratively sampling the latent tessellations t_i from the respective full conditional distributions, i.e. t_i is sampled conditionally on $\{Y_1, \ldots, Y_n\}$ and $\{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$. We show that the latent tessellation t_i can be easily sampled from the full conditional distribution. We refer to Robert and Casella (2004) for a detailed explanation of the Gibbs sampling method.

The sampling algorithm is based on the Blackwell MacQueen representation of an exchangeable sequence of variables with Dirichlet random distribution (Blackwell and MacQueen, 1973). In our case the elements of the sequence are the latent tessellations. The algorithm is initialized by a truncated se-

quence of colored tessellations, one for each observation. In each iteration the tessellation corresponding to a single observation is conditionally generated. The structure of the algorithm has similarities with some of the approaches proposed in literature for fitting Dirichlet mixture models (Lo, 1984); see for example Escobar and West (1995) and MacEachern and Muller (1998).

The Blackwell MacQueen representation allows us to verify that, if $\mathcal{D}$ is a Dirichlet process with parameter $M[P_\lambda \otimes F^{\mathbb{N}}]$ and $\{t_i\}_{i\geq 1} \mid \mathcal{D} \overset{i.i.d.}{\sim} \mathcal{D}$, then for every measurable set $\Delta \in \{\mathcal{T} \otimes \mathcal{B}(\mathbb{R}^{\mathbb{N}})\}$

$$P(t_i \in \Delta \mid t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) = M \frac{P_\lambda \otimes F^{\mathbb{N}}(\Delta)}{M+n-1} + \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{I(t_j \in \Delta)}{M+n-1}\,. \quad (6)$$

Note that it is simple to generate t_i conditionally on $\{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$. The law of t_i conditionally on $\{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$ and $\{Y_i\}_{i=1}^n$ is identical to the conditional distribution (6) restricted to $\{t \in T \otimes \mathbb{R}^{\mathbb{N}} : g(X_i; t) = Y_i\}$, indeed for every measurable Δ

$$\begin{aligned} &P(t_i \in \Delta \mid t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n, \{Y_j\}_{j=1}^n) \propto \\ &P(t_i \in \Delta \mid t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n) P(Y_i \mid t_1, \ldots, t_{i-1}, t_i \in \Delta, t_{i+1}, \ldots, t_n) = \\ &\qquad = P(t_i \in \{t \in \Delta : g(X_i; t) = Y_i\} \mid t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n)\,. \quad (7) \end{aligned}$$

Expressions (6) and (7) allows us to sample t_i from the full conditional:

$$P(t_i \in \Delta \mid t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n, \{Y_j\}_{j=1}^n) = \quad (8)$$

$$= \frac{M P_\lambda \otimes F^{\mathbb{N}}(\{t \in \Delta : g(X_i; t) = Y_i\}) + \sum_{j\leq n;\, j\neq i} I(g(X_i; t_j) = Y_i) I(t_j \in \Delta)}{M F(Y_i) + \sum_{j\leq n;\, j\neq i} I(g(X_i; t_j) = Y_i)}\,.$$

We exploit the fact that it is simple to compute the conditional probabilities of the events $\{t_i = t_j\}$ for every $j \in \{1, \ldots, i-1, i+1, \ldots, n\}$. Moreover, when it is necessary to sample the latent tessellation t_i conditionally on the complementary event $t_i \notin \{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$, the data and $\{t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_n\}$, expression (8) indicates that it suffices (i) to generate a P_λ distributed tessellation t, (ii) to assign color Y_i to the cell $p^t_{X_i}$ and (iii) to randomly color the other cells accordingly with F. Note that at each iteration it is not necessary to sample the entire tessellation t_i; it suffices to

partially generate t_i in such a way to know how a subset of interest $A \subset \mathbb{R}^d$, such that $A \supset \{X_1, \ldots, X_n\}$, is partitioned and colored.

3.2 Simulation example.

The proposed algorithm has been applied to the simulated data set represented in *Figure 2.a*. $n = 100$ points $\{X_1, \ldots, X_{100}\}$ have been randomly selected in $[0,1]^2$ and for each point a Bernoulli r.v. Y_i has been generated accordingly with the response probabilities illustrated in *Figure 2.b*. The surface in the graph is the rescaled density of a mixture of two truncated bivariate Gaussian distributions.

The prior distribution has been specified choosing the parameters $M = 2$, $F(\{0\}) = F(\{1\}) = 1/2$ and $\lambda = 10$. The adopted parameterization implies that the response probabilities are a priori marginally uniformly distributed and their distribution is characterized by the correlation function

$$Corr(f(X_1), f(X_2)) = \exp(-\frac{10}{\pi}\|X_1 - X_2\|)\,.$$

Figure 2.c represents an approximation of the posterior mean

$$\hat{f}(X) = E(\mathcal{D}_X(\{1\}) \mid data)$$

of the response probability function, which has been obtained computing, by means of the proposed Gibbs sampling algorithm, the values of $\hat{f}$ on a grid of 2500 points on the unit square. We note that the Bayesian estimate $\hat{f}$ captures the bimodal shape of the response probability function adopted to generate the data.

[*Figure* 2]

Figure 2.d illustrates a unidimensional section of $\hat{f}$ and the relative 90% credible intervals. The credible intervals have been computed by approximating the quantile functions of the random probabilities $\mathcal{D}_X(\{1\})$ via Monte Carlo method. In order to compute the quantile functions, we have iteratively sampled the latent tessellations $\{t_1, \ldots t_{100}\}$ from their conditional distribution and have exploited the conjugacy property of the Dirichlet prior. This property implies that, for every $X \in [0,1]^2$, $D_X(\{1\})$ conditionally on $\{t_1, \ldots t_{100}\}$ and $\{Y_1, \ldots, Y_{100}\}$ has a $beta(1 + \sum_{i=1}^{100} g(X; t_i)\,;\, 101 - \sum_{i=1}^{100} g(X; t_i))$ distribution.

3.3 Example: Spatial variation in risk of disease

We apply the proposed inferential procedure to an epidemiological data set. The data set reports the postcodes of patients in Northeast England affected with primary biliary cirrhosis (PBC) and of a control group of residents of the region. For a detailed description of the data set we refer to Prince et al. (2001). One of the main findings of the cited study was that the risk of disease varies considerably across the region. Identifying the high-risk areas can substantially contribute to ascertain the environmental risk factors. Controls, as described in Prince et al. (2001), are randomly selected from the region population.

As discussed in Kelsall and Diggle (1998), the problem of identifying the variations of the risk of disease across a specific area can be formalized as a binary regression problem. In principle, it would be desirable to compare the spatial distribution of patients with a specific disease diagnosed in a fixed time interval with the population density across the region. In practice, a complete record of the diagnosed patients and reliable data about the population density, in many cases, are not available to the investigator.

Kelsall and Diggle (1998) assume that unknown proportions of the diseased and of the non-diseased populations (q_1 and q_2) are observed and model the observed cases and controls as two independent non homogeneous Poisson processes with intensities $\eta_1(X) = q_1\nu_1(X)$ and $\eta_2(X) = q_2\nu_2(X)$. They observe that both

$$g(X) \equiv \frac{\nu_1(X)}{\nu_1(X) + \nu_2(X)} \quad \text{and} \quad f(X) \equiv \frac{\eta_1(X)}{\eta_1(X) + \eta_2(X)}$$

have clear interpretations. Note that $g(X)$ can be interpreted as the risk of disease associated with the spatial coordinates X and that f allows us to evaluate if the risk of disease varies considerably across the area of interest. Observe that for every pair of locations (X_1, X_2)

$$\log\left(\frac{g(X_1)}{1-g(X_1)}\right) - \log\left(\frac{g(X_2)}{1-g(X_2)}\right) = \log\left(\frac{f(X_1)}{1-f(X_1)}\right) - \log\left(\frac{f(X_2)}{1-f(X_2)}\right).$$

We are interested in estimating the binary regression function f. A flat regression function would represent a constant risk of disease while possible variations could be determined by environmental factors which vary across

the region.

Figure 3.a illustrates the polygonal approximation of the study area and the residence locations of the 761 patients affected by PBC and 3044 controls in the data set. Figure 3.b gives a synthetic representation of the results obtained from applying the proposed model. The graph illustrates the estimate $\hat{f}$ of the binary regression function. The shades of gray used in the picture show that the estimates $\hat{f}(X)$ vary approximately between 0.1 (lightest tone) and 0.3 (darkest tone). The map suggests a relevant difference in the incidence of PBC between the urban areas of Newcastle and Gateshead, which are associated with the highest values of $\hat{f}$ and are colored in black in Figure 3.b, and the surrounding areas.

[*Figure* 3]

In order to evaluate the strength of the evidence that there is a higher risk of disease in the identified urban centers than in the peripheral areas, we have computed the posterior probabilities that the risk variations suggested by Figure 3.b correspond to analogous variations of the unknown regression function f. These posterior probabilities allow us to evaluate the evidence in favor of the hypothesis that, if the environmental risk factors will remain unchanged, the risk of disease will continue to be higher in the urban centers than in the peripheral areas.

The introduced Gibbs sampler allows us, for any couple (X_1, X_2), to approximate the posterior distribution of the difference $f(X_1) - f(X_2)$. In order to evaluate the evidence of relevant variations of the risk across the region we have approximated the posterior distributions of the random quantities $f(X_1) - f(X_2)$, where X_1 is fixed, maximizes $\hat{f}$ (i.e. $\hat{f}(X_1) = \max \hat{f}(X)$) and is located in the urban center of Newcastle while X_2 varies on a grid of points. This simple procedure identifies the points X_2 for which the data gives strong evidence that the difference $f(X_1) - f(X_2)$ is strictly positive. The criteria that we adopted for distinguishing such points is $P(f(X_1) - f(X_2) > 0 \,|\, data) > 0.95$. The dashed line in Figure 3.b approximately identifies the points that satisfies the inequality. The entire study region with the exception of the identified urban centers and a restricted peripheral area satisfies the adopted criteria.

4 Final remarks

In this article a class of dependent beta r.v.'s $\{f(X)\}_{X\in\mathbb{R}^d}$ indexed by covariates is characterized introducing a Diririchlet random distribution on a space of colored tessellations. More generally the adopted approach allows us to characterize a class of dependent Dirichlet processes $\{D_X\}_{X\in\mathbb{R}^d}$. In recent years several contributions on dependent random probability measures have appeared in literature, see for example De Iorio et. al. (2004), Griffin and Steel (2006) and Dunson and Park (2007). We have used the dependent processes $\{D_X\}_{X\in\mathbb{R}^d}$ for performing nonparametric binary regression. It remains unexplored if the proposed characterization of the random distributions can be usefully adopted for alternative purposes. The random probability measures $\{D_X\}_{X\in\mathbb{R}^d}$ could be used, for example, for defining dependent Dirichlet mixtures of parametric densities.

Acknoledgements. The authors would like to thank two referees for their helpful comments on an earlier draft of this paper.

References

Blackwell, D., MacQueen, J. B. (1973) Ferguson distributions via Pólya-urn schemes. *The Annals of Statistics*, 1, 353-355.

Bordenave, C., Gousseau, Y., Roueff, F. (2006) The dead leaves model: a general tessellation modeling occlusion. *Advances in Applied Probability*, 38, 31-46.

Choudhuria, N., Ghosal, S., Roy, A.(2007) Nonparametric binary regression using a Gaussian process prior. *Statistical Methodology* , 4, 227-243

De Iorio, M., Muller, P., Rosner, G.L., MacEachern, S.N. (2004) An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, 99, 205-215.

Dunson, D.B., Park, J.H. (2007) Kernel stick breaking processes. *Biometrika*, 95, 307-323.

Eaton, M.L. (1981) On the projections of isotropic distributions. *The Annals of Statistics*, 9, 391-400.

Escobar, M.D., West, M. (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association,* 90, 577-588

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209-230.

Griffin, T.S., Steel, M.F.J. (2006) Order-based dependent Dirichlet processes. *Journal of the American Statistical Association*, 101, 179-194.

Lo, A.Y. (1984) On a class of Bayesian nonparametric estimates. *The Annals of Statistics,* 12, 351-357

MacEachern, S., Muller, P. (1998) Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics ,* 2, 223-238

Matheron, G. (1968). Modele sequentiel de partition aleatoire. *Tech. Rep., Centre de Morphologie Mathematique,* Fontainebleau

Miles, R. E. (1964) Random polygons determined by random lines in a plane. *Proc. Nat. Acad. Sci,* 52, 901-907

Miles, R. E. (1972) The random division of space. *Advances in Applied Probability*, Vol. 4, Supplement, 243-266.

Møller, J. (1989) Random tessellations in $\mathbb{R}^d$. *Advances in Applied Probability,* 21 , 3-73.

Muliere, P., Secchi, P., Walker, S. (2005) Partially exchangeable processes indexed by the vertices of a k-tree constructed via reinforcement. *Stochastic processes and their applications*, 115, 661-677

Newton, M.A., Czado, C. and Chappell, R. (1996) Bayesian inference for semiparametric binary regression. *Journal of the American Statistical Association*, 91, 142-153

Kelsall, J.E., Diggle, P.J. (1998) Spatial variation in risk of disease : a nonparametric binary regression approach. *Applied statistics*, 47, 559-573

Okabe, A., Boots, B., Sugihara, K., Nok Chiu, S. (1992) *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams.* John Wiley

Prince, I.M. et al. (2001) The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology* , 34, 1083-1088

Stoyan, D., Kendall, W. S. and Mecke, J. (1995) *Stochastic geometry and its applications.* John Wiley

Walker, S., Muliere, P. (2003) A bivariate Dirichlet process . *Statistics &*

Probability Letters, 64, 1-7

Wood, S., Kohn, R. (1998) A Bayesian approach to robust binary nonparametric regression. *Journal of the American Statistical Association*, 93, 203-213

Hug, D., Reitzner, M., Schneider, R. (2004) The limit shape of the zero cell in a stationary Poisson hyperplane tessellation. *Annals of Probability,* 32, 1140-1167.

Robert, C.P., Casella, G. (2004) *Monte Carlo Statistical Methods,* New York: Springer-Verlag

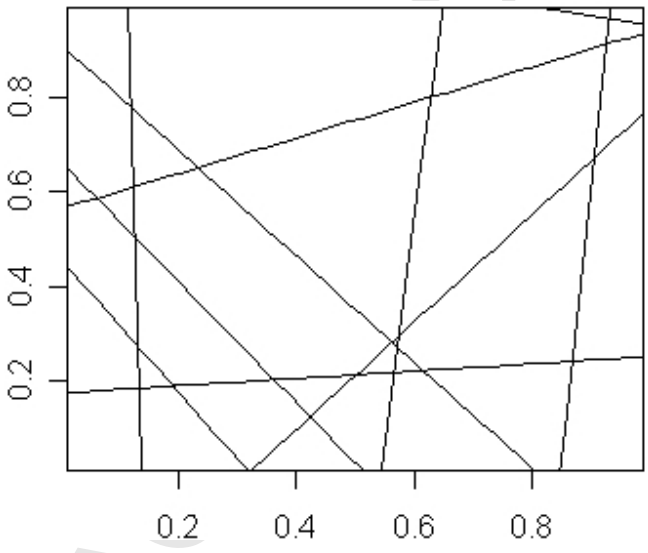
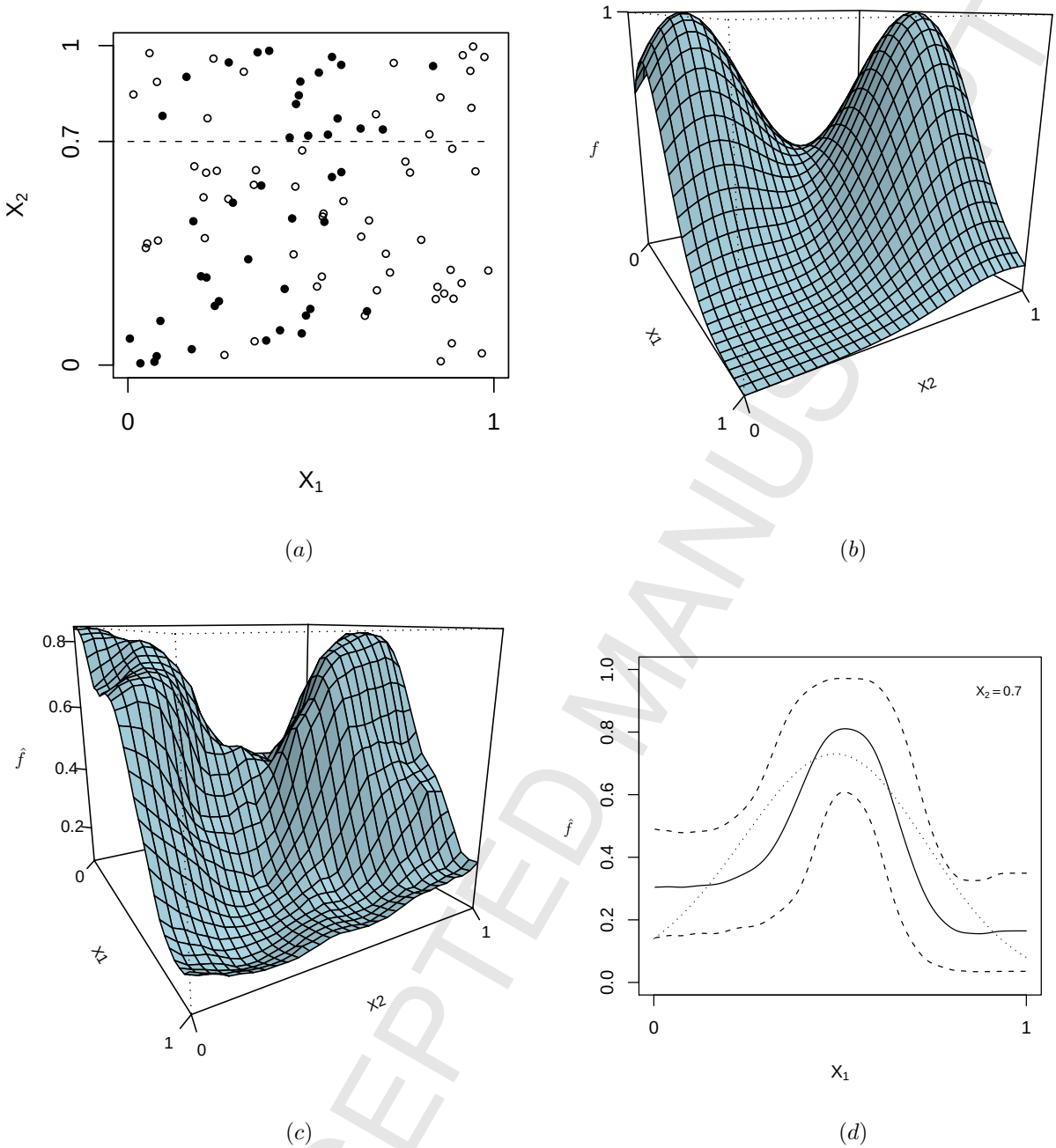


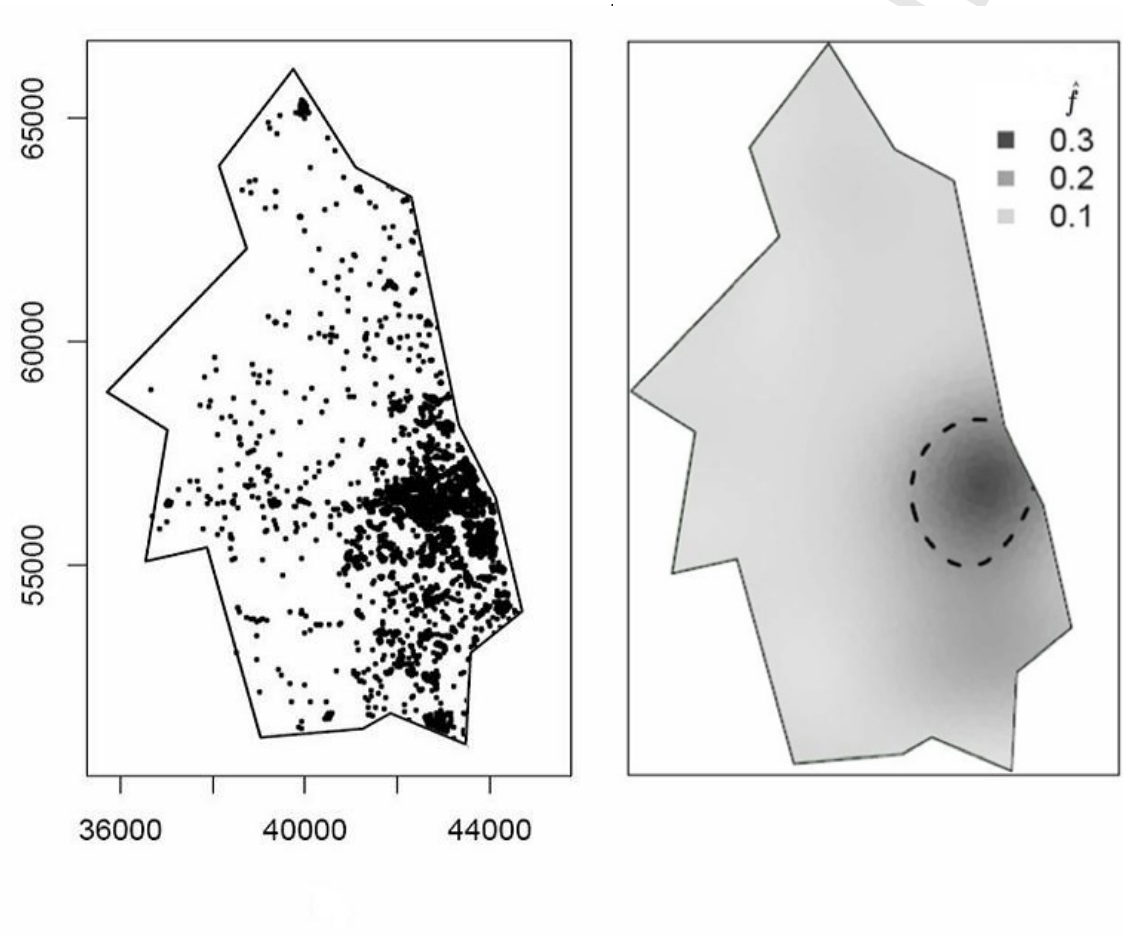
Figure 1. Bidimensional hyperplane tessellation.



(a) (b)

(c) (d)

Figure 2. Panel (a) illustrates the simulated data set, $\circ = 0$ and $\bullet = 1$. Panel (b) represents the response probability function f used for generating the data. Panel (c) illustrates the estimate $\hat{f}$. Panel (d) shows unidimensional sections of $\hat{f}$ (solid line) and f (dotted line), and illustrates the 90% credible intervals of f (dashed lines).



(a) (b)

Figure 3. Panel (a) illustrates a polygonal approximation of the study area and the places of residence of the 3805 individuals involved in the study. The overall size of the study region is 7.581 km^2. Panel (b) illustrates the estimate $\hat{f}$. The sites with the highest values of $\hat{f}$ are colored with the darkest tones of grey.