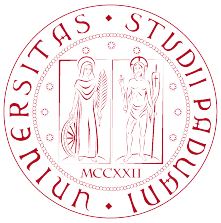


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXIV

BAYESIAN NONPARAMETRIC MODELS FOR COUNT DATA WITH APPLICATIONS TO CUSTOMER BASE MANAGEMENT

Direttore della Scuola: Alessandra Salvan
Supervisore: Bruno Scarpa
Co-supervisore: David B. Dunson

Dottorando: Antonio Canale

Data consegna tesi 30 Gennaio 2012

A Giò

Occhio per occhio,
si finisce per diventare ciechi...

Acknowledgements

At the end of this PhD adventure there are few people that I sincerely want to thank.

First of all thanks to Bruno, a real advisor. He was always patient and comprehensive with me and able to understand my needs while keeping me working hard, pushing my attention to topics that three years ago I would have considered too difficult. He gave me important advices and prevent me to waste time in any way, trying to let my PhD to be more fruitful as possible. We had great time together discussing about Statistics and life in general both in his office in Padua or in a taqueria in San Crisobal de las Casas.

Thanks to David, a brilliant professor and a great teacher. I learned a lot on how to do research working with him. I enjoyed the several discussions we had together and the emails in which he patiently commented my results and pointed out my poor English. It was great to visit him at Duke University for one year.

Thanks also to Alessandra Salvan, the head of the PhD school for her advices, helpfulness and generally for running a great PhD school.

Finally, thanks to my fellow travelers: Riccardo, Nicola, Davide, Monjed, Checca and Marlies. We built a great team in the first year and I learned a lot studying and discussing together. I also thank my friends at Duke and in particular Debdeep, Anirban, Francesca and Andrew for the discussions we had and generally for being welcoming with me.

Padua, January 26, 2012 Antonio Canale

Sommario

Motivati dall'analisi di dati di marketing nelle telecomunicazioni, solitamente multidimensionali, longitudinali e per lo più composti da conteggi, questo lavoro di tesi introduce nuove tecniche bayesiane nonparametriche per la stima delle funzioni di probabilità e la modellazione dei processi stocastici a valori interi. Sono introdotti inoltre i fondamenti teorici per la stima di densità congiunta con variabili su scale di misura miste (continue, conteggio e categoriali) tramite modelli mistura nonparametrica. Sebbene i modelli bayesiani nonparametrici per variabili continue siano ben sviluppati, la letteratura su approcci simili per dati di conteggio è scarsa, mentre quella per dati su diverse scale di misura è praticamente inesistente. L'idea principale di questo lavoro è quella di indurre distribuzioni a priori sugli spazi astratti di interesse tramite distribuzioni a priori su appropriati spazi latenti e funzioni di mappatura. Nello specifico, attraverso *a priori* sullo spazio delle densità countinue è introdotta una nuova classe di *a priori* sullo spazio delle funzioni di probabilità discrete e a scala di misura mista, mentre attravesto *a priori* sullo spazio dei processi stocastici a valori continui è introdotta una classe di *a priori* sui processi stocastici di conteggio. Le proprietà asintotiche di queste procedure sono studiate e, sotto opportune ipotesi, vengono dimostrati risultati sull'ampiezza del supporto e sulla consistenza dell'*a posteriori*. Vengono inoltre sviluppati efficienti algoritmi di campionamento di Gibbs per il calcolo delle *a posteriori*. Le prestazioni dei metodi proposti sono verificate tramite studi di simulazione e applicazioni a dati reali.

Abstract

Motivated by the analysis of telecommunications marketing data, which are multidimensional, longitudinal and mostly consisting in counts, this thesis introduces novel Bayesian nonparametric techniques for the estimation of probability mass functions and count stochastic processes. In addition, the theoretical basis of nonparametric mixture models for mixed-scale density estimation are provided. Mixed-scale data consists in joint continuous, count and categorical variables. Although Bayesian nonparametric models for continuous variables are well developed, the literature on related approaches for counts is limited and that for mixed-scale variables is close to none. The leading idea of this work is to induce prior distributions on the spaces of interest via priors on suitable latent spaces and mapping functions. Precisely a class of priors on the space of the probability mass functions and of the mixed-scale densities is induced through priors on the space of continuous densities and another class of priors on count stochastic process is induced through priors on the space of continuous stochastic processes. Asymptotic properties of these procedures are studied and results in terms of large support and posterior consistency are obtained under suitable assumptions. Efficient Gibbs samplers are developed for posterior computation, and the performance of the proposed methods is assessed in simulation studies and real data applications.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

Every day mobile phone operators collect plenty of information on the usage and on the behavior of their customers. This kind of data are usually multivariate, longitudinal and collected for a very large sample. The information contained by these data is of dramatic interest for companies and a better understanding of the customer behavior through these data is one of the main goal of the companies strategy. Motivated by this application we developed this PhD thesis work.

One of the complications of this data is that the variables consist in counts or in mixed counts and continuous variables. Since we find a lack of theory and methods to deal with count variables, especially in the multivariate and longitudinal context, we decide to develop novel methods to model counts under a Bayesian nonparametric framework while keeping in mind our motivating application and going back to it in testing the proposed models and procedures to fit real data. Bayesian nonparametric methods have recently received a lot of attention in the statistical literature. Personally we find interesting the idea of flexibility and large support induced by prior distribution on a infinite dimensional space, while appealing the possibility of including prior information to the data analysis.

To tackle the challenge of giving a general framework to model longitudinal multidimensional count data, we started to decompose the problem in subproblems. First we decide to drop the time dependence and to study the theoretical and methodological basis to work with counts at one time observation. Later we proceed reducing the dimensionality to one and introducing a flexible framework for count stochastic process estimation. A brief review of the literature that partially deals with these problems is given in the next section while Section 1.2 describes the main contributions of the thesis.

1.1 Overview

Nonparametric methods are well developed in the Bayesian literature for a vast range of applied problems with density estimation, regression and

functional data analysis among others.

Density estimation, under a Bayesian setting, is usually performed using a Dirichlet process mixture of Gaussians kernels (Lo, 1984; Escobar and West, 1995) to obtain a prior for the unknown density. A detailed description of the Dirichelet process (DP) of Ferguson (1973, 1974) and related priors is given in Chapter 2. Such a prior can be chosen to have dense support on the set of densities with respect to Lebesgue measure. Ghosal et al. (1999) show that the posterior probability assigned to neighborhoods of the true density converges to one exponentially fast as the sample size increases, so that consistent estimates are obtained. Similar results can be obtained for nonparametric mixtures of various non-Gaussian kernels using tools developed in Wu and Ghosal (2008).

Nonparametric regression under the Bayesian paradigm can be broadly divided in two main groups: (a) simplify the problem by considering a basis representation and assigning a prior on the basis coefficients. Splines, wavelets and reproducing kernels fall in this category with Bayesian P-splines (Lang and Brezger, 2004) as a standard tool; (b) assume that the regression function is a realization of a stochastic process and use a Gaussian Process prior (Rasmussen and Williams, 2006). Gaussian processes are computationally convenient and as n increases lead to posterior consistency (Ghosal and Roy, 2006).

Functional data consist in modeling n different subject specific trajectories varying through a domain set (usually time or space). A common approach deals with nested and hierarchical basis representation which allow variability in the functions assuming normally distributed basis coefficients. In the Bayesian framework Bigelow and Dunson (2007) and Thompson and Rosen (2007) recently introduce relative approach using adaptive splines. In the presence of prior information about the shape of the trajectories, as in the motivating examples of telecommunications companies, recent developments under semi and nonparametric Bayes are Scarpa and Dunson (2009, 2011).

For nonparametric probability estimation, regression and functional data analysis, there are lack of theory and methods to deal with count data. Even more if we consider mixed-scale data consisting of binary, categorical, continuous and count measurements the literature is close to none.

Few strategies have been proposed in the literature for nonparametric modeling of count distributions having support on $\mathbb{N} = \{0, \ldots, \infty\}$ and all in the univariate case. One consists in mixing Poisson distributions

$$\Pr(Y = j \mid P) = \int \text{Poi}(j; \lambda) dP(\lambda), \quad j \in \mathcal{N},$$

with $\text{Poi}(j; \lambda) = \lambda^j \exp(-\lambda)/j!$ and P a mixture distribution. When P is chosen to correspond to a $\text{Ga}(\phi, \phi)$ distribution on the Poisson rate param-

eter, one induces a negative-binomial distribution, which accounts for overdispersion. Poisson mixtures are reviewed in Karlis and Xekalaki (2005).

A more flexible nonparametric approach consists in choosing the DP as mixing distribution letting $P \sim \mathrm{DP}(\alpha P_0)$, with P_0 a base probability measure over the real line. Krnjajic et al. (2008) recently considered a related approach motivated by a case control study. At a glance, the Dirichlet process mixture (DPM) of Poissons seems the natural counts counterpart of the DPM of Gaussians used for continuous density estimation. However, the resulting prior on the count distributions is quite inflexible, as the Poisson kernel has a single parameter ruling out both location and scale. Clearly this model cannot consistently estimate underdispersed count distributions.

As an alternative one can use the almost sure discreteness property of the DP and avoid the mixture specification. With this approach one can let $y_i \sim P$ with $P \sim \mathrm{DP}(\alpha P_0)$ and P_0 corresponding to parametric count distribution, such as a Poisson. Carota and Parmigiani (2002) proposed a generalization of this approach in which they modeled the base distribution as dependent on covariates through a Poisson log-linear model. Although this model is clearly flexible, there are some major disadvantages. Given iid draws $y^n = (y_1, \ldots, y_n)^{\mathrm{T}}$, in fact, the resulting posterior distribution is

$$(P \,|\, y^n) \sim \mathrm{DP}\left((\alpha + n)\left\{\alpha P_0 + \sum_i \delta_{y_i}\right\}\right),$$

with δ_y a degenerate distribution with all its mass at y. The posterior is centered on a mixture with weight proportional to α on the Poisson base P_0 and weight proportional to n on the empirical probability mass function. There is no allowance for smooth deviations from the base.

If we found some proposals for the univariate case, the literature on multivariate methods for count data is very short and mostly relies on multivariate Poisson models (Johnson et al., 1997) which have the unpleasant characteristic of not allowing negative correlation. Copula models are an alternative approach to model the dependence among multivariate data with the proposal of Nikoloulopoulos and Karlis (2010) that directly deals with multivariate counts. A very flexible copula model that considers variables having different measurement scales (counts, continuous, binary) is proposed by Hoff (2007). Unfortunately the latter method is focused only on modeling the association among variables with the marginals treated as a nuisance and hence one cannot do any inference on the marginals or on any conditional distribution. Other flexible approaches include to consider mixtures of Poissons (Meligkotsidou, 2007) and random effects model, which incorporates shared latent factors in Poisson log-linear models for each individual count (Moustaki and Knott, 2000; Wedel et al., 2003). The latter models can deal also with mixed scale variables by defining a separate generalized linear model for each variable, with shared latent variables to induce dependence structure

(Sammel et al., 1997; Dunson, 2000, 2003). This framework assumes that the observed variables are independently drawn from distributions in the exponential family conditionally on latent variables. In marginalizing out the latent variables, one obtains a multivariate distribution with essentially unknown properties and computation can be quite challenging. In certain cases, pitfalls can arise due to the dual role of the latent factors in controlling the dependence structure and the shape of the marginal distributions.

Given these issues, it is quite appealing to consider flexible nonparametric models to estimate joint mixed-scale distributions. Somewhat surprisingly given the considerable applied interest, the literature on nonparametric estimation for this is very small. Some frequentist proposal can be found in the papers of Li, Racine and co-authors (Li and Racine, 2003; Hall et al., 2004; Ouyang et al., 2006; Li and Racine, 2008) that use a kernel smoothing approach and in the recent work of Efromovich (2011). At the moment in which we are writing this dissertation it seems that no Bayesian nonparametric literature is available on this topic.

With count regression we model a response count variable $y \in \mathbb{N}$ conditionally on some explanatory variables. From a Bayesian perspective a way of seeing regression is to assume that the regression function is a realization of a stochastic process $y = \{y(s), s \in \mathcal{D}\}$. Here y is a collection of count random variables indexed by $s \in \mathcal{D}$ with $\mathcal{D}$ a domain set usually corresponding to time or space and $y(s)$ a random variable observed at a specific time or location s.

The literature on stochastic processes is rich both from a frequentist and Bayesian point of view. Common choices includes the Gaussian processes (GP) (discussed later in Section 2.2.1) and Lévy processes, such as the Poisson, Wiener, beta or gamma process. GP provides a convenient and well studied choice for real value stochastic processes.

Integer valued stochastic processes, also known as count processes, are widely studied but mostly rely on Poisson hierarchical specifications. For example, Frühwirth-Schnatter and Wagner (2006) consider $y(s) \sim \text{Poisson}\{\lambda(s)\}$ with the Poisson mean $\lambda(s)$ varying over $\mathcal{D}$ according to a latent process. Rue et al. (2009) recently developed an integrated nested Laplace approximation to the posterior for a broad class of latent Gaussian structured additive regression models. The observed variables are assumed to belong to an exponential family (Poisson for counts), with the means given an additive model having Gaussian and Gaussian process priors on the unknown components. Although such models have a flexible mean structure, the Poisson assumption is clearly restrictive, having one single parameter ruling out both mean and variance. This leads to a pitfall in which the dependence structure in the data is perfectly confounded with the degree of overdispersion in the marginals in that both are induced through the latent Gaussian

process. Such models unfortunately cannot accommodate correlated counts that are under-dispersed. In addition, even if the model assumptions are approximately correct, computation is challenging for Poisson latent process models. This is particularly the case for count functional data in which we have $y_i = \{y_i(s), s \in \mathcal{D}\}$, with y_i the count stochastic process for subject i, for $i = 1, \ldots, n$.

Also here copula models are a useful approach to separate the marginal from the dependence structure. Wilson and Ghahramani (2010) recently proposed a Gaussian copula process model to characterize dependence between arbitrarily many random variables independently of their marginals, and applied their framework to stochastic volatility models. However, it is not clear how to apply their framework without using Poisson marginals, and even in this case conceptual difficulties and substantial computational hurdles may arise. Rodríguez et al. (2010) proposed a latent stick-breaking process, which is a nonparametric Bayes approach for a stochastic process with an unknown common marginal distribution modeled via a stick-breaking prior. They considered a spatial count process application, with the marginal modeled via a mixture of Poissons and the spatial dependence characterized through a latent Gaussian process. This successfully separates the marginal and dependence structure, but the marginal model is nonetheless restrictive in being characterized as a mixture of Poissons, computation is intensive, and count functional data are not accommodated.

1.2 Main contributions of the thesis

The main contributions of this thesis can be found in Chapter 3 in which a new class of prior distributions on the space of probability mass functions is introduced, in Chapter 4 where we study the asymptotic properties of Bayesian procedures of mixed-scale density estimation and in Chapter 5 in which the nonparametric count regression is approached treating the regression function as a stochastic count process.

In Chapter 3 we propose the leading idea of this work, based on rounding continuous objects, in this case, continuous density functions. Precisely we introduce a new class of prior distributions on the abstract space of the probability mass functions through a prior on a latent space and suitable mapping functions. Theoretical results on the topology on the two abstract spaces in exam are given showing that suitable assumptions on the underlying prior lead to large support of the induced prior with almost all count distributions falling within its Kullback-Leibler support. This is shown to imply both weak and strong posterior consistency thanks to Schwartz (1965) Theorem on one side and to the relations between the strong and weak topology in the space in exam on the other side. A general efficient Gibbs sampler is developed for posterior computation for the entire class of prior. Focusing

on rounded Gaussian for simplicity and on rounded skew normal (Azzalini, 1985) for better fit asymmetric patterns in the data, particular Gibbs samplers are developed and implemented in R and C. Simulation studies are performed to assess performance of the methods. Generalization of the modeling framework to account for multivariate count data are also shown. This natural extension is of particular interest since usual parametric and semiparametric models for multivariate count distribution are extremely inflexible. Part of the results of Chapter 3 are discussed in an accepted article on the Journal of the American Statistical Association (Canale and Dunson, 2011).

In Chapter 4 we generealize the framework of Chapter 3 to jointly model continuous, count and categorical variables under a nonparametric prior. For the proposed class of priors, we provide sufficient conditions for large support, strong consistency and rates of posterior contraction.

In Chapter 5, considering the longitudinal nature of our telecommunications data, we developed Bayesian nonparametric methods to model count stochastic processes. We introduce a novel class of Bayesian nonparametric count process models, which are constructed through rounding real-valued stochastic processes. Theoretical results on large support and posterior consistency are established under suitable assumptions on the stochastic process and on the observation points, and suitable computational algorithms are developed modeling the underlying process as a regression function estimated through P-splines (Jullion and Lambert, 2007). This rounded P-spline approach is then extended with a hierarchical representation in the case of n related count processes.

Applications of the proposed methods to customer base management in telecommunications market can be found both in Chapter 3 and Chapter 5. Despite the particular application context, the methodology can be applied in any other settings dealing with counts such as toxicology (number of tumors), ecology (number of species or animals in a given area) or epidemiology (number of ill patients) among others. Examples related to toxicology are also reported in the main chapters.

Chapter 2

Background

2.1 Customer base management and Statistics

Statistics is applied in several company sectors, ranging from production to sales. The interest here is on marketing analysis (Lehmann et al., 1998) and particularly on customer base management of telecommunications companies.

Telecommunications market is a high profitable market world wide. Particularly in emerging countries mobile markets has an high rate of development and telecommunications companies still have the acquisition of new customers as marketing goal. China for example has one of the largest mobile communication network in the world with the subscriber number grown from only 3 millions in 1990 to over 641 millions by the end of 2008 (MIIT, 2009) with a current population of 1,336 millions of people. European and North American mobile telephone market instead have already grown from the early 90's and the market is now saturated. The average number of mobile phone subscriptions per 100 inhabitants in Europe, for example, stood at 122 in 2008 (Eurostat, 2011), meaning that there are more subscriptions than people. In this context, clearly, companies paradigm has arguably changed from an acquisition orientation to a retention orientation.

In this market context, customer base management has become one of the main important business strategies. Customer base management is the marketing branch that manages the company's interactions with its customers which overall goal is to retain and increase the value of existing customers, while enticing former or new clients. Under this framework, it is evident that the knowledge of the behavior and of the characteristics of customers becomes dramatically important.

Defining homogeneous clusters of customers, for example, is a key to perform *ad hoc* marketing campaigns or promotions limiting the cost of the marketing actions only towards those customers with high potential positive outcome. Also, customer profiling is very important in defining the

positioning of products and the marketing strategy. Particularly in mature and highly competitive markets, customers exercise their right of switching and hence the churn rate, the measure of the number of individuals quitting the company, is one of the main marketing indicators to control. Churn, in fact, is costly. In 2004 in the US wireless market, the retention cost of a customer was estimated at 60$ while the one to acquire a new one at 400$ (Strouse, 2004). There are several strategies to control the churn rate but they can be divided into two main fields. The first consists in creating barriers which discourage customers to change company. The impossibility of the mobile number portability, for example, was in Italy the main barrier until 2002. The second way of contrasting churn consists in preventing it. The identification of potential churners, in fact, leads specific retention activities such as loyalty programs and promotions.

It becomes evident that a quantitative knowledge of these phenomena is necessary and hence sophisticated statistical tools are needfuls. From a concrete point of view, statisticians have to face marketing challenges using the data available. Nowadays telecommunications companies storage terabytes of data for each customer since every action of the customer can be potentially recorded. In mobile phone market, for example, the daily number of text messages, outgoing calls, services, Internet connections, downloaded applications and so forth is recorded for each customer.

Standard data mining tools (Hastie et al., 2001; Azzalini and Scarpa, 2012), can be used to perform general tasks such as a cluster analysis, churn or profit predictions but sometimes these tools are used inappropriately. In churn prediction, for example, a common data mining practice is to use the information about the traffic of past months to predict the churn. Usual methods (Nath and Behara, 2003, e.g.), unfortunately, do not treat the usage observations as time dependent variables, ignoring any sort of auto-correlation. Considering the longitudinal nature of these data can lead, for example, to a better prediction of the churn or to the customer base profiling based on time series clustering. In the biomedical context there are some attempts to use this approach (Wang et al., 2000; Li et al., 2004; Bigelow and Dunson, 2007; Dunson et al., 2008) but none of them has been applied to churn prediction nor considered count variables.

2.1.1 A real telecommunications dataset

In this section we introduce a dataset of a European UTMS service provider and a set of possible concrete problems that lack of statistical methods to be solved. The company is the first UMTS based provider in the market and it built its brand image pushing the role of the video calls.

The dataset contains informations about 29,315 randomly chosen customers active during 18 months. It consists of both longitudinal records of traffic and usage and static variables such as demographic informations

of who activated the contract (sex, age, region) or information on the contract itself (type, distribution channel, mobile number portability). A last binary variable records if the contract was still active in the 19th month or if otherwise the customer churned is also available. The total proportion of deactivated customers is equal to 6.52%. Among the longitudinal variables we found the monthly number of text and multimedia messages sent, the duration and the number of the incoming, outgoing to landlines, outgoing to numbers of other operators, outgoing to numbers of the same operator, and video calls. These longitudinal records are hence multidimensional and on a mixed scale of measure including count and continuous variables. The mean trajectories for the counts, which are the object of this thesis work, are reported in Figures 2.1–2.2. Table 2.1 reports some descriptive statistics for each months.

For a better knowledge of the customer base it is sometimes of interest to estimate the distribution of the number of events, e.g. outgoing phone calls to landlines, made by customers stratified by some variable such as age, employment or geographic area. Such a distribution is usually highly skewed, zero inflated and fat tailed. To estimate the probability mass function of the monthly number of outgoing phone calls for the customers divided by geographic area, in Section 3.5.1 we use a Bayesian nonparametric mixture of rounded skew normal kernels.

Another applied problem consists in forecasting a count variable y_{i1}, using data on $y_{i2}, \ldots, y_{ip}$. In our case, for example let y_{i1} the number of outgoing calls to landlines, y_{i2} to mobile numbers managed by competing operators and y_{i3} to mobile numbers of the same operator; y_{i4} and y_{i5} are then the total number of multimedia and text messages sent. For the company, both the point forecast of y_{i1} or the estimation of the probability that y_{i1} equals zero or is greater than a given threshold, is of high interest since this kind of calls has a high cost for the service provider. In Section 3.5.2 we jointly model the multivariate probability mass function of $\mathbf{y} = (y_{i1}, \ldots, y_{i5})$ using a Bayesian mixture and then predicting the value of y_{i1} given $y_{i2}, \ldots, y_{i5}$.

A very important task, as already discussed, is the churn prediction using the dynamic variables on traffic, e.g. the outgoing number of video calls. Our dataset presents only 1,284 deactivations over all the 29,315 customers which is a peculiar characteristic of this context. In Section 5.5.1 we analyze the outgoing video calls traffic to predict churn.

Table 2.1: Descriptive statistics for the count variables for each month.

	Video calls			Calls to same operator			Calls to other operators			Calls to landlines			Text message			Multimedia message		
	Mean	Sd	Range	Mean	Sd	Range	Mean	Sd	Range	Mean	Sd	Range	Mean	Sd	Range	Mean	Sd	Range
1	4.97	9.59	222	27.86	47.6	962	28.73	46.63	666	10.17	17.61	249	21.36	57.46	1270	1.09	4.11	110
2	4.86	8.68	176	31.9	51.58	882	34.51	51.87	738	12.04	19.69	328	29.35	68.75	1341	1.77	6.68	253
3	5.03	9.9	263	29.19	47.91	614	29.48	45.02	508	10.79	17.39	204	25.85	65.19	1290	1.64	7.27	303
4	4.82	10.24	223	29.7	58.32	2816	27.69	46.18	867	10.34	18.01	246	22.58	61.48	1433	1.44	7.07	323
5	5.13	12.21	443	32.68	54.42	758	31.89	53.3	815	11.36	19.51	254	25.63	70.82	2326	1.38	6.98	334
6	4.1	7.98	206	21.73	35.76	618	18.3	29.65	385	6.72	12.26	165	13.84	38.97	824	0.87	4.14	153
7	5.05	9.96	196	38.77	62.96	1030	32.58	52.91	1086	11.7	20.66	314	23.37	60.89	1050	1.29	6.5	328
8	4.81	9.72	202	38.6	61.27	728	33.51	53.43	957	11.57	19.03	238	25.01	75.77	1875	1.32	6.42	264
9	4.99	10.78	281	39.05	61.33	744	32.58	50.01	873	11.03	18.43	301	25.67	77.64	2242	1.4	6.15	251
10	5.04	10.76	246	35.17	56.73	682	29.42	46.15	421	9.27	16.02	213	25.97	75.26	2285	1.61	6.64	175
11	4.82	10.31	183	36.93	59.54	598	30.17	49.73	698	11.02	19.53	308	25.83	74.02	1377	1.47	7.79	259
12	4.75	11.06	243	37.42	61.75	750	28.68	51.41	751	11.1	20.27	458	23.5	69.98	1802	1.26	8.42	296
13	4.55	10.81	257	34.58	56.02	592	27.05	45.07	645	10.52	19.27	280	22.87	70.34	1766	1.26	8.08	277
14	4.25	9.13	173	35.73	57.93	636	28.93	45.49	501	11.55	19.97	269	30.8	74.83	1431	1.51	8.1	302
15	4.11	7.41	125	36.32	59.68	658	28.96	46.06	518	11	19.58	259	27.61	73.31	1989	1.59	7.91	226
16	4.62	11.88	421	35.54	56.33	678	27.08	42.16	450	10.86	18.92	236	24.46	74	1609	1.39	8.1	269
17	4.69	10.78	313	39.05	66.43	1018	30.51	56.79	918	11.13	19.52	258	27.56	86.3	1710	1.58	9.98	227
18	3.46	5.98	138	33.31	51.21	480	31.48	51	623	11.69	19.88	234	17.76	46.52	776	1.3	7.17	126

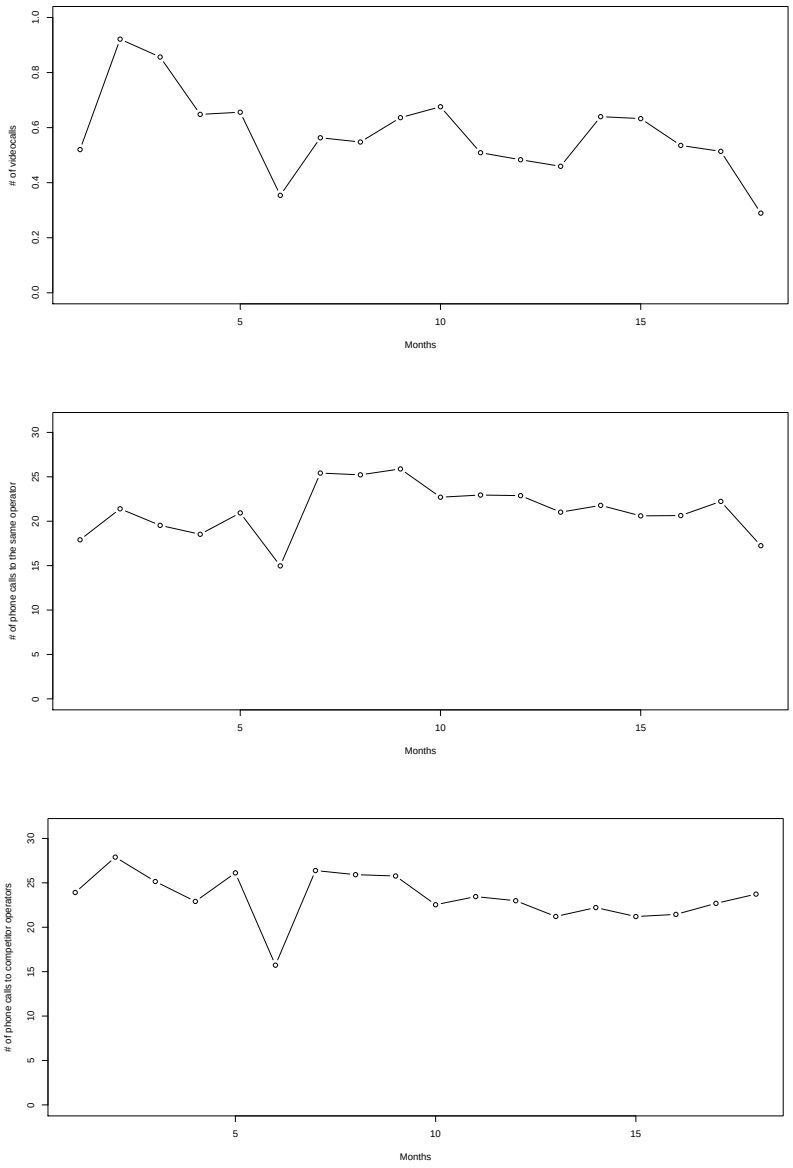


Figure 2.1: Mean trajectories over 18 consecutive months of outgoing video calls (a), outgoing phone calls to the same operator (b) and outgoing phone calls to competitor operators.

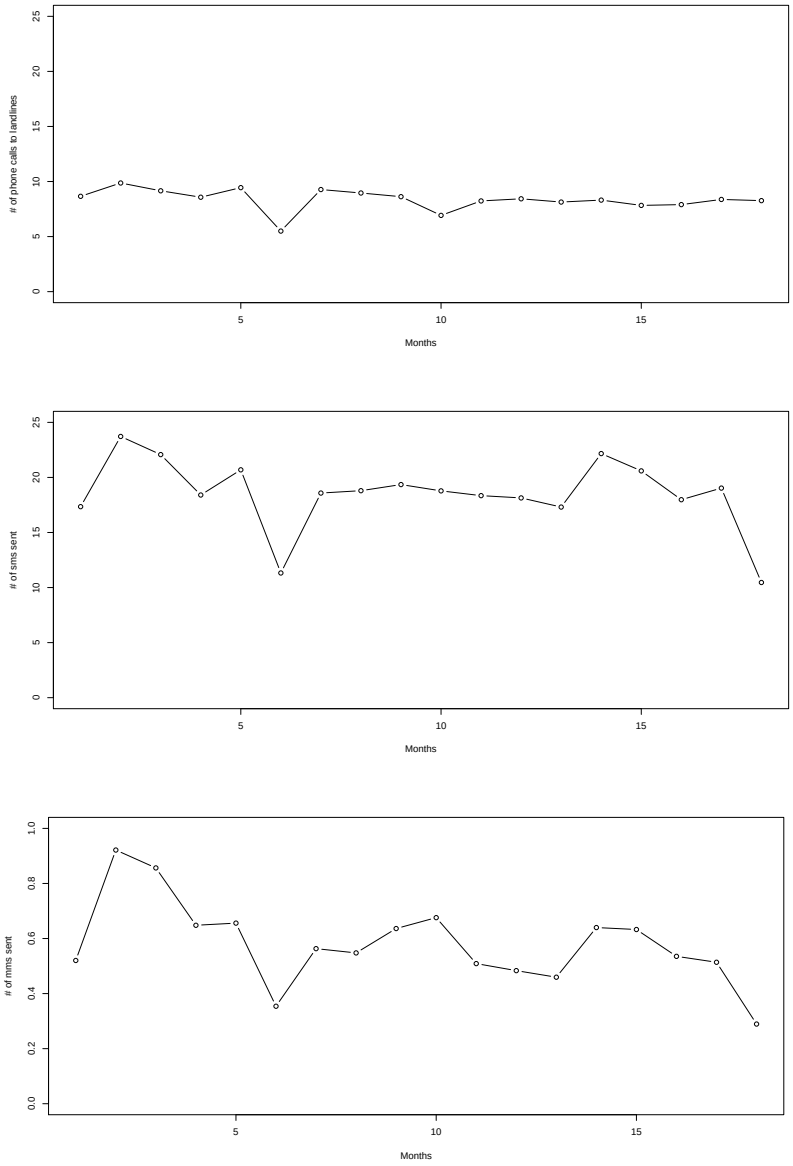


Figure 2.2: Mean trajectories over 18 consecutive months of outgoing calls to landlines (a), number of text message (b) and multimedia message (c) sent.

2.2 Statistical tools

In this section we move our attention to the statistical tools used to develop this thesis project. We are clearly not going to give a comprehensive picture of the methods but rather some basic references.

Bayesian nonparametric methods have recently received a lot of attention in the statistical literature. The first theoretical results go back to the 70's but it is in the past twenty years that the scientific literature on this topic dramatically increases also pushed by the considerable success in a lot of applied fields such as biostatistics or machine learning. After reviewing some results on the Gaussian process, that will be used in Chapter 5, we give a brief review the main Bayesian nonparametric tools used in this dissertation. A more comprehensive review on Bayesian nonparametric methods can be found in the book of Ghosh and Ramamoorthi (2003) and in the book edited by Hjort, Holmes, Müller and Walker (2010).

2.2.1 The Gaussian process

A Gaussian process (GP) is a stochastic process $\{y(t); t \in T\}$, where T is a domain space (usually time or space), for which for every finite set of indices $t_1, \ldots, t_n$ the vector $\{y(t_i), \ldots y(t_n)\}$ is normally distributed. In particular we write

$$y(t) \sim \mathcal{GP}(\mu(t), k(t, t^{'}))$$

where $\mu(t)$ is the mean function and $k(t, t^{'})$ is the covariance function of the process. Mean and covariance functions are defined such that

$$E[y(t)] = \mu(t), \quad E[(y(t) - \mu(t))^T (y(t^{'}) - \mu(t^{'}))] = k(t, t^{'}).$$

A common choice for the covariance function consists in the so called squared exponential covariance function, i.e.

$$k(t, t^{'}) = \exp\left(-\frac{1}{2}|t - t^{'}|^2\right).$$

The squared exponential covariance function plays an important role when the GP is used as mean function in a regression context. It can be shown in fact that a GP regression with such a covariance function corresponds to a Bayesian linear regression model with an infinite number of basis functions (Zhu et al., 1998).

In the Bayesian literature, there have been substantial theoretical and computational advances for GP models in recent years. For example, Ghosal and Roy (2006) show that the GP is appealing in providing a prior that can be specified to generate functions that are within an arbitrarily small neighborhood of any continuous function with positive probability and van der

Vaart and van Zanten (2009) study asymptotic properties including posterior consistency and rates of convergence. From the computational point of view, Banerjee et al. (2008) and Murray and Adams (2010) develop improved methods for posterior computation.

Historically the GP was introduced in the Bayesian framework in the late 70's as nonparametric regression priors (O'Hagan, 1978; Wahba, 1978). Nonetheless these groundbreaking papers, GP modeling remained hidden until the early 90's when the machine learning community start to use the GP for regression and classification. The standard introductory reference to the topic is in fact the book of Rasmussen and Williams (2006).

2.2.2 The Dirichlet process and related priors

The Dirichlet process was introduced by Ferguson (1973, 1974) with the idea of introducing a prior for nonparametric problems. Ferguson (1973) wrote

> *There are two desirable properties of a prior distribution for nonparametric problems.*
>
> - *The support of the prior distribution should be large – with respect to some suitable topology on the space of probability distributions on the sample space.*
> - *Posterior distributions given a sample of observations from the true probability distribution should be managed analytically*

With this in mind he introduced the DP, a probability distribution on the space of probability measures. Let for example $\mathcal{Y}$ a space and $\mathcal{B}$ the Borel σ-algebra on $\mathcal{Y}$. Let denote the base measure with P_0 a finite non null measure on $(\mathcal{Y}, \mathcal{B})$ and $\alpha \in \mathbb{R}$ be the concentration parameter characterizing prior precision. Then P is a Dirichlet process, namely $P \sim DP(\alpha P_0)$ if for any partition $(B_1, \ldots, B_k)$ of $\mathcal{B}$ we have

$$(P(B_1), \ldots, P(B_k)) \sim \text{Dirichlet}(\alpha P_0(B_1), \ldots, \alpha P_0(B_k)).$$

With this characterization, one can think to P as a random probability measure on $(\mathcal{Y}, \mathcal{B})$ and hence to the DP as a prior for the space of probability measure on $(\mathcal{Y}, \mathcal{B})$. The DP is conjugate with the multinomial likelihood. In a total nonparametric setting the posterior distribution of a DP given a sample $x_1, \ldots, x_n$ of iid observation from P is again a DP with precision parameters $\alpha + n$ and base measure $P_0 + \sum \delta_{x_i}$ where δ_x is a point mass of one in the points x. Some properties about moments of the DP easily follow. Let $A \subset \mathcal{Y}$, then the expectation $E[P(A)] = P_0(A)$ meaning that the expected draw from a DP is the base measure. Also observe that $\text{var}[P(A)] = P_0(A)(1 - P_0(A))/(\alpha + 1)$ meaning that the high precision parameter α let the prior to be concentrated around its mean.

A very important property of the DP is that if $P \sim DP(\alpha P_0)$, P is almost surely discrete, even when P_0 is purely non-atomic. This property can result disappointing if we want to estimate a density, which is by definition non-atomic. Nonetheless this is not a great drawback since even the empirical histogram, which is itself discrete, converges uniformly to any true distribution. In addition it can be shown that the topological support of the DP, i.e. the smallest closed set of probability one is quite big. The support contains in fact all the probability distributions that share the same support of the base measure P_0. Practically in the univariate context, if $P_0 = N(0, 1)$ the DP can generate any density function defined on the real line.

The almost sure discreteness of the process has, in addition, some appealing characteristics. Let $X_1, \ldots, X_p$ an iid sample from P and $P \sim DP(\alpha, P_0)$. Blackwell and MacQueen (1973) studied a sequential representation of the DP, that is

$$\begin{aligned} X_1|P &\sim P_0 \\ X_2|P, X_1 &\sim DP\left(\alpha + 1, \frac{\alpha}{\alpha+1}P_0 + \frac{1}{\alpha+1}\delta_{X_1}\right) \end{aligned}$$

that is, the second observation X_2 given X_1, is a new draw from P_0 with probability $\alpha/(\alpha + 1)$ and is equal to X_1 with probability $1/(\alpha + 1)$. Generalizing this we have

$$X_{j+1} \sim \begin{cases} \delta_{\theta_h}, & \text{with probability } \frac{n_h}{\alpha+j},\ h = 1, \ldots, k \\ P_0, & \text{with probability } \frac{\alpha}{\alpha+j} \end{cases}$$

where k is the number of distinct observations $\theta_1, \ldots, \theta_k$ and n_h is the number of X_i equal to θ_h. The different θ_h can be considered as clusters in which the observations fall. This representation, known as Blackwell-MacQueen generalized Polya urn scheme, has a key role in practical application since it can be used in Markov chain Monte-Carlo (MCMC) simulation from the posterior. This representation also stress the role of α in determining k and hence in characterizing the clustering structure of the DP. A colorful metaphors of the Polya-Urns scheme is the so called Chinese restaurant process. Consider a Chinese restaurant with an infinite number of tables, each with infinite number of seats. In this imaginary restaurant the first customer seats at an unoccupied table with probability 1. The general $j + 1$ customer seats at a new table with probability $\alpha/(\alpha + j)$ or choses one of the occupied tables with probability proportional to $n_h/(\alpha + j)$. Note that even if MCMC algorithms naturally produce at each iteration a clustering structure of the observations, the posterior interpretation of this clusters is not trivial. In fact both the actual number of occupied clusters and their composition varies at each iteration. This problem is typically referred to in the literature as the label switching problem (Stephens, 2000; Jasra et al.,

2005). If the goal of inference is cluster specific, some strategies have been proposed to solved the label switching problem. One technique consists in relabel the clusters at each MCMC iteration using a post-processing algorithm. This approach unfortunately tends to be time-consuming. Other approaches include to put identifiability constraints (Diebolt and Robert, 1994) and to perform *a posteriori* a hierarchical cluster analysis using a distance matrix depending from the MCMC output and the complete linkage principle (Medvedovic and Sivaganesan, 2002).

The stick breaking representation of Sethuraman (1994) is another useful representation of the DP, in which

$$P = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}, \;\; \theta \overset{iid}{\sim} P_0,$$

and $\pi_1 = V_1$, $\pi_h = V_h \prod_{l<h}(1 - V_l)$ with $V_h \sim \text{beta}(1, \alpha)$. The mechanism that generates the weight π_h gives the name to this representation since it may be thought as breaking a stick of length one into infinitely many pieces with length proportional to the sequence of weight.

A useful prior for density estimation can be constructed using the DP. Assume that the data $y_1, \ldots, y_n$ are iid from f and

$$f = \int K(y; \theta) dP(\theta), \quad P \sim DP(\alpha P_0)$$

where K is a positive kernel that integrates to one. The DP prior on P and the structure above induce a prior on f known as the Dirichlet process mixture (DPM) prior, a tool particularly useful in Bayesian models for density estimation. When K corresponds to the Gaussian distribution (Lo, 1984; Escobar and West, 1995) we get mathematical tractability and nice asymptotic properties in terms of large support and posterior consistency. Under a DPM of Gaussian, Ghosal et al. (1999) derive sufficient conditions on the prior and the true distribution f_0 in order to achieve strong posterior consistency with Tokdar (2006) strongly relaxing their conditions assuming a DP location-scale mixture of univariate Gaussians. Under the same modeling framework, Ghosal and van der Vaart (2001, 2007) give the rate of convergence for Bayesian univariate density estimation. Posterior consistency and Bayesian asymptotic are discussed in the next section.

2.2.3 Asymptotic topics in Bayesian inference

Some of the theoretical results of this work are in terms of posterior consistency. The notion of consistency in the Bayesian context can result inapt since it is commonly associated with frequentist estimators theory. Let us briefly analyze this apparent contradiction.

The aim of statistical inference under both the frequentist and the Bayesian reasoning is to draw meaningful conclusions about unknown objects. For

years there has been a strong division between frequentists and Bayesians not only for a different epistemological point of view on how to acquire knowledge, but also for a total different idea on the essence of the object to be known. If frequentists assume that there exists a fixed truth that generates the data, historically Bayesians avoid the idea of a true data generating process and rather rely on a subjective interpretation of the probability. For sake of this short introduction let oversimplify the problem and assume a parametric model indexed by the parameter θ. While frequentists assume that there exists a fixed true parameter θ_0, Bayesians assume that θ is a random variable itself. The formers will build an estimator $\hat{\theta}(X)$ of θ using the observed data X, being confident that if the experiment is repeated a number of times, approximately this many realizations of the estimator will be close to θ_0 even though nothing is known about the current $\hat{\theta}(X)$. The latter would consider a probability distribution of θ reflecting their own prior knowledge about θ. After the observation of the data, the prior belief is updated through the conditional distribution of the parameter given data. How could hence a Bayesian procedure be consistent if no θ_0 exists?

One possible approach to the Bayesian reasoning consists in not deny the existence of a fixed ground truth but rather admit the impossibility of fully discover it. The prior distribution can be hence seen more as a quantification of the uncertainty about the unobservable θ_0 while the posterior distribution reflects the remaining uncertainty about θ after observing the data. Hopefully, as we observe more and more data, this uncertainty will tend to decrease and our knowledge should get closer to the unobservable θ_0. Posterior consistency formalize this point of view.

Roughly speaking, posterior consistency means that the posterior distribution concentrates around θ_0 as $n \to \infty$. A more precise definition of posterior consistency is:

Definition 2.1. *Let Θ be an arbitrary topological space parameterizing a statistical model. Let $X_1, \ldots, X_n$ be a sample of size n from such a statistical model, $\mathcal{B}$ the Borel σ-algebra on Θ and Π a probability measure on $(\Theta, \mathcal{B})$ with $\Pi(\cdot \mid X_1, \ldots, X_n)$ being the induced posterior distribution given the data. The posterior distribution is consistent at $\theta_0 \in \Theta$ if for every neighborhood (defined with respect to a given topology on Θ) U of θ_0, said U^C its complement, we have*

$$\Pi(U^C \mid X_1, \ldots, X_n) \to 0$$

in probability under P_0^n or almost surely with respect to P_0^n, where P_0^n is a probability measure induced by θ_0.

Another view of posterior consistency is that of Bayesian robustness. Assume that two statisticians choose two different sets of prior distributions; as $n \to \infty$ we would like to have the same posterior inference. As formalized

in Diaconis and Freedman (1986) two different priors will agree *a posteriori* if and only if consistency holds.

In this dissertation, Θ will be an infinite-dimensional parameter space, and hence posterior consistency results will not follow trivially. A brief review of posterior consistency also from an historical point of view follows.

The first result in this direction is due to Doob (1949). Doob's theorem states that if exists a consistent estimator of θ, then the posterior distribution of θ will tend to concentrate near the true θ_0 with probability 1 under the joint distribution of the data and parameter. Even being groundbreaking, this results does not give any information about consistency at a specific θ_0 of interest. Moreover in the infinite-dimensional case, the set of θ_0 where consistency holds may be topologically very small.

The celebrated posthumous paper of Schwartz (1965) gives a result that is crucial for further posterior consistency studies in infinite dimensional spaces. Assume henceforth that Θ is $\mathcal{L}$, the space of densities with respect to a σ-finite measure on $\mathbb{R}$. Schwartz theorem gives sufficient conditions on the true $f_0 \in \mathcal{L}$ and the prior in order to get consistency of posterior distributions in the case of iid random variables. We give some definition before stating the theorem. Let Π be a prior on $\mathcal{L}$, $KL(f, g)$ the Kullback-Leibler divergence $\int f \log(f/g) d\mu$ and $\mathcal{K}_\epsilon(f)$ a ϵ size neighborhood $\{g : KL(f, g) < \epsilon\}$.

Definition 2.2. *Let $f_0 \in \mathcal{L}$. f_0 is said to be in the KL support of the prior Π, if for all $\epsilon > 0$, $\Pi(\mathcal{K}_\epsilon(f_0)) > 0$.*

Definition 2.3. *Let U be a neighborhood of $f_0 \in \mathcal{L}$. The sequence of test functions $\{\Phi_n\}$ is uniformly consistent for testing $H_0 : f = f_0$ versus $H_1 : f \in U^C$, if for $n \to \infty$*

$$E_{f_0}\{\Phi_n\} \to 0, \qquad \inf_{f \in U^C} E_f\{1 - \Phi_n\} \to 0.$$

Theorem 2.1 (Schwartz)**.** *Let Π be a prior on $\mathcal{L}$. If $f_0 \in \mathcal{L}$ and U a neighborhood around f_0 satisfy*

1. *f_0 is in the K-L support of Π,*
2. *there exists a uniformly consistent sequence of tests as in Definition 2.3,*

then $\Pi(U \mid X_1, \ldots X_n) \to 1$ a.s $P_{f_0}^\infty$

The two conditions of Schwartz theorem can be interpreted as follow. The first condition requires that non null prior probability is assigned to a neighborhood of the true f_0. Clearly if the prior distribution gives zero probability to a given region of the parameter space, the posterior will also give null probability to it. The second condition is an identifiability condition. The insight of the existence of the sequence of test is that as n increases

we should be able to better identify the true model that lies in the infinite dimensional parameter space.

Note that Theorem 2.1 is stated in a general form as what concern the neighborhood U. The neighborhood must be defined with respect to a particular topology. If U is a weak neighborhood the following result holds.

Theorem 2.2 (Schwartz)**.** *Let* Π *be a prior on* $\mathcal{L}$*. If* f_0 *is in the KL support of* Π *and* U *a weak neighborhood around* f_0 *then* $\Pi(U \mid X_1, \ldots, X_n) \to 1$ *a.s* $P_{f_0}^{\infty}$

The above result is based on the fact that if U is a weak neighborhood the construction of the sequence of test is an easy task. Nonetheless if U is a L_1 or Hellinger neighborhood such a sequence of test is not easy to construct.

Barron et al. (1999) and Ghosal et al. (1999) establish posterior consistency without invoking Schwartz's conditions. The new condition is on the size of the parameter space measured in terms of L_1-metric entropy, defined as

Definition 2.4. *Let* $\mathcal{G} \subset \mathcal{L}$*. For* $\delta > 0$ *the* L_1*-metric entropy* $J(\delta, \mathcal{G})$ *is defined as the logarithm of the minimum of all* k *such that there exist* $f_1, \ldots, f_k \in \mathcal{L}$ *such that* $\mathcal{G} \subset \bigcup_{j=1}^{k} \{f \in \mathcal{L} : \int |f(x) - f_j(x)|dx < \delta\}$*.*

We report Ghosal et al. (1999) theorem below.

Theorem 2.3 (Ghosal, Ghosh and Ramamoorthi)**.** *Let* Π *be a prior on* $\mathcal{L}$*. Suppose* $f_0 \in \mathcal{L}$ *is in the KL support of* Π *and let* $U = \{f \in \mathcal{L} : \int |f(x) - f_0(x)|dx < \epsilon\}$ *for each* $\epsilon > 0$*,* U *a strong neighborhood around* f_0 *If for each* ϵ *there is a* $\delta < \epsilon$*,* c_1*,* $c_2 > 0$*,* $\beta < \epsilon^2/2$*, and a sequence of set* $\mathcal{L}_n \subset \mathcal{L}$ *such that, for* n *large,*

- $\Pi(\mathcal{L}_n) < c_1 \exp(-c_2 n)$*,*
- $J(\mathcal{L}_n, \delta) < n\beta$*,*

then the posterior is strongly consistent at f_0*.*

Using Ghosal et al. (1999) Theorem to prove strong posterior consistency in non-compact spaces, a critical step is to introduce a compact subset $\mathcal{L}_n$ that is indexed by the sample size n and that grows to fill the entire space as $n \to \infty$. This sequence of subsets is typically referred to as a sieve. The size of this sieve in terms of L_1-metric entropy is required to grow slower than linearly in n, and the prior probability assigned outside of $\mathcal{L}_n$ (to $\mathcal{L}_n^C$) is required to decrease exponentially fast in n. Those conditions can be used to construct an exponentially consistent sequence of tests and then to upper bound the numerator and lower bound the denominator of $\Pi(U^C \mid X_1, \ldots X_n)$ ensuring it to be exponentially small end hence achieving strong posterior consistency.

Chapter 3

Probability mass function estimation

We focus now on the Bayesian probability mass function estimation. In this chapter we introduce the principal idea of this thesis, consisting in rounding continuous objects, here density functions. The topology on the space of count distributions is studied showing that under suitable assumptions on the underlying prior we achieve weak and strong posterior consistency. We introduce computational tools for MCMC simulation from the posterior and apply the modeling framework both in the univariate and multivariate case. We show how the proposed method can be used in some customer base management problems. Part of the results presented here are already discussed in Canale and Dunson (2011).

3.1 Rounded kernel mixture priors

3.1.1 Rounding continuous distributions

In the univariate case, letting $y \in \mathbb{N}$ denote a count random variable, our goal is to specify a prior Π for the probability mass function p of this random variable. Following the philosophy of Ferguson (1973), nonparametric priors for unknown distributions should be interpretable, have large support and lead to straightforward posterior computation. We propose a simple approach that induces Π through first choosing a prior Π^* for the density f of a continuous random variable $y^* \in \mathcal{Y}$ and then rounding $y^* \in \mathcal{Y}$ to obtain $y \in \mathbb{N}$. Here, $\mathcal{Y}$ is either the real line $\mathbb{R}$ or a measurable subset.

Let $y = h(y^*)$, where $h(\cdot)$ is a rounding function defined so that $h(y^*) = j$ if $y^* \in (a_j, a_{j+1}]$, for $j = 0, 1, \ldots, \infty$, with $a_0 < a_1 < \ldots$ an infinite sequence of pre-specified thresholds that defines a disjoint partition of $\mathcal{Y}$. For example, when $\mathcal{Y} = \mathbb{R}$ one can simply choose $\mathbf{a} = \{a_j\}_{j=0}^{\infty}$ as $\{-\infty, 0, 1, 2, \ldots, \infty\}$. The probability mass function p of y is $p = g(f)$, where $g(\cdot)$ is a rounding

function having the simple form

$$p(j) = g(f)[j] = \int_{a_j}^{a_{j+1}} f(y^*)\mathrm{d}y^* \quad j \in \mathcal{N}. \tag{3.1}$$

The thresholds a_j are such that $a_0 = \min\{y^* : y^* \in \mathcal{Y}\}$, $a_\infty = \max\{y^* : y^* \in \mathcal{Y}\}$ and hence $\int_{a_0}^{a_\infty} f(y^*)\mathrm{d}y^* = 1$. Examples of $a_0, \ldots, a_\infty$ include $0, 1, 2, \ldots, \infty$ for an f defined on $\mathcal{Y} = \mathbb{R}^+$ and $0, 1/2, \ldots, 1 - 1/2^h, \ldots$ for an f defined on $\mathcal{Y} = [0, 1]$.

Relating ordered categorical data to underlying continuous variables is quite common in the literature. For example, Albert and Chib (1993) proposed a very widely used class of data augmentation Gibbs sampling algorithms for probit models. In such settings, one typically lets $a_0 = -\infty$ and $a_1 = 0$, while estimating the remaining $k - 2$ thresholds, with k denoting the number of levels of the categorical variable. A number of authors have relaxed the assumption of the probit link function through the use of nonparametric mixing. See for example Kottas et al. (2005), Jara et al. (2007) and Gill and Casella (2009).

In our case we fix *a priori* a sequence of thresholds relying on flexibility in nonparametric modeling of f to induce a flexible prior on p. In order to assign a prior Π on the space of count distributions, it is sufficient under this formulation to specify a prior Π^* on the space $\mathcal{L}$ of densities with respect to Lesbesgue measure on $\mathcal{Y}$.

3.1.2 Large support and posterior consistency

The involved mapping functions are both surjective and hence the inverse mapping $g^{-1}(\cdot)$ of a point $p \in \mathcal{C}$, where $\mathcal{C}$ is the space of the probability mass functions on $\mathbb{N}$ will correspond to an uncountably infinite set of densities in $\mathcal{L}$. Similarly, the inverse mapping $h^{-1}(\cdot)$ of a point $y \in \mathbb{N}$ will correspond to a subset of $\mathcal{Y}$ containing infinitely many y^*s. The existence of at least one element in $\mathcal{L}$ for every $p \in \mathcal{C}$ is ensured by the following lemma.

Lemma 3.1. *For every count measure $p_0 \in \mathcal{C}$ and rounding function $g(\cdot)$ defined in* (3.1)*, there exists at least one $f_0 \in \mathcal{L}$ such that $g(f_0) = p_0$.*

Proof. The lemma is trivially proved by defining f_0 as a step function of the form

$$f_0(x) = \frac{p_0(0)}{a_1 - b}\mathbb{1}_{[b,a_1)}(x) + \sum_{h=1}^{\infty} \frac{p_0(h)}{a_{h+1} - a_h}\mathbb{1}_{[a_h,a_{h+1})}(x),$$

where $\mathbb{1}_A(x)$ is 1 iff $x \in A$ and b is an arbitrary number such that (b, a_1) is in the domain of f. □

Lemma 3.2 demonstrates that the mapping $g : \mathcal{L} \to \mathcal{C}$ maintains Kullback-Leibler neighborhoods. As it is formalized in Theorem 3.3, this property

implies that the induced prior $p \sim \Pi$ assigns positive probability to all Kullback-Leibler neighborhoods of any $p_0 \in \mathcal{C}$ if at least one element of the set $g^{-1}(p_0)$ is in the KL support of the prior Π^*. By using conditions of Wu and Ghosal (2008), the KL condition becomes straightforward to demonstrate for a broad class of kernel mixture priors Π^*.

Lemma 3.2. *Assume that the true density of a count random variable is p_0 and choose any f_0 such that $p_0 = g(f_0)$. Let $\mathcal{K}_\epsilon(f_0) = \{f : KL(f_0, f) < \epsilon\}$ be a Kullback-Leibler neighbourhood of size ϵ around f_0. Then the image $g(\mathcal{K}_\epsilon(f_0))$ contains values $p \in \mathcal{C}$ in a Kullback-Leibler neighbourhood of p_0 of at most size ϵ.*

Proof. Let f a general element of $\mathcal{K}_\epsilon(f_0)$ and denote $p = g(f)$ its image on $\mathcal{C}$, hence

$$KL(f_0, f) = \int_{a_0}^{a_\infty} f_0(x) \log\left(\frac{f_0(x)}{f(x)}\right) \mathrm{d}x < \epsilon. \tag{3.2}$$

If we discretize the integral (3.2) in the infinite sum of integrals on disjoint subset of the domain of f we have

$$\sum_{h=0}^{\infty} \int_{a_h}^{a_{h+1}} f_0(t) \log\left(\frac{f_0(t)}{f(t)}\right) \mathrm{d}t < \epsilon.$$

Using the condition (see Theorem 1.1 of Ghurye (1968))

$$\int_A g_1(t)\mathrm{d}t \times \log\left(\frac{\int_A g_1(t)\mathrm{d}t}{\int_A g_2(t)\mathrm{d}t}\right) \leq \int_A g_1(t) \log\left(\frac{g_1(t)}{g_2(t)}\right) \mathrm{d}t$$

for each $A \in \mathcal{A}$, countable family of disjoint measurable sets of $\mathcal{Y}$ and $g_1, g_2 \in \mathcal{L}$, we get

$$p_0(j) \log \frac{p_0(j)}{p(j)} \leq \int_{a_j}^{a_{j+1}} f_0(t) \log\left(\frac{f_0(t)}{f(t)}\right) \mathrm{d}t$$

and hence

$$\sum_{j=0}^{\infty} p_0(j) \log \frac{p_0(j)}{p(j)} \leq \int_{a_0}^{a_\infty} f_0(x) \log\left(\frac{f_0(x)}{f(x)}\right) \mathrm{d}x < \epsilon,$$

that gives the result. □

Theorem 3.3. *Given a prior Π^* on $\mathcal{L}_{\Pi^*} \subseteq \mathcal{L}$ such that all $f \in \mathcal{L}_{\Pi^*}$ are in the Kullback-Leibler support of Π^*, then all $p \in \mathcal{C}_\Pi = g(\mathcal{L}_{\Pi^*})$ are in the Kullback-Leibler support of Π.*

Proof of Theorem 3.3. For every $f \in \mathcal{L}_{\Pi^*}$ by Lemma 3.2 we have

$$\Pi(\mathcal{K}_\epsilon(p)) \geq \Pi(g(\mathcal{K}_\epsilon(f))) = \Pi^*(\mathcal{K}_\epsilon(f)) > 0.$$

□

Corollary 3.4. *From Schwartz Theorem 2.2, the posterior probability of any weak neighborhood around the true data-generating distribution $p_0 \in \mathcal{C}_\Pi$ converges to one exponentially fast as $n \to \infty$.*

Theorem 3.5 using the fact that in $\mathcal{C}$ the weak and strong topology are topological equivalent, states that weak consistency implies L_1 consistency. The Kullback-Leibler property results hence to be sufficient also for strong consistency.

Theorem 3.5. *Given a prior $p \sim \Pi$ for a probability mass function $p \in \mathcal{C}$, if the posterior $\Pi(\cdot \,|\, y_1, \ldots, y_n)$ is weakly consistent, then it is also strongly consistent in the L_1 sense.*

Proof. In $\mathcal{C}$ weak convergence of sequences implies pointwise convergence by definition. In addition, Schur's property holds in $\mathcal{C}$ and hence weak convergence of sequences implies also strong convergence. Weak and strong metrics are hence topologically equivalent since $p_n \to p$ weakly iff $p_n \to p$ in L_1. Topologically equivalent metrics generate the same topology and this implies that the balls nest, i.e. that for any $p \in \mathcal{C}$ and radius $r > 0$, there exist positive radii r_1 and r_2 such that

$$S_{r_1}(p) \subseteq W_r(p) \quad \text{and} \quad W_{r_2}(p) \subseteq S_r(p)$$

where $S_r(p)$ and $W_r(p)$ are respectively strong and weak open neighborhoods of p of radius r. It follows that for any L_1 neigborhood S there exists a weak neighborhood W such that $S^C \subseteq W^C$. Hence the posterior probability of S^C is

$$\Pi(S^C \,|\, y_1, \ldots, y_n) \leq \Pi(W^C \,|\, y_1, \ldots, y_n).$$

Since the right hand side of the last equation goes to zero with P_{p_0}-probability 1, it follows that also

$$\Pi(S_r^C \,|\, y_1, \ldots, y_n) \to 0$$

with P_{p_0}-probability 1 and this concludes the proof. □

3.1.3 Rounded mixture of Gaussian

Mixtures of Gaussians is the leading approach in Bayesian continuous density estimation. It seems hence reasonable to adopt a modification of this approach and hence let

$$f(y^*; P) = \int N(y^*; \mu, \tau^{-1}) dP(\mu, \tau), \quad P \sim DP(\alpha P_0), \tag{3.3}$$

where $N(\cdot; \mu, \tau^{-1})$ is a normal kernel having mean μ and precision τ and P_0 chosen to be Normal-Gamma. Let Π^* denote the prior on f induced through

(3.3) and let Π denote the resulting prior on p induced through (3.1) with the thresholds chosen as $a_0 = -\infty$ and $a_j = j - 1$ for $j \in \{1, 2, \dots\}$. Let F the cumulative distribution function of f.

Note that the mixture model prior described satify the condition on the KL support requested in Wu and Ghosal (2008). Hence large support, weak and strong posterior consistency follows under the theory of Section 3.1.2.

In the following we first state some results on how to elicit prior knowledge about the random p and we then introduce a Gibbs sampling algorithm for posterior computation.

Eliciting the thresholds

When prior informations for p are available we can center p on an arbitrary probability mass function q simply by moving around the thresholds. The expectation of p can in fact be easily computed. Clearly

$$\mathrm{E}\{p(j)\} = \mathrm{E}\left\{\int_{a_j}^{a_{j+1}} f(y^*; P)dy^*\right\} = \mathrm{E}\{F(a_{j+1})\} - \mathrm{E}\{F(a_j)\}.$$

Marginalizing out prior $P \sim DP(\alpha P_0)$ with $P_0 = N(\mu; \mu_0, \kappa\tau^{-1})Ga(\tau; \nu/2, \nu/2)$, using Fubini's theorem we get

$$\mathrm{E}\big\{F(a_j)\big\} = \int F(a_j; P)dDP(P; \alpha P_0) = \int \int_{-\infty}^{a_j} f(y^*; P)dy^*\, dDP(P; \alpha P_0).$$

Hence

$$\begin{aligned}\mathrm{E}\{F(a_j)\} &= \mathrm{E}\left\{\sum_{h=1}^{\infty} \pi_h \Phi(a_j; \mu_h, \tau_h^{-1})\right\} \\ &= \int_{R\times R^+} \Phi(a_j; \mu, \tau^{-1})N(\mu; \mu_0, \tau^{-1}\kappa)Ga(\tau; \nu/2, \nu/2)\, d\mu\, d\tau \\ &= \int_0^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{a_j} N(y^*; \mu, \tau^{-1})N(\mu; \mu_0, \tau^{-1}\kappa)Ga(\tau; \nu/2, \nu/2)\, d\mu\, d\tau\, dy^*.\end{aligned} \tag{3.4}$$

Marginalizing out μ from (3.4) we get

$$\mathrm{E}\left\{F(a_j)\right\} = \int_{-\infty}^{a_j}\int_0^{\infty} N(y^*; \mu_0, (\kappa+1)/\tau)Ga(\tau; \nu/2, \nu/2)\, d\tau\, dy^*.$$

while marginalizing out τ we obtain

$$\mathrm{E}\left\{F(a_j)\right\} = \int_{-\infty}^{a_j} t_\nu(y^*; \mu_0, \kappa+1)dy^*$$

that gives

$$\mathrm{E}\{p(j)\} = \mathcal{T}_\nu(a_{j+1}; \mu_0, \kappa+1) - \mathcal{T}_\nu(a_j; \mu_0, \kappa+1) \tag{3.5}$$

where $\mathcal{T}_\nu(\cdot;\xi,\omega)$ is the cdf of a non central Student-t distribution with ν degrees of freedom, location ξ and scale ω. Hence, the expected probability of $y=j$ is simply a difference in t cdfs having ν degrees of freedom, mean μ_0, and scale $\kappa+1$. Setting $\mu_0=0$ and $\kappa=1$ for identifiability, the prior for p can be centered to have expectation exactly equal to an arbitrary pmf q; a simple iterative algorithm for choosing $\mathbf{a}$ to enforce $\mathrm{E}\{p(j)\}=q(j)$, for $j=0,1,\ldots$ is:

$$\begin{aligned} a_0 &= -\infty \\ a_1 &= \mathcal{T}_\nu^{-1}(q(0);0,2) \\ &\cdots \\ a_j &= \mathcal{T}_\nu^{-1}(\sum_{h=0}^{j-1} q(h);0,2). \end{aligned}$$

Although we can conceptually define an infinite sequence of thresholds, practically it is sufficient to define $\mathrm{E}\{p(j)\}=q(j)$, for $j=0,1,\ldots,J$ with $\sum_{j=0}^{J} q(j)=1-\epsilon$ and let the remaining a_j for $j=J+1,\ldots$ to be equispaced with unit step.

Given $\mathbf{a}$, the prior variance of $p(j)$ can be computed along similar lines. Let $F_D(a,b)=F(b)-F(a)$, $\Phi(a;\xi,\omega)$ the cumulative distribution function of a normal with mean ξ and variance ω, $\Phi_D(a,b;\xi,\omega)=\Phi(b;\xi,\omega)-\Phi(a;\xi,\omega)$ and $\mathcal{T}_{D,\nu}(a,b;\xi,\omega)=\mathcal{T}_\nu(b;\xi,\omega)-\mathcal{T}_\nu(a;\xi,\omega)$,

$$\begin{aligned} \mathrm{Var}\{p(j)\} &= \mathrm{Var}\{F_D(a_j,a_{j+1})\} \\ &= \mathrm{E}\{F_D(a_j,a_{j+1})^2\} - \mathrm{E}\{F_D(a_j,a_{j+1})\}^2. \end{aligned} \tag{3.6}$$

The second moment of $F_D(a_j,a_{j+1})$ can be derived as

$$\begin{aligned} \mathrm{E}\left\{F_D(a_j,a_{j+1})^2\right\} =& \mathrm{E}\left\{\left(\sum_{h=1}^{\infty}\pi_h\Phi_D(a_j,a_{j+1};\mu_h,\tau_h^{-1})\right)^2\right\} \\ =& \sum_{h=1}^{\infty}\mathrm{E}\left\{\left(\pi_h\Phi_D(a_j,a_{j+1};\mu_h,\tau_h^{-1})\right)^2\right\} + \\ &+ 2\sum_{k\neq l}\mathrm{E}\left\{\pi_k\pi_l\Phi_D(a_j,a_{j+1};\mu_k,\tau_k^{-1})\Phi_D(a_j,a_{j+1};\mu_l,\tau_l^{-1})\right\} \\ =& \sum_{h=1}^{\infty}\mathrm{E}\left\{\pi_h^2\right\}\mathrm{E}\left\{\Phi_D(a_j,a_{j+1};\mu_h,\tau_h^{-1})^2\right\} + \\ &+ 2\sum_{k\neq l}\mathrm{E}\left\{\pi_k\pi_l\right\}\mathrm{E}\left\{\Phi_D(a_j,a_{j+1};\mu_k,\tau_k^{-1})\Phi_D(a_j,a_{j+1};\mu_l,\tau_l^{-1})\right\}. \end{aligned}$$

Using the stick-breaking construction of the π_h and the results on the variance of the beta distribution we have

$$\mathrm{E}\{F_D(a_j,a_{j+1})^2\} = \frac{1}{\alpha+1}\mathrm{E}\left\{\Phi_D(a_j,a_{j+1};\mu,\tau^{-1})^2\right\} + \frac{\alpha}{\alpha+1}\mathrm{E}\left\{\Phi_D(a_j,a_{j+1};\mu,\tau^{-1})\right\}^2$$

where the expectations are with respect to $(\mu, \tau) \sim P_0$. This leads to

$$\begin{aligned}\mathrm{Var}\{p(j)\} &= \frac{1}{\alpha+1}\left[\mathrm{E}\left\{\Phi_D(a_j, a_{j+1}; \mu, \tau^{-1})^2\right\} - \mathrm{E}\left\{\Phi_D(a_j, a_{j+1}; \mu, \tau^{-1})\right\}^2\right] \\ &= \frac{1}{\alpha+1}\left[\mathrm{E}\left\{\Phi_D(a_j, a_{j+1}; \mu, \tau^{-1})^2\right\} - \left\{\mathcal{T}_{D,\nu}(a_j, a_{j+1}; \mu_0, \kappa+1)\right\}^2\right].\end{aligned}$$

The expected value of the squared normal cdf is with respect to P_0 and can be computed numerically.

From empirical evidence the method is enough robust to the prior specification to make this prior elicitation more a nice algebraical result rather than a practical used procedure. As it will be clear from the simulation study of Section 3.4.1 the DP mixture of rounded Gaussians is flexible enough to pragmatically let $a_0 = -\infty$ and $a_j = j - 1$.

A Gibbs sampling algorithm

From the computational point of view we can rely on existing results adapting any existing MCMC algorithm developed for DPMs of Gaussians. We focus here on the blocked Gibbs sampler of Ishwaran and James (2001), with $f(y^*) = \sum_{h=1}^{N} \pi_h N(y^*; \mu_h, \tau_h^{-1})$ with $\pi_1 = V_1$, $\pi_h = V_h \prod_{l<h}(1 - V_l)$, V_h independent Beta(1,α) and $V_N = 1$. See Walker (2007) and Yau et al. (2010) for specifications that avoid the truncation. The blocked Gibbs sampling steps are reported in Algorithm 1. The algorithm has been implemented in a R function with the core code written in C.

3.1.4 Rounded mixture of skew normal

Motivated by the analysis of telecommunications count data, which are often positive skewed and concentrated near zero, we introduce in this section another choice for the kernel: the skew normal (Azzalini, 1985). The skew normal (SN) distribution is a flexible distribution allowing several degrees of skewness. We aspect that using a flexible kernel within a flexible mixture model will lead to a lower effective number of cluster components.

The SN is a family of distributions which generalize the Gaussian adding a third parameter ruling out the shape of the distribution. In the last two decades the skew normal distribution has been widely studied in the statistical literature. See Azzalini (2005) for a comprehensive theoretical and applied review. Skew normal mixture models, that we will investigate here, are studied in Lin et al. (2007) and more recently in Frühwirth-Shnatter and Pyne (2010). Under a Bayesian nonparametric setup the master thesis of Cavatti Vieira (2011) gives some interesting results on continuous density estimation using the DP mixture of skew normal.

To give some notation let X be distributed as a skew normal with location ξ, scale ω and shape λ, written $X \sim SN(\xi, \omega, \lambda)$. The density function

Algorithm 1 Gibbs sampling algorithm: rounded mixture of Gaussians

Step 1: Generate each y_i^* from the full conditional posterior

- Generate $u_i \sim U\Big(\Phi(a_{y_i}; \mu_{S_i}, \tau_{S_i}^{-1}), \Phi(a_{y_i+1}; \mu_{S_i}, \tau_{S_i}^{-1})\Big)$
- Let $y_i^* = \Phi^{-1}(u_i; \mu_{S_i}, \tau_{S_i}^{-1})$

Step 2: Update S_i from its multinomial conditional posterior with

$$\Pr(S_i = h|-) = \frac{\pi_h p(y_i|\mu_h, \tau_h^{-1})}{\sum_{l=1}^{N} \pi_l p(y_i|\mu_l, \tau_l^{-1})},$$

where $p(j|\mu_h, \tau_h^{-1}) = \Phi(a_{j+1}|\mu_h, \tau_h^{-1}) - \Phi(a_j|\mu_h, \tau_h^{-1})$.
Step 4: Update the stick-breaking weights using

$$V_h \sim \text{Be}\left(1 + n_h, \alpha + \sum_{l=h+1}^{N} n_l\right)$$

Step 4: Update (μ_h, τ_h) from its conditional posterior

$$(\mu_h, \tau_h^{-1}) \sim \text{N}(\hat{\mu_h}, \hat{\kappa}_h \tau_h^{-1})\text{Ga}(\hat{a}_{\tau_h}, \hat{b}_{\tau_h})$$

with $\hat{a}_{\tau_h} = a_\tau + n_h/2$, $\hat{b}_{\tau_h} = b_\tau + 1/2(\sum_{i:S_i=h}(y_i^* - \bar{y}_h^*) + n_h/(1+\kappa n_h)(\bar{y}_h^* - \mu_0)^2)$, $\hat{\kappa}_h = (\kappa^{-1} + n_h)^{-1}$ and $\hat{\mu}_h = \hat{\kappa}_h(\kappa^{-1}\mu_0 + n_h\bar{y}_h^*)$.

of X is

$$SN(X;\xi,\omega,\lambda) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\lambda\frac{x-\xi}{\omega}\right), \tag{3.7}$$

where $\phi(x)$ is the density function of a standard normal and $\Phi(\cdot)$ is the distribution function of a standard normal, $\xi \in \mathbb{R}$, $\omega \in \mathbb{R}^+$ and $\lambda \in \mathbb{R}$. Note that for $\lambda = 0$ the density reduces to the normal $N(x;\xi,\omega^2)$.

Several extensions and alternative formulations have been proposed in connection with the skew normal distribution. A commendable work of unification of such proposal is carried out in Arellano-Valle and Azzalini (2006) under the general formulation known as unified skew normal and denoted with the acronym SUN. A SUN distribution has density

$$SUN_{d,m}(X;\xi,\gamma,\bar{\omega},\Omega^*) = \phi_d(y-\xi;\Omega)\frac{\Phi_m(\gamma+\Delta^{\mathrm{T}}\bar{\Omega}^{-1}\omega^{-1}(y-\xi);\Gamma-\Delta^{\mathrm{T}}\bar{\Omega}^{-1}\Delta)}{\Phi_m(\gamma;\Gamma)}, \tag{3.8}$$

where $\bar{\omega} = \omega 1_d$ and Ω^* can be partitioned as

$$\Omega^* = \begin{pmatrix}\Gamma & \Delta^{\mathrm{T}}\\ \Delta & \bar{\Omega}\end{pmatrix}.$$

Among the many generation mechanisms of model (3.7), we use the convolution Lemma of Azzalini (1986). Denoted by $HN(\omega^2)$ the half normal distribution with scale ω^2 (and variance $\omega^2(1-2/\pi)$), we recall the following lemma.

Lemma 3.6 (Azzalini)**.** *Let $\eta \sim HN(\sigma^2)$, $U \sim N(0,\sigma^2)$ with U independent from η and $|\delta| < 1$. If*

$$X = \delta\eta + \sqrt{1-\delta^2}U$$

then $X \sim SN(0,\sigma,\lambda)$ with $\lambda = \delta/\sqrt{1-\delta^2}$.

Lemma 3.6 allows us to treat a skew normal variable $X \sim SN(\xi,\omega,\lambda)$ with a hierarchical representation in which, conditionally on a realization η from a half normal distribution, X is normal with mean $\xi+\delta\eta$ and variance $(1-\delta^2)\omega^2$.

Going back to our rounded mixture prior we let

$$f(y^*;P) = \int SN(y^*;\xi,\omega,\lambda)dP(\xi,\omega,\lambda), \quad P \sim DP(\alpha P_0) \tag{3.9}$$

where $\alpha > 0$ and

$$P_0(\xi,\omega,\lambda) = N(\xi;\xi_0,\kappa\omega^2) \times \mathrm{Ga}(\omega^{-2};a,b) \times N(\lambda;0,\psi). \tag{3.10}$$

Equation (3.10) follows the lines of the expression for the mixing measure of model (3.3). In the DPM of Gaussians, in fact, P_0 was a normal-inverse-gamma following the standard convenient choice for the prior of the parameters of a single Gaussian distribution. Unfortunately, such a golden standard

for the skew normal does not exist, particularly for the prior on the shape parameter, and several proposals have been made in this direction. An objective Bayes viewpoint is studied in Liseo and Loperfido (2006) introducing the Jeffrey's reference prior for the shape parameter, showing that it has unbounded support and that it is proper. Arellano-Valle et al. (2009) propose a skew normal distribution also as prior showing conjugancy within the SUN family of Arellano-Valle and Azzalini (2006). Even if our proposal is similar to that of Cavatti Vieira (2011), it was developed independently. In addition our sampling method, discussed below, is new not only concerning the data augmentation step that relates our count observations with the latent continuous variables. It is hence of self interest also for continuous density estimation using skew normal mixtures.

A Gibbs sampling algorithm

For posterior computation we assume $f(y^*) = \sum_{h=1}^{N} \pi_h SN(y^*; \xi_h, \omega_h, \lambda_h)$ with $\pi_1 = V_1$, $\pi_h = V_h \prod_{l<h}(1 - V_l)$, V_h independent Beta(1,α) and $V_N = 1$. In addition we will use the result of Lemma 3.6 to gain conjugacy for the location and scale parameters of each component of the mixture. The distributions for the shape parameters are in closed form and belong to the SUN class (3.8). The full conditional posterior distributions are specified in Algorithm 2.

The seventh step of Algorithm 2, in particular, can be done in closed form, using the result of the following lemma that is a particular case of the situation described in Section 2.1 of Arellano-Valle and Azzalini (2006).

Lemma 3.7 (Arellano-Valle-Azzalini)**.** *Let* $V_0 \sim N_q(0, I_q(1-\Delta^T\Delta)+\Delta\Delta^T)$, $V_1 \sim N(0,1)$ *with* V_0 *independent from* V_1*. If*

$$Y = \Delta|V_0| + \sqrt{1 - \Delta^T\Delta}V_1$$

then $Y \sim SUN_{1,q}(0, 0, 1, \Omega)$ *with*

$$\Omega = \begin{pmatrix} I_{n_q}(1 - \Delta^T\Delta) + \Delta\Delta^T & \Delta \\ \Delta^T & 1 \end{pmatrix}.$$

Cavatti Vieira (2011) proposed similar prior distributions while not relying on the stochastic representation discussed in Lemma 3.6. This leads the full conditional posterior for ω_h to be not in closed form and hence to perform the MCMC simulation from the posterior using a Metropolis-Hastings algorithm that is clearly computationally less efficient than our Gibbs sampler.

3.2 Multivariate rounded kernel mixtures prior

Multivariate count data are quite common in a broad class of disciplines and particularly in the setting introduced in Section 2.1.1.

Algorithm 2 Gibbs sampling algorithm: rounded mixture of skew normals

Step 1: Generate each y_i^* from the underline continuous distribution under the constraints $a_{y_i} < y_i^* < a_{y_i+1}$;

Step 2: Sample S_i, the class indicator from the multinomial

$$\Pr(S_i = h|-) = \frac{\pi_h p(y_i|\xi_h, \omega_h, \lambda_h)}{\sum_{l=1}^{N} \pi_l p(y_i|\xi_l, \omega_l, \lambda_l)}$$

with $h = 1, \ldots, H$ and H the number of occupied clusters.

Step 3: Sample α using Escobar and West (1995) given n and H, the number of occupied clusters

Step 4: Update the stick-breaking weights using

$$V_h \sim \text{Be}\left(1 + n_h, \alpha + \sum_{l=h+1}^{H} n_l\right)$$

where n_h is the sample size of the hth cluster.

Step 5: Update

$$\eta_i \sim N(\delta_{S_i}(y_i^* - \xi_{S_i}), \omega_{S_i}^2(1 - \delta_{S_i}^2))$$

where δ_h is $\lambda_h/\sqrt{\lambda_h^2 + 1}$.

Step 6: Sample (ξ_h, ω_h) from

$$N\left(\hat{\mu}_h, \hat{\kappa}_h\omega_h^2\right) \text{InvGam}(a + n_h/2 + 1, b + \hat{b}_h)$$

where

$$\hat{\mu}_h = \frac{\kappa \sum_{S_i=h}(y_i^* - \delta_h\eta_i) + (1 - \delta_h^2)\xi_0}{n_h + \kappa\omega^2(1 - \delta_h^2)}, \quad \hat{\kappa}_h = \frac{\kappa(1 - \delta_h^2)}{n_h\kappa + (1 - \delta_h^2)},$$

$$\hat{b}_h = \frac{\sum_{S_i=h} \eta_i^2 - 2\delta_h \sum_{S_i=h} \eta_i(y_i^* - \xi_h) + \sum_{S_i=h}(y_i^* - \xi_h)^2 + (1 - \delta_h^2)(\xi_h - \xi_0)^2}{2(1 - \delta_h^2)}.$$

Step 7: Sample λ_h from

$$\lambda_h \sim SUN_{1,n_h}(\lambda_h; 0, 0, \phi, \Omega_h)$$

where $z_i^* = \sqrt{\phi}(y_i^* - \xi_h)/\omega_h$, z_h is the vector of size n_h with all z_i^* belonging to cluster h, $\Delta_h = z_h^*(1 + z_h^{*T} z_h^*)^{-1/2}$ and

$$\Omega_h = \begin{pmatrix} I_{n_h}(1 - \Delta_h^T\Delta_h) + \Delta_h\Delta_h^T & \Delta_h \\ \Delta_h^T & 1 \end{pmatrix}.$$

Our approach easily generalizes in a multivariate setting. The multivariate rounded kernel mixture prior introduced below can flexibly characterize the entire joint distribution including the marginals and dependence structure, while leading to straightforward and efficient computation. For ease of computation in the multivariate case we focus in using underlying multivariate Gaussian kernels.

3.2.1 Properties in the multivariate context

The rounding idea of Section 3.1.1 can be easily generalized into its multivariate counterpart. Assume that the multivariate count vector $\mathbf{y} = (y_1, \ldots, y_p)$ is the transformation through a threshold mapping function h of a latent continuous vector $\mathbf{y}^* = (y_1^* \ldots, y_p^*)$. In a general setting we have

$$\mathbf{y} = h(\mathbf{y}^*) \tag{3.11}$$

$$\mathbf{y}^* \sim f(\mathbf{y}^*) = \int \mathrm{K}_p(\mathbf{y}^*; \theta, \Omega) d\mathrm{P}(\theta, \Omega), \quad P \sim \tilde{\Pi}, \tag{3.12}$$

where $\mathrm{K}_p(\cdot; \theta, \Omega)$ is a p-variate kernel with location θ and scale-association $p \times p$ matrix Ω and $\tilde{\Pi}$ is a prior over the space of probability measures over $\mathbb{R}^p \times M_p$ with M_p the space of positive definite $p \times p$ matrices. The mapping $h(\mathbf{y}^*) = \mathbf{y}$ implies that the probability mass function p of $\mathbf{y}$ is

$$p(y_1 = J_1, \ldots, y_p = J_p) = p(\mathbf{J}) = g(f)[\mathbf{J}] = \int_{A_J} f(\mathbf{y}^*) d\mathbf{y}^* \quad \mathbf{J} \in \mathbb{N}^p \tag{3.13}$$

where $A_J = \{\mathbf{y}^* : a_{1,J_1} \leq y_1^* < a_{1,J_1+1}, \ldots, a_{p,J_p} \leq y_p^* < a_{p,J_p+1}\}$ defines a disjoint partition of the sample space. Marginally this formulation is the same of that in (3.1).

Remark 3.1. *Lemma 3.2 and Theorem 3.3 demonstrate that in the univariate case the mapping $g : \mathcal{L} \to \mathcal{C}$ maintains Kullback-Leibler neighborhoods and hence the induced prior Π assigns positive probability to all Kullback-Leibler neighbourhoods of any $p_0 \in \mathcal{C}$. This property holds also in the multivariate case.*

The true p_0 is in the KL support of our prior, and hence we obtain weak and strong posterior consistency following the theory of Section 3.1.1, as long as there exists at least one multivariate continuous density $f_0 = g^{-1}(p_0)$ that falls in the KL support of the mixture prior for f described in (3.11)-(3.12).

Also in the multivariate context we could rely on existing results on large support and posterior consistency developed for multivariate continuous density estimation. Unfortunately results for multivariate density estimations are scarce. The only available results on asymptotic properties of Bayesian procedures for multivariate continuous density estimation are presented by Ghosal and co-authors (Wu and Ghosal, 2010; Shen and Ghosal,

2011). In both papers the models considered are quite limited in scope in focusing on DP location mixtures of Gaussian kernels while assigning an independent prior for the covariance matrix. In the following theorem we assume that K_p corresponds to a multivariate Gaussian kernel. It is in terms of mixture prior for multivariate continuous density estimation and gives sufficient condition on the true multivariate continuous f_0 to ensure that it falls within the KL support of the prior $f \sim \Pi^*$ induced by (3.12). It modifies Theorem 2 of Wu and Ghosal (2010).

Theorem 3.8. *Let f_0 be a density over $\mathbb{R}^p$ with respect to Lesbesgue measure and let Π^* denote the prior on f induced from (3.12) with K_p corresponding to a multivariate Gaussian kernel, $P \sim \tilde{\Pi}$ and $\tilde{\Pi}$ an arbitrary prior with weak support over the space of probability measures over $\mathbb{R}^p \times M_p$. Assume the following*

1. *$0 < f(\mathbf{y}^*) < M^*$ for some constant M^* and all $\mathbf{y}^* \in \mathbb{R}^p$;*
2. *$|\int f_0(\mathbf{y}^*) \log f_0(\mathbf{y}^*) d\mathbf{y}^*| < \infty$;*
3. *for some $\delta > 0$, $\int f_0(\mathbf{y}^*) \log \frac{f_0(\mathbf{y}^*)}{\phi_\delta(\mathbf{y}^*)} d\mathbf{y}^* < \infty$, where $\phi_\delta(\mathbf{y}^*) = \inf_{||\mathbf{y}^{*\prime} - \mathbf{y}^*|| < \delta} f_0(\mathbf{y}^*)$;*
4. *for some $\eta > 0$, $\int ||\mathbf{y}^*||^{2p(1+\eta)} f_0(\mathbf{y}^*) d\mathbf{y}^* < \infty$.*

Then f_0 is in the KL support of Π^*.

Proof. The proof follows Theorem 2 of Wu and Ghosal (2010) in first bounding the density of a multivariate normal density with general covariance matrix Σ by

$$\left(\frac{\lambda_1(\Sigma)}{\lambda_p(\Sigma)}\right)^{\frac{p-1}{2}} \phi(y; 0, \lambda_1(\Sigma) I_p) \leq \phi(y; 0, \Sigma) \leq \left(\frac{\lambda_d(\Sigma)}{\lambda_1(\Sigma)}\right)^{\frac{p-1}{2}} \phi(y; 0, \lambda_p(\Sigma) I_p),$$

and then showing thanks to this that, for P belonging to an open set of the space of probability measures over $\mathbb{R}^p \times M_p$ we have

$$\int f_0(y) \log \frac{f_0(y)}{\int \phi(y : \theta, \Sigma) \mathrm{d}P(\theta, \Sigma)} \mathrm{d}y \leq \epsilon.$$

□

This result will be also useful in Chapter 4 which generalizes the modeling framework discussed in this chapter to the mixed-scale data, i.e. when we observe a multivariate vector $\mathbf{y}$ containing both continuous and discrete (counts, categorical, binary) variables.

3.2.2 Multivariate rounded mixture of Gaussians

Continue to assume that K_p corresponds to a multivariate Gaussian kernel. Let furthermore $\tilde{\Pi}$ be DP(αP_0), with P_0 corresponding to a normal inverse-Wishart base measure. Under this formulation we can obtain the following Gibbs sampler for posterior computation. Also here, we rely on existing methods for posterior computation and in Algorithm 3 we report a modification of the Gibbs sampler with auxiliary parameters of Neal (2000).

Algorithm 3 Gibbs sampling algorithm: multivariate rounded mixture of Gaussians

Step 1: Generate each $\mathbf{y}_i^*$ from the full conditional posterior
for j in $1, \ldots, p$ **do**
Generate $u_{ij} \sim U\Big(\Phi(a_{y_{ij}-1}; \tilde{\mu}_{i,j}, \tilde{\sigma}^2_{i,j}), \Phi(a_{y_ij}; \tilde{\mu}_{i,j}, \tilde{\sigma}^2_{i,j})\Big)$, where

$$\tilde{\mu}_{i,j} = \mu_{S_i,j} + \Sigma_{S_i,12}\Sigma^{-1}_{S_i,22}(\mathbf{y}^*_{-j} - \mu_{S_i,-j})$$
$$\tilde{\sigma}^2_{i,j} = \sigma^2_{S_i,j} - \Sigma_{S_i,12}\Sigma^{-1}_{S_i,22}\Sigma_{S_i,21}$$

are the usual conditional expectation and conditional variance of the multivariate normal.
Let $y^*_{ij} = \Phi^{-1}(u_{ij}; \tilde{\mu}_{i,j}, \tilde{\sigma}^2_{i,j})$
end for
Step 2:
for i in $1, \ldots, n$ **do**
Let k_i the number of occupied clusters
Draw $(\mu_{k_i+1}, \Sigma_{k_i+1}) \sim P_0$
Update S_i from

$$P(S_i = h) = \begin{cases} \frac{n_h}{n-1+\alpha} p(\mathbf{y}_i; \mu_h, \Sigma_h) & \text{for } h = 1, \ldots, k_i \\ \frac{\alpha}{n-1+\alpha} p(\mathbf{y}_i; \mu_{k_i+1}, \Sigma_{k_i+1}) \end{cases}$$

end for
Step 3: Update (μ_h, Σ_h) from their conditional posteriors.

3.3 Bayesian count curve fitting

Our modeling framework modifies the Bayesian curve fitting approach presented by Müller, Erkanli and West (1996) to accommodate count data.

Consider a standard nonparametric regression problem in which we want to estimate $s(x)$, a smooth trajectory for a response variable y given the observation of one or more explanatory variables x such that it minimizes some criterion. For example, if y is continuous one typically minimizes $E[\{y - s(x)\}^2 \mid x]$ obtaining $s(x) = E(y \mid x)$. If y is a count variable, a more

natural loss function is the absolute deviation loss, namely $E[|y - s(x)| \mid x]$ that leads to $s(x) = \text{Median}(y \mid x)$.

In both the situations if one can estimate the joint distribution of $z = (y, x)$, or better the conditional distribution of $(y \mid x)$, then $s(x)$ immediately follows. For sake of explanation assume henceforth that both y and x are univariate.

A lot of approaches deal with this problem under a nonparametric settings and generalizing the idea of local linear regression (see e.g. Hastie et al., 2001, for a general review of standard nonparametric smoothing approaches). In those settings the regression function $s(x)$ can be written with the form

$$s(x) = \sum_j w_j(x) m_j(x),$$

where w_js are weights assigned to m_js, some precise linear functions of x. Müller et al. (1996) proposed to view $s(x)$ as deriving from a joint distribution of a mixture model like

$$(x, y) \sim \sum_j w_j f_j(x, y),$$

where w_j are weights summing to one, $f_j(x, y)$ are probability density functions having $f(y \mid x)$ as conditional density function with means $m_j(x)$. More precisely they assumed a DPM of Gaussians prior for joint distribution of (y, x) and built a Gibbs sampling scheme for posterior computation. Under this settings and thanks to the closure under conditioning of the Gaussian family, predictive values of y can be generated in an MCMC chain and the predictive regression estimate can be obtained by averaging the simulated ys.

Our rounded DPM of Gaussians modifies this framework to accommodate count data. Assume to model jointly $\mathbf{y} = (y, x)$ as described in Section 3.2.2. Then, in Algorithm 3 it is sufficient to add a further step in which we generate under its conditional predictive distribution y_{n+1} given x, i.e.

$$(y^*_{n+1} \mid x) \sim w_0(x) f_0(y^* \mid x) + \sum_{j=1}^{k} w_j(x) f_j(y^* \mid x), \quad y_{n+1} = h(y^*_{n+1})$$

where f_0 is the conditional density of y^* given x based on the normalized base measure P_0, and f_j are the conditional Gaussian densities of y given x under the joint Gaussian distributions. The current number of occupied clusters in the DP mixture is denoted by k and the $k+1$ weights w_j $(j = 0, 1, \ldots, k)$ are functions of the marginal densities of x under the base measure P_0 and the the joint Gaussians. Taking the average or the median of the Montecarlo replicates of $(y_{n+1} \mid x)$, for a grid of values of x we can obtain different regression functions. This framework can be generalized in the case in which z is a mixed scale vector, having discrete and continuous variables. Some theoretical developments for the mixed scale case are presented in Chapter 4.

3.4 Simulation studies

To assess the performance of the proposed approach, in this section we report the results of some simulation studies.

3.4.1 Probability mass function estimation simulation

To test the performance of the rounded mixture of Gaussians (RMG) of Section 3.1.3 and of the rounded mixture of skew normal (RMSN) of Section 3.1.4 in the probability mass function estimation, we compared them with four different approaches: the empirical probability mass function (E), two Bayesian nonparametric approaches, with the first assuming a Dirichlet process prior with a Poisson base measure (DP) and the second using a Dirichlet process mixture of Poisson kernels (DPM-Pois), and lastly the maximum likelihood estimate under a Poisson model (MLE). Several simulations have been run under different simulation settings leading to qualitatively similar results. In what follows we report the results for the five following scenarios:

(a) $y = h(y^*)$, $y^* \sim 0.4N(25, 1.5)+0.15N(20, 1)+0.25N(24, 1)+0.2N(21, 2)$;

(b) $y \sim \text{Poi}(12)$;

(c) $y \sim 0.4\text{Poi}(1) + 0.25\text{Poi}(3) + 0.25\text{Poi}(5) + 0.1\text{Poi}(13)$;

(d) $y \sim \text{ConPoi}(30, 3)$, where with $\text{ConPoi}(\lambda, \nu)$ is the Conway-Maxwell-Poisson distribution (Shmueli et al., 2005);

(e) $y = h(y^*)$, $y^* \sim 0.6\text{Ga}(2, 0.5) + 0.4\text{Ga}(4, 1.5)$.

A plot for each of the five true probability mass functions is reported in Figure 3.1.

For each case, we generated samples of size $n = 10, 25, 50, 100, 300$. Each of the five analysis approaches were applied to R=1,000 replicated data sets under each scenario. The methods were compared based on a Monte Carlo approximation to the mean Bhattacharyya (1943) distance (BCD) and Kullback-Leibler divergence (KLD) calculated as

$$\text{BCD} = \frac{1}{R}\sum_{r=1}^{R}\left(\sum_{j=\max(0,\min(y)-B)}^{\max(y)+B} -\log\left(\sqrt{p(j)\hat{p}_r(j)}\right)\right),$$

$$\text{KLD} = \frac{1}{R}\sum_{r=1}^{R}\left(\sum_{j=\max(0,\min(y)-B)}^{\max(y)+B} p(j)\log\left(p(j)/\hat{p}_r(j)\right)\right).$$

where we take the sums across the range of the observed data $\pm$ a buffer of $B = 10$.

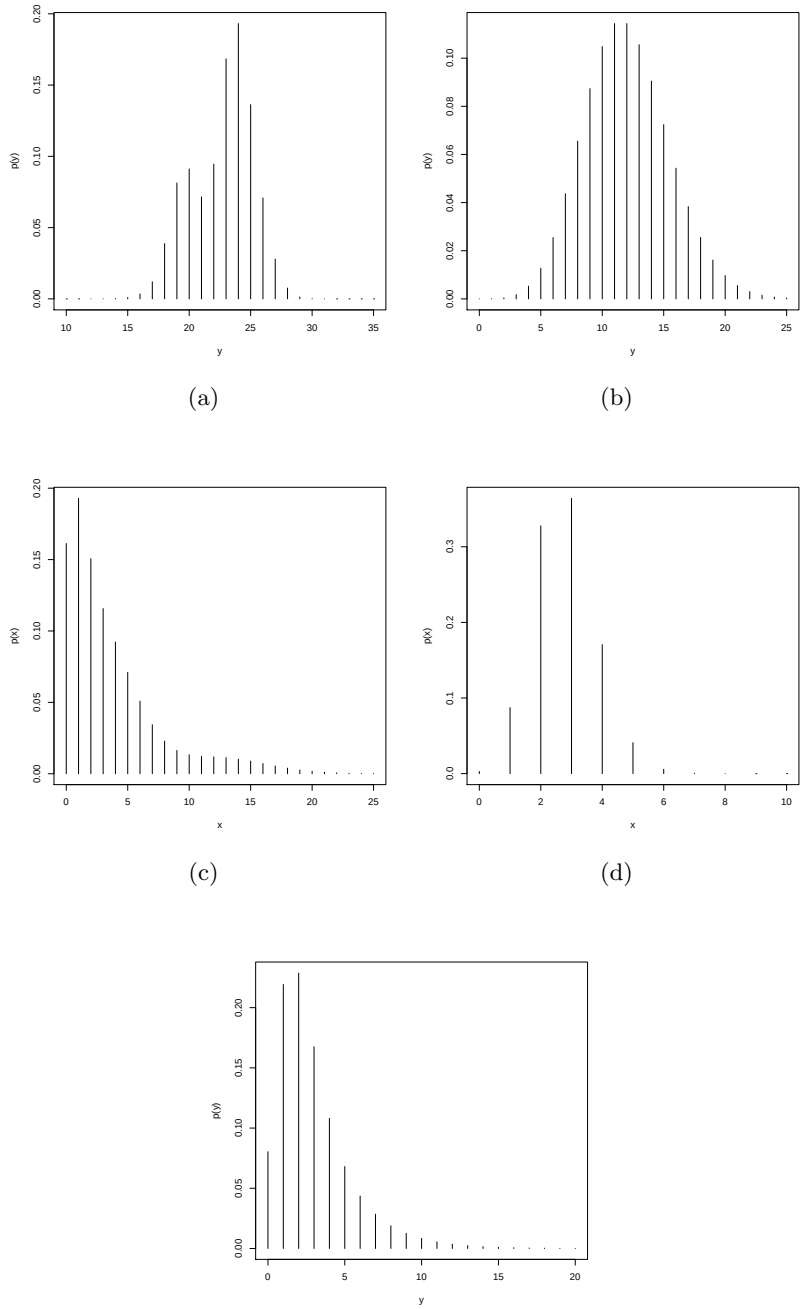


Figure 3.1: Real probability mass functions in the simulation study: rounded mixture of Gaussians (a), Poisson (b), mixture of Poissons (c), Conway-Maxwell Poisson (d) and rounded mixture of gammas (e).

The empirical coverage of 95% credible intervals for the $p(j)$s is also calculated and reported in Figures 3.2–3.6. These intervals were estimated as the 2.5th to 97.5th percentiles of the samples collected after burn-in for each $p(j)$, with a small buffer of $\pm 10^{-8}$ added to accommodate numerical approximation error.

In implementing the blocked Gibbs sampler for the rounded mixture of Gaussians, the first 1,000 iterations were discarded as a burn-in and the next 10,000 samples were used to calculate the posterior mean of $\hat{p}(j)$. For the hyperparameters, as a default empirical Bayes approach, we chose $\mu_0 = \overline{y}$, the sample mean, and $\kappa = s^2$, the sample variance, and $a_\tau = b_\tau = 1$. The precision parameter of the DP prior was set equal to one as a commonly used default and the truncation level N is set to be equal to the sample size of each sample. We also tried reasonable alternative choices of prior, such as placing a gamma hyperprior on the DP precision (Escobar and West, 1995), for smaller numbers of simulations and obtained similar results. The values of $p(j)$ for a wide variety of js were monitored to gauge rates of apparent convergence and mixing. The trace plots showed excellent mixing, and the Geweke (1992) diagnostic suggested very rapid convergence.

The DP approach used Poi($\overline{y}$) as the base measure, with $\alpha = 1$ or $\alpha \sim \text{Ga}(1, 1)$ considered as alternatives. For fixed α, the posterior is available in closed form, while for $\alpha \sim \text{Ga}(1, 1)$ we implemented a Metropolis-Hastings normal random walk to update $\log \alpha$, with the algorithm run for $10,000$ iterations with a $1,000$ iterations burn-in.

The blocked Gibbs sampler (Ishwaran and James, 2001) was used for posterior computation in the DPM-Pois model, with the first 1,000 iterations discarded as a burn-in and the next 10,000 samples used to calculate the posterior mean $\hat{p}(j)$. A gamma base measure with hyperparameters $a = b = 1$ was chosen within the DP while the precision parameter was fixed to $\alpha = 1$.

The results of the simulation are reported in Table 3.1. The proposed method performs generally better, in terms of BCD and KLD, than the other methods when the truth is underdispersed and clearly not Poisson, as in the first and last scenarios.

When we simulated Poisson data the MLE under a Poisson model and the DPM of Poissons performs slightly better than the proposed RMG and RMSN in very small samples. However, even in modest sample sizes of $n = 25$, the RMG approach was surprisingly competitive when the truth was Poisson. Interesting, when the truth was a mixture of Poissons as in scenario (c) we obtained much better performance for the RMG approach than the DPM-Pois model. The ∞ recorded for the empirical estimation is due to the presence of $p(j)$ exactly equal to zero if we do not observe any $y = j$.

The Gaussian kernel performs better than the skew normal one but such a difference tends to decrease as n grows. This behaviour was expected and

it is probably due to the fact that the skew normal pays its flexibility in terms of one additional parameter to be estimated. To better understand the differences between the RMG and the RMSN, we reported the posterior average number of occupied clusters in the mixture in Table 3.2. Expecially for high n and skew scenarios like (e), the average number of occupied clusters is smaller when using the skew normal. In fact in approximating an asymmetric density, a mixture of distributions that already allows different levels of skewness will be more efficient than a mixture of symmetric distributions, which will rather require a larger number of components to fit the asymmetric pattern. In practical applications, hence, the prior knowledge of the skewness of the probability mass function to be estimated, the available sample size and the necessity to interpret the model based clustering naturally obtained in the DPM, can help in choosing one kernel rather than another. In absence of strong prior information on the skewness, the Gaussian kernel is a robust and convenient choice.

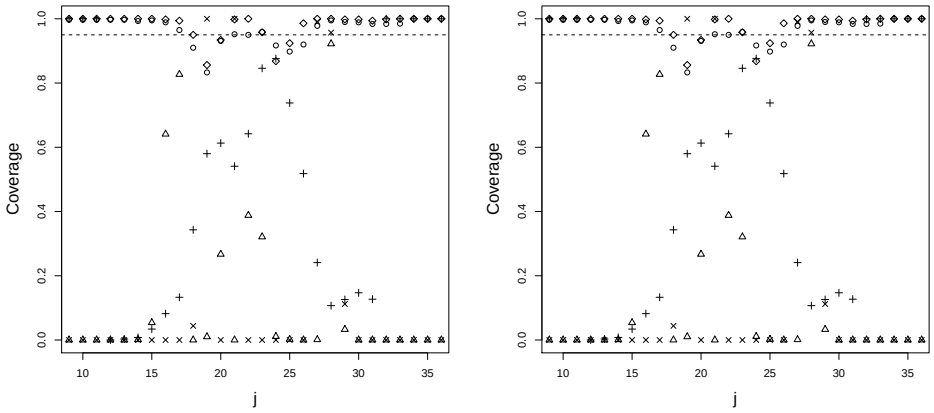
The effective coverage of the credible intervals for $p(j)$ for the RMG and the RMSN fluctuates around the nominal value for all the scenarios and sample sizes with a slighlty better performance of the RMG. Using the Dirichlet process prior we get an effective coverage that is either strongly less than the nominal levels, or much too high, due to too wide credible intervals. For DP-Pois, we obtain coverage close to the nominal level only at the values of j such that the true $p(j)$ is high enough so that substantial numbers of observations fall at that value.

Table 3.1: Bhattacharya coefficient and Kullback-Leibler divergence from the true distribution for samples from the five scenarios

		Scenario (a)		Scenario (b)		Scenario (c)		Scenario (d)		Scenario (e)	
n	Method	BCD	KLD	BCD	KLD	BCD	KLD	BCD	KLD	BCD	KLD
10	RMG	0.04	0.16	0.04	0.17	0.03	0.11	0.04	0.12	0.04	0.14
	RMSN	0.1	0.28	0.1	0.25	0.08	0.19	0.09	0.23	0.09	0.24
	E	0.24	∞	0.35	∞	0.39	∞	0.09	∞	0.24	∞
	$DP_{\alpha=1}$	0.14	0.68	0.19	0.9	0.16	1.14	0.06	0.25	0.13	0.74
	$DP_{\alpha\sim Ga(1,1)}$	0.11	0.47	0.11	0.49	0.12	0.87	0.05	0.22	0.1	0.54
	MLE	0.13	0.37	0.01	0.05	0.11	0.78	0.07	0.21	0.04	0.24
	DPM-Pois	0.26	0.69	0.09	0.29	0.15	0.43	0.26	0.67	0.15	0.42
25	RMG	0.02	0.08	0.02	0.08	0.02	0.06	0.02	0.06	0.02	0.08
	RMSN	0.05	0.13	0.04	0.11	0.04	0.1	0.04	0.1	0.03	0.11
	E	0.09	∞	0.14	∞	0.23	∞	0.03	∞	0.13	∞
	$DP_{\alpha=1}$	0.07	0.34	0.1	0.57	0.1	0.9	0.02	0.11	0.07	0.5
	$DP_{\alpha\sim Ga(1,1)}$	0.06	0.24	0.06	0.29	0.08	0.68	0.02	0.1	0.06	0.35
	MLE	0.13	0.36	0.01	0.02	0.11	0.76	0.06	0.2	0.03	0.18
	DPM-Pois	0.18	0.5	0.02	0.06	0.02	0.1	0.21	0.55	0.11	0.33
50	RMG	0.01	0.05	0.01	0.04	0.01	0.04	0.01	0.04	0.01	0.05
	RMSN	0.03	0.08	0.03	0.07	0.02	0.05	0.02	0.06	0.02	0.06
	E	0.04	∞	0.07	∞	0.17	∞	0.02	∞	0.09	∞
	$DP_{\alpha=1}$	0.03	0.16	0.06	0.33	0.06	0.69	0.01	0.06	0.04	0.34
	$DP_{\alpha\sim Ga(1,1)}$	0.03	0.12	0.04	0.18	0.05	0.54	0.01	0.05	0.04	0.25
	MLE	0.13	0.35	<0.01	0.01	0.11	0.75	0.06	0.19	0.03	0.16
	DPM-Pois	0.16	0.44	0.03	0.09	0.12	0.28	0.19	0.51	0.10	0.27
100	RMG	0.01	0.03	0.01	0.02	<0.01	0.02	0.01	0.03	0.01	0.03
	RMSN	0.02	0.06	0.02	0.05	0.01	0.03	0.02	0.05	0.01	0.03
	E	0.02	∞	0.03	∞	0.13	∞	0.01	∞	0.07	∞
	$DP_{\alpha=1}$	0.02	0.08	0.03	0.18	0.03	0.47	0.01	0.03	0.02	0.22
	$DP_{\alpha\sim Ga(1,1)}$	0.02	0.07	0.02	0.11	0.03	0.37	0.01	0.03	0.02	0.17
	MLE	0.13	0.35	<0.01	0.01	0.11	0.75	0.06	0.19	0.03	0.15
	DPM-Pois	0.54	1.41	<0.01	0.01	0.01	0.03	0.18	0.48	0.09	0.23
300	RMG	<0.01	0.01	<0.01	0.01	<0.01	0.01	<0.01	0.02	<0.01	0.01
	RMSN	0.01	0.03	0.01	0.02	<0.01	0.02	0.01	0.03	<0.01	0.01
	E	0.01	∞	0.01	∞	0.1	∞	<0.01	∞	0.05	∞
	$DP_{\alpha=1}$	0.01	0.03	0.01	0.07	0.01	0.17	0.26	2.29	0.01	0.18
	$DP_{\alpha\sim Ga(1,1)}$	0.05	0.29	0.01	0.35	0.05	0.74	0.03	0.16	0.01	0.15
	MLE	0.13	0.35	<0.01	<0.01	0.1	0.74	0.06	0.19	0.03	0.09
	DPM-Pois	0.14	0.4	0.01	0.05	0.1	0.21	0.17	0.47	0.06	0.13

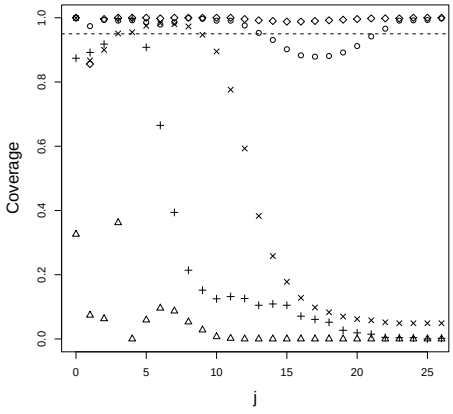
Table 3.2: Average number of occupied clusters in the RMG and RMSN

n	Method	Scenario (a)	Scenario (b)	Scenario (c)	Scenario (d)	Scenario (e)
10	RMSN	3.01	2.82	3.30	3.19	3.22
	RMG	3.05	2.85	3.34	3.23	3.26
25	RMSN	2.90	2.46	3.84	3.38	3.86
	RMG	3.20	2.72	4.24	3.73	4.26
50	RMSN	2.73	2.45	4.11	3.30	4.25
	RMG	3.23	2.90	4.87	3.91	5.03
100	RMSN	2.98	2.88	4.37	3.48	4.67
	RMG	3.71	3.59	5.44	4.34	5.82
300	RMSN	2.93	2.83	4.29	3.42	5.59
	RMG	3.71	3.59	5.44	4.34	7.09

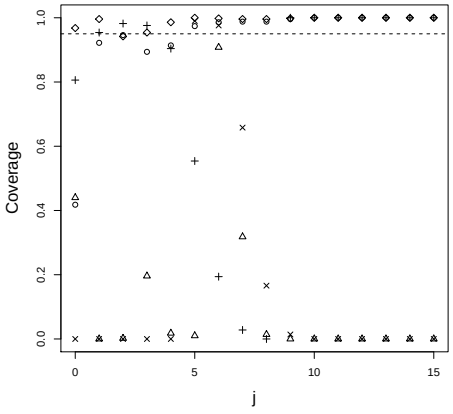


(a)

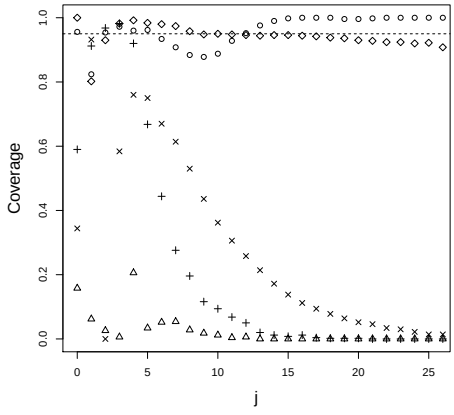
(b)



(c)



(d)



(e)

Figure 3.2: Coverage of 95% credible intervals for $p(j)$ under the four scenarios with $n = 10$. Points represent the RMG method, squares the RMSN, cross-shaped dots the DP with $\alpha = 1$, triangles the DP with $\alpha \sim Ga(1, 1)$ and x-shaped dots the DPM of Poisson.

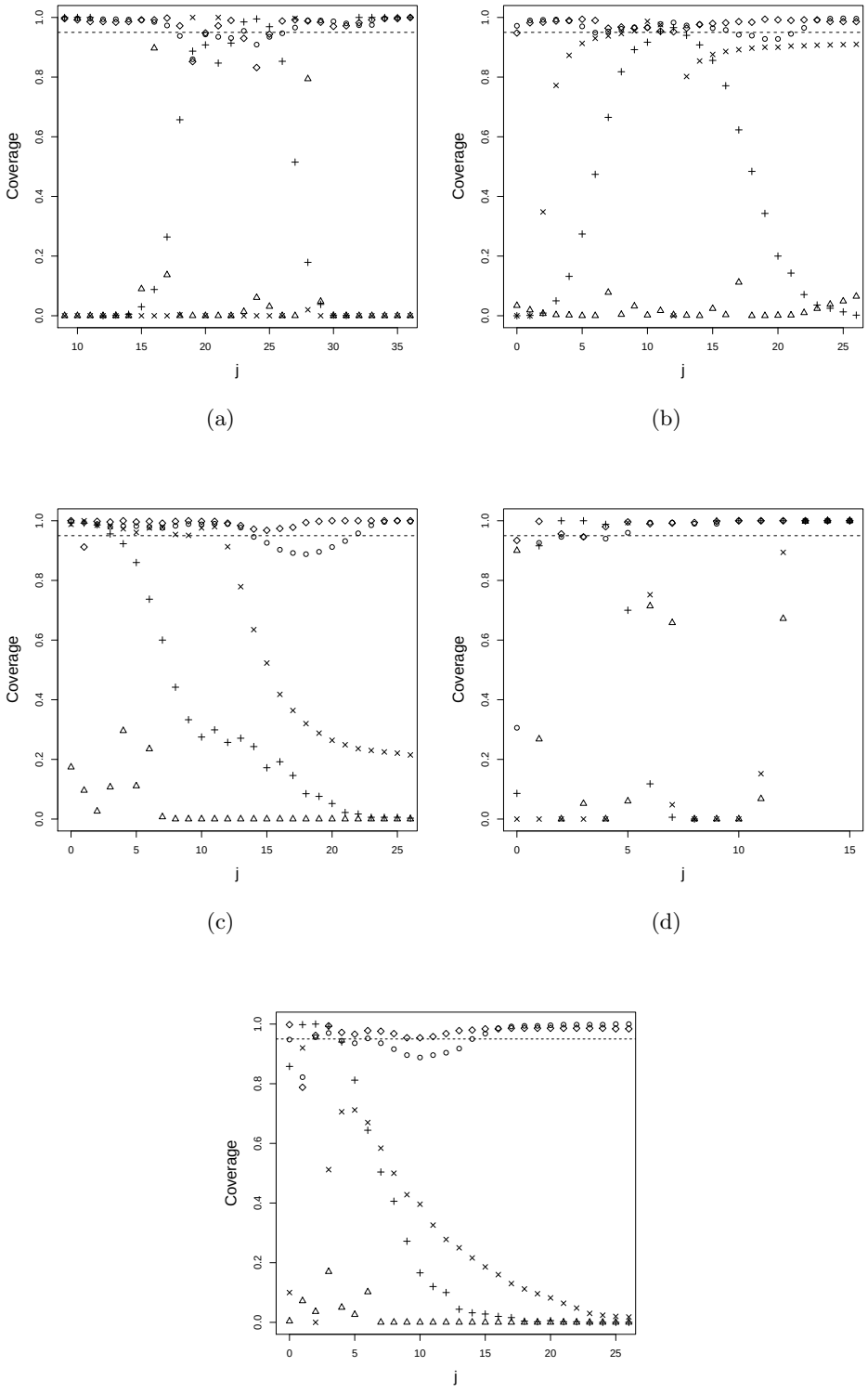


Figure 3.3: Coverage of 95% credible intervals for $p(j)$ under the four scenarios with $n = 25$. Points represent the RMG method, squares the RMSN, cross-shaped dots the DP with $\alpha = 1$, triangles the DP with $\alpha \sim Ga(1, 1)$ and x-shaped dots the DPM of Poisson.

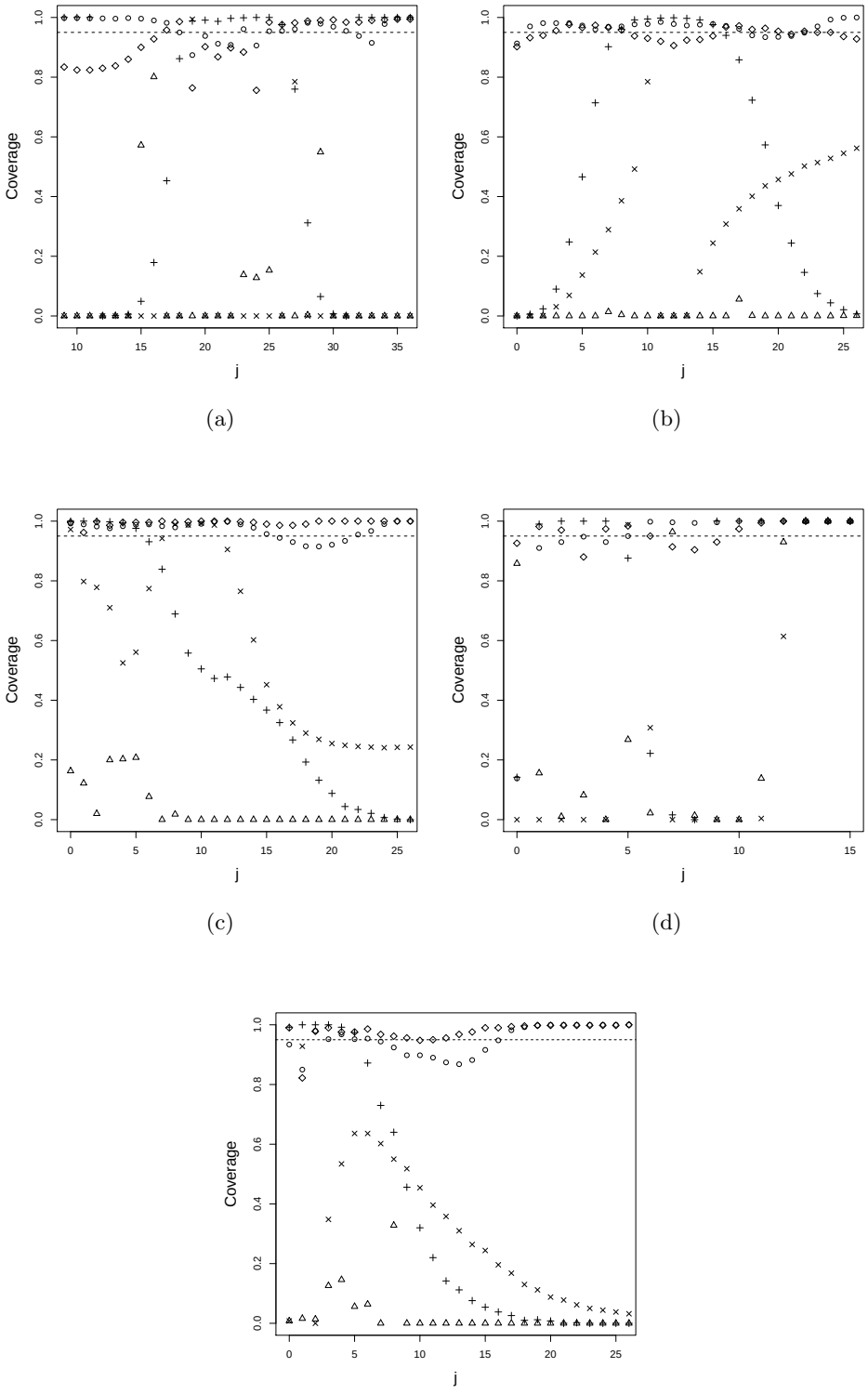


Figure 3.4: Coverage of 95% credible intervals for $p(j)$ under the four scenarios with $n = 50$. Points represent the RMG method, squares the RMSN, cross-shaped dots the DP with $\alpha = 1$, triangles the DP with $\alpha \sim Ga(1, 1)$ and x-shaped dots the DPM of Poisson.

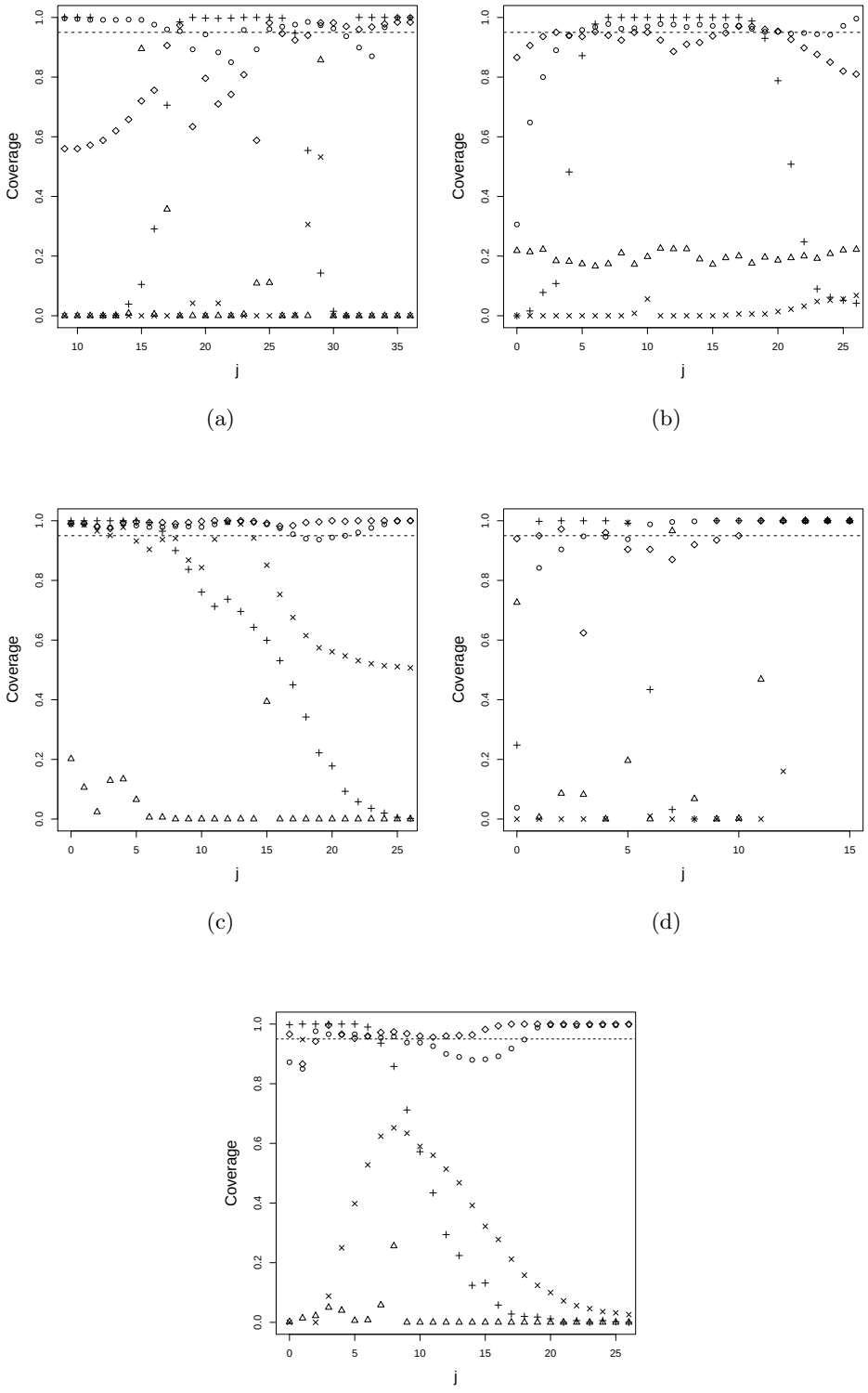
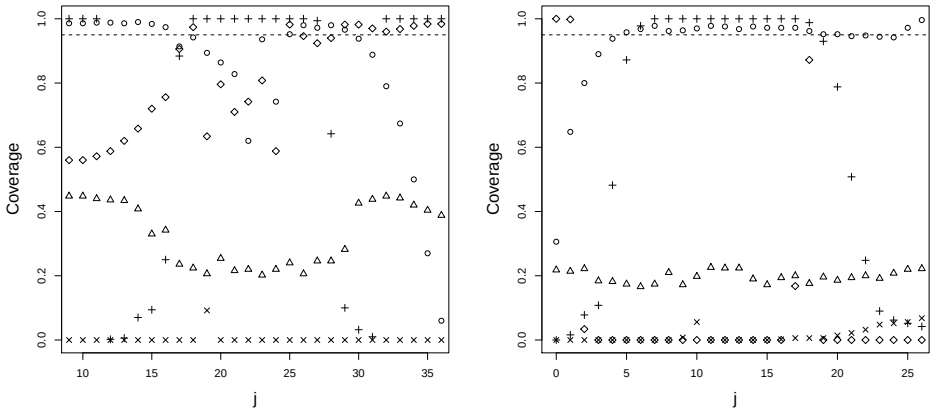
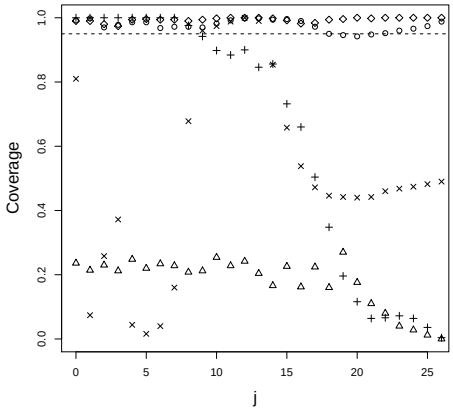


Figure 3.5: Coverage of 95% credible intervals for $p(j)$ under the four scenarios with $n = 100$. Points represent the RMG method, squares the RMSN, cross-shaped dots the DP with $\alpha = 1$, triangles the DP with $\alpha \sim Ga(1, 1)$ and x-shaped dots the DPM of Poisson.

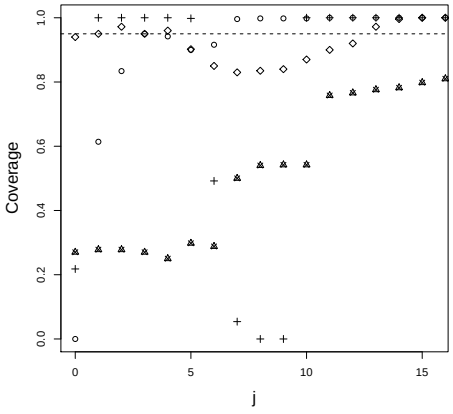


(a)

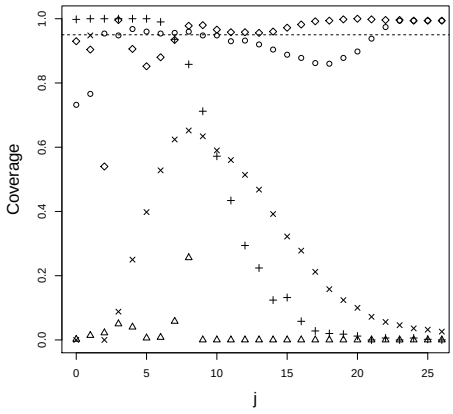
(b)



(c)



(d)



(e)

Figure 3.6: Coverage of 95% credible intervals for $p(j)$ under the four scenarios with $n = 300$. Points represent the RMG method, squares the RMSN, cross-shaped dots the DP with $\alpha = 1$, triangles the DP with $\alpha \sim Ga(1, 1)$ and x-shaped dots the DPM of Poisson.

3.4.2 Out of sample prediction simulation

In testing the model of Section 3.2.2 if used for Bayesian regression, we perform a simulation study. We assumed two scenarios, the first consists in generating data from the mixture

$$\mathbf{y}_i^* \sim \sum_{h=1}^{3} \pi_h N(\mu_h, \Sigma_h),$$

with $\pi = (\pi_1, \pi_2, \pi_3) = (0.14, 0.40, 0.46)$, $\mu_1 = (35, 82, 95)$, $\mu_2 = (-2, 1, 2.5)$, $\mu_3 = (12, 29, 37)$ and variance-covariance matrices

$$\Sigma_1 = \begin{pmatrix} 3 & -0.6 & 0.25 \\ -0.6 & 3 & 0.7 \\ 0.25 & 0.7 & 2 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1 & 0.5 & 0.4 \\ 0.5 & 1 & -0.4 \\ 0.4 & -0.4 & 0.7 \end{pmatrix}, \quad \Sigma_3 = 7.5 \cdot \Sigma_2,$$

with the positive observations floored and all negative values set equal to zero leading to a multivariate zero-inflated count distribution and the second scenario consisting in the mixture of multivariate Poisson distributions (Johnson et al., 1997):

$$\pi \text{Poi}_3(\lambda_1, \lambda_{01}) + (1 - \pi)\text{Poi}_3(\lambda_2, \lambda_{02})$$

with $\lambda_1 = (1, 8, 15)$, $\lambda_2 = c(8, 1, 3)$, $\lambda_{01}^{-1} = \lambda_{02} = 2$ and $\pi = 0.7$. For each scenario we simulate 100 dataset. The samples were then randomly and equally split into training and test subsets, with the Gibbs sampler applied to the training data and the results used to predict y_{i1} given y_{i2} and y_{i3} in the test sample.

We used the model (3.12) with multivariate Gaussian kernel and mixing distribution $P \sim \text{DP}(\alpha P_0)$ with base measure $P_0 = \text{N}_p(\mu; \mu_0, \kappa_0\Sigma)\text{Inv-W}(\Sigma; \nu_0, \mathbf{S}_0)$. The hyperparameters were specified as follows:

$$\begin{aligned} &\mu_0 \sim N_3(\bar{\mathbf{y}}^*, \hat{\mathbf{S}}), \ \mathbf{S}_0 \sim \text{InvWishart}(4, \Psi_0), \\ &\Psi_0 = I_3, \ \ \nu_0 = 4, \ \ \kappa_0 \sim \text{Gamma}(0.01, 0.01), \ \ \alpha = 1 \end{aligned} \tag{3.14}$$

with $\overline{\mathbf{y}}^* = (1 - \hat{p}_0)\overline{\mathbf{y}}_+ - \hat{p}_0\overline{\mathbf{y}}_+$, $\hat{p}_0$ the proportion of zeros in the training sample, $\overline{y}_+$ the mean of the non-zero values and $\hat{\mathbf{S}} = \text{diag}(s_1^2, s_2^2, s_3^3)$ with s_j the empirical variance of y_{ij}, $i = 1, \ldots, n$. The Gibbs sampler was run for $10,000$ iterations with the first $4,000$ discarded. We assessed predictive performance using the absolute deviation loss, which is more natural than squared error loss for count data and hence with the median of the posterior predictive distribution of y_{i1}.

We compare our approach with prediction under an oracle based on the true models, Poisson log-linear regressions fit with maximum likelihood, generalized additive models (GAM) (Hastie et al., 2001) with spline smoothing

function and generalized latent trait model (GLTM) (Moustaki and Knott, 2000; Dunson, 2003) with Poisson responses. The generalized latent trait model assumed a single latent variable which was assigned a standard normal prior, while a vague normal prior with mean 0 and variance 20 was assigned to the factor loadings with one of them constrained to be positive for identifiability. The out of sample prediction was made taking the median of a MCMC chain of length $12,000$ after a burn in of $3,000$ iterations from the posterior predictive distribution of y_{i1} in the test set. The results are reported in Table 3.3.

An additional gain of our approach is a flexible characterization of the whole predictive distribution of y_{i1} given y_{i2}, y_{i3} and not just the point prediction $\hat{y}_{i1}$. In addition to median predictions, it is often of interest in applications to predict subjects having zero counts or counts higher than a given threshold q. Based on our results, we obtained much more accurate predictions of both $y_{i1} = 0$ and $y_{i1} > q$ than either the log-linear Poisson model or the GAM approach when the true model is not a mixture of multivariate Poissons and prediction with similar degree of precision when the truth is a mixture of multivariate Poissons. As an additional competitor for predicting $y_{i1} = 0$ and $y_{i1} > q$, we also considered logistic regression, logistic GAM and a logistic latent trait model with the same prior specification as before fitted to the appropriate dichotomized data. Based on a 0-1 loss function that classified $y_{i1} = 0$ if the probability (posterior for our Bayes method and fitted estimate for the logistic GLM and GAM) exceeded 0.5, we compute the misclassification rate out-of-sample in Table 3.4.

Table 3.3: Mean absolute deviation errors for the prediction obtained with the RMG prior, the Oracle prediction, the linear regressions and the generalized latent trait model.

	Scenario 1	Scenario 2
RMG	2.44	1.42
oracle	1.36	1.28
GAM	2.72	1.55
GLM	5.34	1.98
GLTM	9.68	4.98

3.5 Applications to real data

3.5.1 Marketing segmentation

We apply now the methods earlier described to the telecommunications data introduced in Section 2.1.1. Telecommunications companies are often interested in understanding possible different behaviors of distinct geographic areas. To partially reach this goal, we estimate the probability mass func-

Table 3.4: Misclassification rate out-of-sample based on the proposed method, GAM, generalized linear regressions, oracle and generalized latent trait models for samples under scenario 1 (S1) and scenario 2 (S2).

		RMG		GAM		GLM		Oracle	GLTM
		Median[a]	0-1 Loss[b]	Poisson	Logistic	Poisson	Logistic	-	0-1 Loss[b]
S1	$y_{i1} = 0$	0.02	0.08	0.42	0.14	0.42	0.20	0.00	0.44
	$y_{i1} > 20$	0.02	0.10	0.02	0.40	0.08	0.30	0.00	0.50
	$y_{i1} > 25$	0.04	0.02	0.04	0.06	0.06	0.08	0.02	0.56
	$y_{i1} > 35$	0.06	0.06	0.06	0.08	0.06	0.14	0.06	0.48
S2	$y_{i1} = 0$	0.14	0.12	0.12	0.12	0.12	0.12	0.12	0.86
	$y_{i1} > 10$	0.16	0.12	0.16	0.12	0.16	0.12	0.12	0.50

a = prediction based on posterior median, b = prediction based on 0-1 loss

tions of the number of outgoing phone calls to landlines for three different geographic areas. We perform the analysis taking the subsample of the data of size 2,050 introduced in Section 2.1.1, we use the rounded mixture of skew normal described in Section 3.1.4.

We separately fit a rounded mixture of skew normals model for each of the three geographic areas. We empirically let the hyperparameters ξ_0 and κ to be respectively the sample median and variance in the three geographic areas, while letting $\alpha = 1$, $\psi = 10$ and $\nu_1 = \nu_2 = 1$. For computational reasons we perform the analysis taking subsamples of size 500 for each region.

Figure 3.7 (a) reports the three posterior mean cumulative mass functions for the three regions. The posterior estimates for the first and the second region are very similar. We obtained qualitatively comparable results also considering different randomly chosen subsamples of size 500. From this evidence we decided to merge the two regions and to fit again the model. Figure 3.7 reports the posterior mean cumulative mass functions along with the 95% credible bands (b), and the posterior mean probability mass functions (c) for the merged geographic areas (henceforth region A) and for the other area (region B) with panel (d) zooming for values around 10. The difference between the two regions, in fact, is stronger for values around 10, where also the credible bands are better separated.

Companies are often interested in identifying groups of customers with similar characteristics in terms of usage behaviour. The DPM precedure provides, as side effect, a clustering structure of the observations. Nonetheless, as outlined in Section 2.2.2, the direct interpretation of this clustering structure is difficult and we must face the label switching problem (Stephens, 2000; Jasra et al., 2005). Among the possible solutions to the problem we focus here on the proposal of Medvedovic and Sivaganesan (2002). This approach uses a hierarchical clustering procedure based on a dissimilarity matrix D^h for region $h \in \{A, B\}$ obtained from the proportion of MCMC samples in which two units were assigned to the same mixture component.

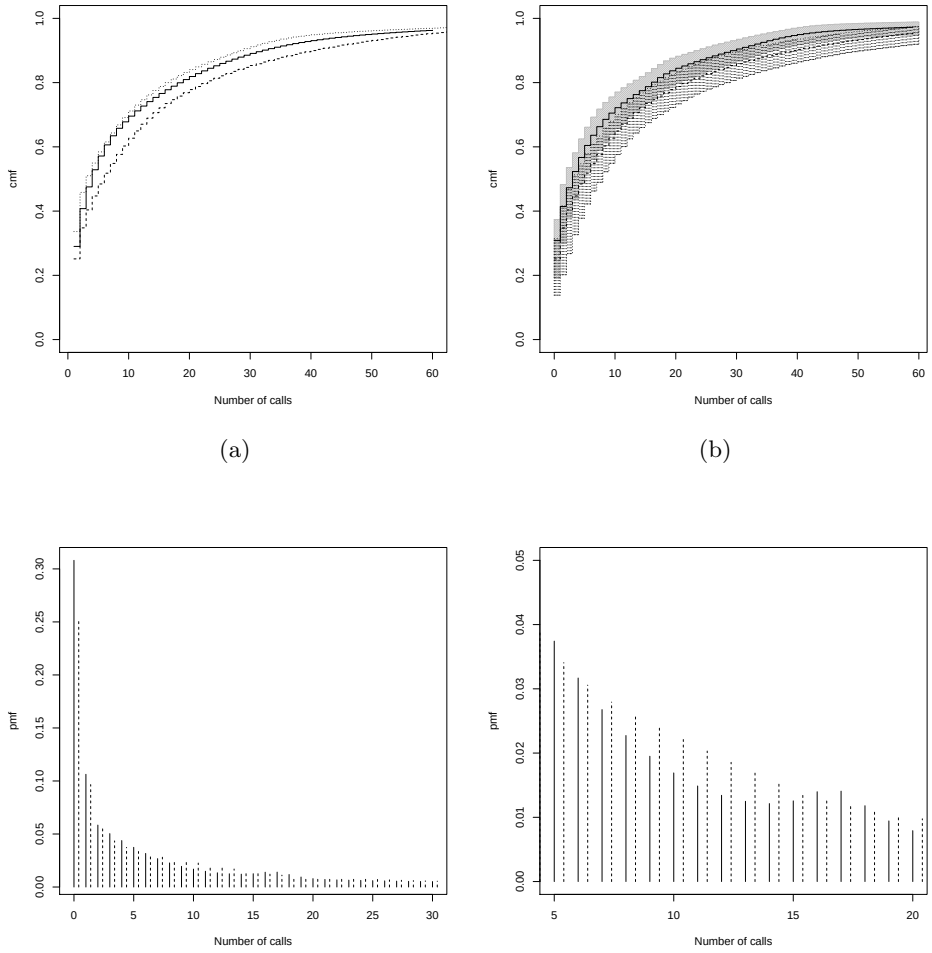
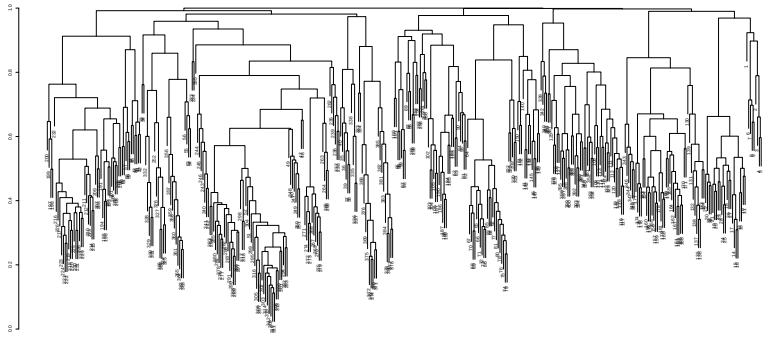
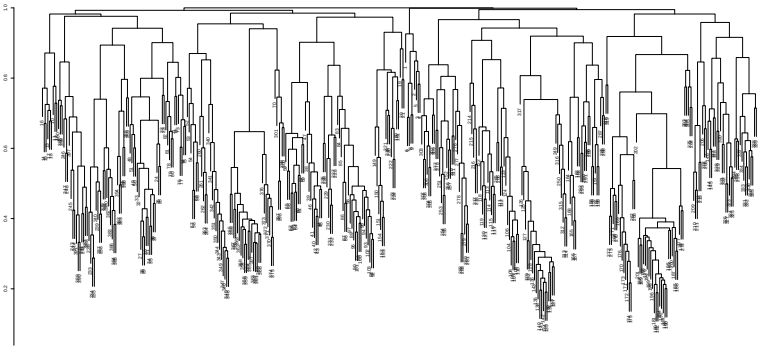


Figure 3.7: Posterior estimates of the probability mass functions: cumulative probability mass functions in the three geographic areas (a), cumulative probability mass functions and 95% credible bands(b), probability mass functions (c)–(d) for region A and B.



(a)



(b)

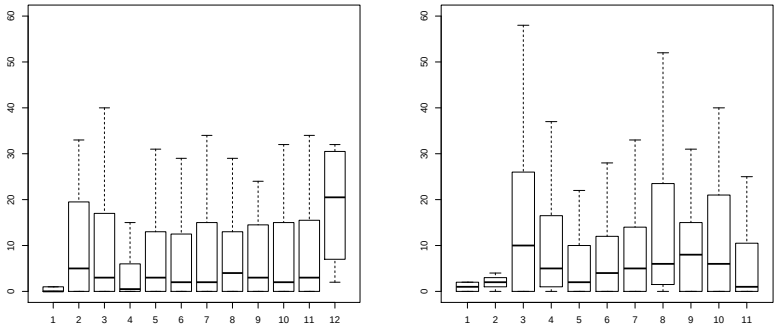
Figure 3.8: Dendrogram of region A (a) and B (b).

The i, j element $d_{i,j}$ of D^h is

$$d_{i,j} = \frac{\text{\# of MCMC samples where } S_i \neq S_j}{\text{\# of MCMC iterations after burn in}},$$

and S_i is the cluster indicator for subject i.

The dendrograms of a hierarchical cluster analysis with the complete linkage are reported in Figure 3.8. The posterior average number of clusters in the DP mixtures is respectively 11.5 for region A and of 10.8 for region B. The boxplots of the final clusters are reported in Figure 3.9, while Table 3.5 reports some descriptive statistics. Particularly interesting from a marketing perspective are cluster 1 of region A and cluster 1 and 2 of region B, corresponding to a proportion of low traffic customers and cluster 12 of region A which instead represents a portion of high usage customers.



(a) (b)

Figure 3.9: Boxplot of the final clusters of region A (a) and B (b).

3.5.2 Phone traffic prediction

We focus again on the same subset of $n = 2,050$ customers used in the previous section and described in Section 2.1.1. We consider now the multivariate observation $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip})$ representing usage in a month for card i. Specifically, we have the number of outgoing calls to landlines (y_{i1}), to mobile numbers of competing operators (y_{i2}) and to mobile numbers of the same operator (y_{i3}), as well as the total number of MMS (y_{i4}) and SMS (y_{i5}) sent.

We focus on the forecast of y_{i1}, using data on $y_{i2}, \ldots, y_{i5}$ with the technique proposed in Section 3.3 by first estimating the joint probability distribution of the multivariate $\mathbf{y}$.

The zero-inflation of the data is automatically accommodated by our method through using thresholds that assign negative underlying y^*_{ij} values to $y_{ij} = 0$. Excess mass at zero is induced through Gaussian kernels located at negative values.

We can model the data assuming the model in (3.12) with hyperparameters specified as in (3.14) and computation implemented as in Section 3.4.2. A training and test set of equal size are chosen randomly. Trace plots of y_{i1} for different individuals exhibit excellent rates of convergence and mixing, with the Geweke (1992) diagnostic providing no evidence of lack of convergence.

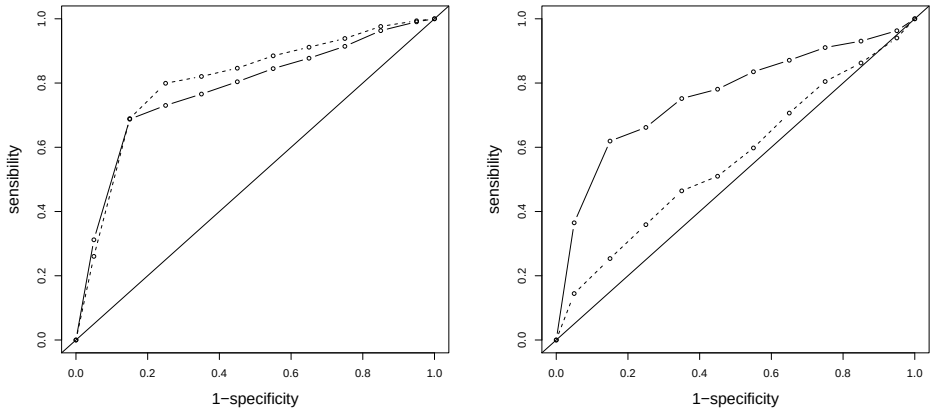
The method is compared with Poisson GLM and GAM and with a generalized latent trait model with prior as in Section 3.4.2. The out-of-sample median absolute deviation (MAD) value was 8.08 for our method, which is lower than the 8.76 obtained for the best competing method (Poisson GAM). The generalized latent trait model turns out to have a too restrictive structure with poor performance both computationally and in terms of prediction

Table 3.5: Descriptive statistics of the final clusters.

Region	Clus.	1st Qu.	Median	Mean	3rd Qu.	Max.	Proportion
A	1	0	0	4	1	24	0.02
	2	0	5	22.74	19.5	227	0.06
	3	0	3	11.79	16.5	71	0.07
	4	0	0.5	4.6	4.75	23	0.03
	5	0	3	15.85	13	216	0.13
	6	0	2	9.92	12.5	80	0.21
	7	0	2	11.29	15	59	0.10
	8	0	4	9.46	13	43	0.10
	9	0	3	17.33	14.5	278	0.11
	10	0	2	9.24	15	64	0.11
	11	0	3	10.6	15.25	63	0.05
	12	9.5	20.5	18.75	29.75	32	0.01
B	1	0	1	3.67	2	21	0.02
	2	1	2	4	3	22	0.02
	3	0	10	15.65	26	58	0.04
	4	1	5	17.93	16.5	280	0.14
	5	0	2	8.3	9.75	56	0.13
	6	0	4	9.86	12	92	0.14
	7	0	5	11.34	14	86	0.19
	8	1.75	6	15.14	22.25	113	0.09
	9	0	8	15.49	15	137	0.12
	10	0	6	16.18	21	83	0.04
	11	0	1	9.41	10.5	59	0.07

(MAD of 10.63). These results were similar for multiple randomly chosen training-test splits.

An important goal for the company is to predict customers with no outgoing calls and highly profitable customers. The estimated multivariate probability mass function allow us to compute the probability if these events. We predict such customers using Bayes optimal prediction under a 0-1 loss function. Using optimal prediction of zero-traffic customers, we obtained lower out-of-sample misclassification rates than the Poisson GAM, but had comparable results to logistic GAM as illustrated in the ROC curve in Figure 3.10 (a). Our expectation is that the logistic GAM will have good performance when the proportion of individuals in the subgroup of interest is $\approx 50\%$, but will degrade relative to our approach as the proportion gets closer to 0% or 100%. In this application, the proportion of zeros was 69% and the sample size was not small, so logistic GAM did well. The results for predicting highly profitable customers having more than 40 calls per month are consistent with this. In fact, even if the proportion of customers that make more than 40 calls were low () in the training set, out method is able to produce a good estimate of this probability. As illustrated in Figure 3.10 (b), it is clear that our approach had dramatically better predictive performance than the logistic GAM.



(a) (b)

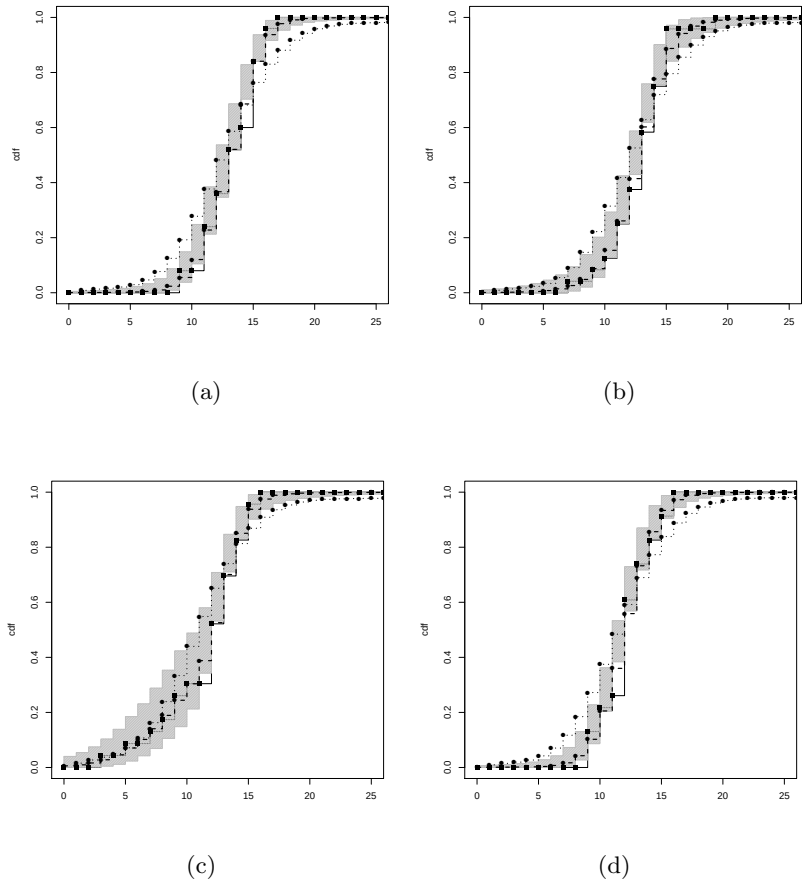
Figure 3.10: ROC curves for predicting customers having outgoing calls to landlines equal to zero (a) or more than 40 (b). The continuous line is for our proposed approach and the dotted lines are for the logistic GAM. Both classifications are based on a 0-1 loss function that classify $y_{i1} = 0$ or $y_{i1} > 40$ if the posterior (estimated) probability is greater than 1/2.

3.5.3 Developmental toxicity study

To show that the proposed methods can be applied in any situation involving counts, we also perform an application to a toxicity study. We consider now data from a developmental toxicity study of ethylene glycol in mice conducted by the National Toxicology Program (Price et al., 1985). As in many biological applications in which there are constraints on the range of the counts, the data are underdispersed having mean 12.54 and variance 6.78.

Pregnant mice were assigned to dose groups of 0, 750, 1,500 or 3,000 mg/kg per day, with the number of implants measured for each mouse at the end of the experiment. Group sizes are 25, 24, 23 and 23, respectively. The scientific interest is in studying a dose response trend in the distribution of the number of implants. To address this, we first estimate the probability mass function within each group using the RMG methodology of Section 3.1.3. Trace plots showed rapid convergence and excellent mixing, with the Geweke (1992) diagnostic failing to show lack of convergence.

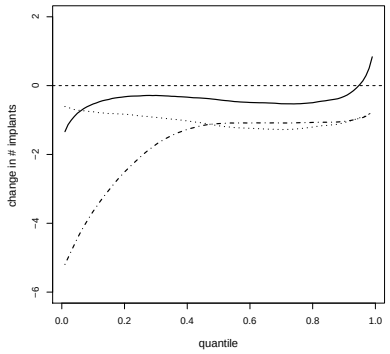
Figure 3.11 shows the estimated and empirical cumulative distribution functions in each group along with 95% pointwise credible intervals and the estimates from a DPM of Poissons analysis. Clearly, the DPM of Poissons provided a poor fit to the data and hence poor characterization of changes with dose, while the proposed RMG method provided an excellent fit for each



(a) (b)

(c) (d)

Figure 3.11: Posterior estimates for the cumulative distribution function for (a) the control group and (b)–(d) the dose groups. Black solid line for the empirical cumulative distribution function, dashed line for the RMG estimation and dotted for the DPM of Poisson. Gray shading for 95% posterior credible bands for the RMG



(a)

Figure 3.12: Posterior mean for the changes in the percentiles (x-axis) between the control group and 750 mg/kg (continuous line), 1,500 mg/kg (dash-dotted line) and 3,000 mg/kg (dotted line) dose groups.

group. To summarize changes in the distribution of the number of implants with dose, we estimated summaries of the posterior distributions for changes in each percentile between the control group and each of the exposed groups, with the results shown in Figure 3.12. In each of the dose levels, the exposure led to a stochastic decrease in the distribution of the number of implants, with an estimated decrease in the number of implants at each percentile (there is a minor exception at high percentiles in the 750 mg/kg group). The estimated posterior probabilities of a negative average change across the percentiles was 0.72, 0.99 and 0.94 in the 750, 1,500 and 3,000 mg/kg groups, respectively. These results were consistent with Mann-Whitney pairwise comparison tests that had p-values of 0.23, 0.04 and 0.06 for stochastic decreases in the low, medium and high dose groups. In contrast, likelihood ratio tests under a Poisson model failed to test any significant differences between the control and exposed groups.

Chapter 4

Mixed scale data

In our example, the data consist of count variables recording the number of events made by customers and of continuous variables such as the average duration of the phone calls or the total cost for customer. The data are hence on a mixed scale of measure. This situation is common in a broad class of applications.

Ideally, to estimate unknown joint distributions for mixed scale data, one could rely on the modelling framework introduced in Chapter 3 using a straightforward modification of the mapping function $h(\cdot)$ defined in (3.11). If this could be done from a modeling point of view, theoretically the mixed-scale framework gives rise to some complications. Strong posterior consistency, in particular, does not trivially follow given the KL condition because the topological equivalence of the weak and strong metrics used in Theorem 3.5 is not satisfied in the space of mixed-scale densities. For this reason, in this chapter, we develop the theoretical basis for Bayesian nonparametric estimation of mixed-scale density showing appealing theoretical properties, such as large support, posterior consistency and near optimal rates of convergence.

4.1 Preliminaries and notation

Our focus is on modeling of joint probability distributions of mixed scale data $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T$, where $\mathbf{y}_1 = (y_{1,1}, \ldots, y_{1,p_1}) \in \mathbb{R}^{p_1}$ is a $p_1 \times 1$ vector of continuous observations and $\mathbf{y}_2 = (y_{2,p_1+1}, \ldots, y_{2,p}) \in Q$ where $Q = \bigotimes_{j=1}^{p_2} \{0, 1, \ldots, q_j - 1\}$ is a $p_2 \times 1$ vector of discrete variables having $\mathbf{q} = (q_1, \ldots, q_{p_2})^T$ as the respective number of levels and $p_2 = p - p_1$. Clearly $\mathbf{y}_2$ can include binary variables ($q_j = 2$), categorical variables ($q_j > 2$) or counts ($q_j = \infty$). Hence, $\mathbf{y}$ is a $p \times 1$ vector of variables having mixed measurement scales. A graphical representation of a mixed scale density with $p_1 = p_2 = 1$ and $q_1 = 6$ is reported in Figure 4.1. We let $y \sim f$, with f denoting the joint density with respect to an appropriate dominating measure μ to be defined

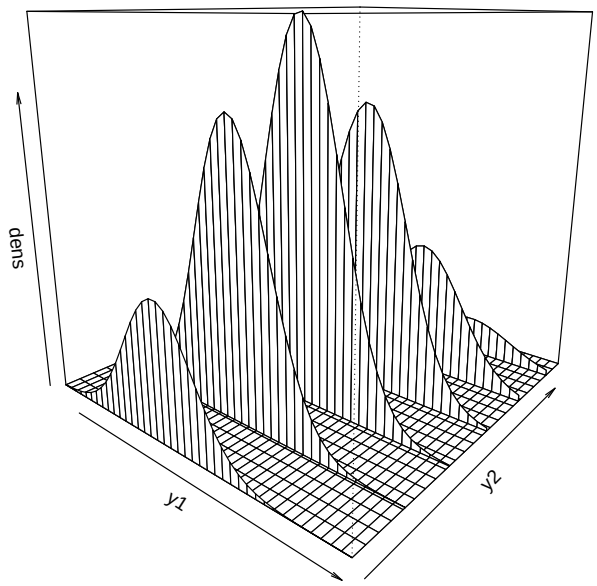


Figure 4.1: Mixed scale density $p_1 = p_2 = 1$ and $q_1 = \infty$

below. The set of all possible such joint densities is denoted $\mathcal{F}$. Following a Bayesian nonparametric approach, we propose to specify a prior $f \sim \Pi$ for the joint density having large support over $\mathcal{F}$.

For the continuous variables, we let $(\Omega_1, \mathcal{S}_1, \mu_1)$ denote the σ-finite measure space having $\Omega_1 = \mathbb{R}^{p_1}$, $\mathcal{S}_1$ the Borel σ-algebra of subsets of Ω_1, and μ_1 the Lebesgue measure. Similarly for the discrete variables we let $(\Omega_2, \mathcal{S}_2, \mu_2)$ denote the σ-finite measure space having $\Omega_2 \subset \mathbb{N}^{p_2}$, a subset of the p_2-dimensional set of natural numbers, $\mathcal{S}_2$ containing all non-empty subsets of Ω_2, and μ_2 the counting measure. Then, we let $\mu = \mu_1 \times \mu_2$ be the product measure on the product space $(\Omega, \mathcal{S}) = (\Omega_1, \mathcal{S}_1) \times (\Omega_2, \mathcal{S}_2)$. To formally define the joint density f, first let ν denote a σ-finite measure on $(\Omega, \mathcal{S})$ that is absolutely continuous with respect to μ. Then, by the Radon-Nikodym theorem there exists a function f such that $\nu(A) = \int_A f d\mu$.

In studying properties of a prior Π for the unknown density f, such as large support and posterior consistency, it is necessary to define notions of distance and neighborhoods within the space of densities $\mathcal{F}$. Letting $f_0 \in \mathcal{F}$ denote an arbitrary density, such as the true density that generated the

data, the Kullback-Leibler divergence of f from f_0 can be defined as

$$d_{KL}(f_0, f) = \int f_0 \log(f_0/f) d\mu = \int_{\Omega_1} \int_{\Omega_2} f_0 \log(f_0/f) d\mu_1 \, d\mu_2$$
$$= \int_{\mathbb{R}^{p_1}} \sum_{\mathbf{y}_2 \in Q} f_0(\mathbf{y}_1, \mathbf{y}_2) \log \left(\frac{f_0(\mathbf{y}_1, \mathbf{y}_2)}{f(\mathbf{y}_1, \mathbf{y}_2)} \right) d\mu_1(\mathbf{y}_1)$$

with the integrals taken in any order from Fubini's theorem. Another topology is induced by the L_1-metric. If f and f_0 are probability distributions with respect to the product measure μ, their L_1-distance is defined as

$$d_1(f_0, f) = \int |f_0 - f| d\mu = \int_{\Omega_1} \int_{\Omega_2} |f_0 - f| d\mu_1 \, d\mu_2$$
$$= \int_{\mathbb{R}^{p_1}} \sum_{\mathbf{y}_2 \in Q} |f_0(\mathbf{y}_1, \mathbf{y}_2) - f(\mathbf{y}_1, \mathbf{y}_2)| d\mu_1(\mathbf{y}_1).$$

4.2 Consistency in multivariate mixed-scale density estimation

In order to induce a prior $f \sim \Pi$ for the density of the mixed scale variables, we let

$$\mathbf{y} = h(\mathbf{y}^*), \quad \mathbf{y}^* \sim f^*, \quad f^* \sim \Pi^*, \tag{4.1}$$

where $h : \mathbb{R}^p \to \Omega$, $\mathbf{y}^* = (y_1^*, \ldots, y_p^*)^T \in \mathbb{R}^p$, $f \in \mathcal{L}$, $\mathcal{L}$ is the set of densities with respect to Lesbesgue measure over $\mathbb{R}^p$, and Π^* is a prior over $\mathcal{L}$. In Chapter 3 we used the notation $\mathcal{L}$ to denote the set of densities with respect to Lesbesgue measure over $\mathbb{R}$, while we refer here to the set of densities with respect to Lesbesgue measure over $\mathbb{R}^p$. This abuse of notation will keep the details light without leading to misunderstanding since it is clear that the focus of this chapter is on densities over p-dimensional spaces. To generalize the mapping h to this settings, we let

$$h(\mathbf{y}^*) = \left\{ h_1(\mathbf{y}_1^*)^T, h_2(\mathbf{y}_2^*)^T \right\}^T, \tag{4.2}$$

where $h_1(\mathbf{y}_1^*) = \mathbf{y}_1^*$ is the identity function and h_2 are thresholding functions that replace the real-valued inputs with non-negative integer outputs by thresholding the different inputs separately. Let $A^{(j)} = \{A_1^{(j)}, \ldots, A_{q_j}^{(j)}\}$ denote a prespecified partition of $\mathbb{R}$ into q_j mutually exclusive subsets, for $j = 1, \ldots, p_2$, with the subsets ordered so that $A_h^{(j)}$ is placed before $A_l^{(j)}$ for all $h < l$. Then, letting $A_{\mathbf{y}_2} = \{\mathbf{y}_2^* : y_{2,j}^* \in A_{y_{2j}}^{(j)}, j = 1, \ldots, p_2\}$, the mixed scale density f is defined as

$$f(y) = g(f^*) = \int_{A_{\mathbf{y}_2}} f^*(\mathbf{y}^*) \mathrm{d}\mathbf{y}^*. \tag{4.3}$$

The function $g : \mathcal{L} \to \mathcal{F}$ defined in (4.3) is a mapping from the space of densities with respect to Lesbesgue measure on $\mathbb{R}^p$ to the space of mixed-scale densities $\mathcal{F}$.

Clearly the properties of the induced prior $f \sim \Pi$ will be driven largely by the properties of $f^* \sim \Pi^*$. Lemma 1 shows that the mapping $g : \mathcal{L} \to \mathcal{F}$ maintains Kullback-Leibler neighborhoods. The proof is omitted as being a straightforward modification of that for Lemma 3.2.

Lemma 4.1. *Choose any f_0^* such that $f_0 = g(f_0^*)$ for any fixed $f_0 \in \mathcal{F}$. Let $\mathcal{K}_\epsilon(f_0^*) = \{f^* : d_{KL}(f_0^*, f^*) < \epsilon\}$ be a Kullback-Leibler neighborhood of size ϵ around f_0^*. Then the image $g(\mathcal{K}_\epsilon(f_0^*))$ contains values $f \in \mathcal{F}$ in a Kullback-Leibler neighborhood of f_0 of at most size ϵ.*

As discussed in Section 2.2.3, large support of the prior plays a crucial role in posterior consistency under the theory of Schwartz (1965). Given f_0 in the KL support of the prior, in fact, it is sufficient to ensure the existence of an exponentially consistent sequence of tests for the hypothesis $H_0 : f = f_0$ versus $H_1 : f \in U^C(f_0)$ where $U(f_0)$ is a neighborhood of f_0. Ghosal et al. (1999) show that the existence of such a sequence of tests is guaranteed by balancing the size of a sieve and the prior probability assigned to its complement.

We now provide sufficient conditions for L_1 posterior consistency for priors in the class proposed in expression (4.1). Our Theorem 4.2 builds on Theorem 8 of Ghosal et al. (1999). The main differences are that we define the sieve $\mathcal{F}_n$ as $g(\mathcal{L}_n)$, where $\mathcal{L}_n$ is a sieve on $\mathcal{L}$ and that we require conditions on the prior probability in terms of the underlying Π^*. The proof relies on the same steps of Ghosal et al. (1999) which give an upper bound for the L_1-metric entropy $J(\delta, \mathcal{L}_n)$ defined as the logarithm of the minimum number of δ-sized L_1 balls needed to cover $\mathcal{L}_n$.

Theorem 4.2. *Let Π be a prior on $\mathcal{F}$ induced by Π^* as described in expression (4.1). Suppose f_0 is in the KL support of Π and let $U = \{f \in \mathcal{F} : ||f - f_0|| < \epsilon\}$. If for each $\epsilon > 0$, there is a $\delta < \epsilon$, c_1, $c_2 > 0$, $\beta < \epsilon^2/8$ and there exist sets $\mathcal{L}_n \subset \mathcal{L}$ such that for n large*

(i) $\Pi^*(\mathcal{L}_n^C) \leq c_1 e^{-nc_2}$;

(ii) $J(\delta, \mathcal{L}_n) < n\beta$

then $\Pi(U \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) \to 1$ a.s. P_{f_0}.

Proof. The proof consists of two steps. First we need to determine the size of the parameter space of $\mathcal{F}$, measured in terms of L_1-metric entropy. Then we need to show that the conditions of an unpublished work of Barron (Ghosh and Ramamoorthi, 2003, see Theorem 4.4.3 of) are satisfied. We start introducing two lemmas that are useful to determine the size of the

parameter space of $\mathcal{F}$ in terms of L_1-metric entropy. The first shows that the L_1 topology is maintained under the mapping g and the second bounds the L_1-metric entropy of a sieve.

Lemma 4.3. *Assume that the true data generating density is $f_0 \in \mathcal{F}$. Choose any f_0^* such that $f_0 = g(f_0^*)$. Let $U(f_0^*) = \{f^* : ||f_0^* - f^*|| < \epsilon\}$ be a L_1 neighborhood of size ϵ around f_0^*. Then the image $g(U(f_0^*))$ contains values $f \in \mathcal{F}$ in a L_1 neighborhood of f_0 of at most size ϵ.*

The proof is omitted since it follows directly from the definition of L_1 neighborhood and from Fubini's theorem.

Lemma 4.4. *Let $\mathcal{L}_n \subset \mathcal{L}$ denote a compact subset of $\mathcal{L}$, with $J(\delta, \mathcal{L}_n)$ the L_1-metric entropy corresponding to the logarithm of the minimum number of δ-sized L_1 balls needed to cover $\mathcal{L}_n$. Letting $\mathcal{F}_n = g(\mathcal{L}_n)$, we have $J(\delta, \mathcal{F}_n) \leq J(\delta, \mathcal{L}_n)$.*

Proof. Let $k = \exp\{J(\delta, \mathcal{L}_n)\}$ be the number of δ balls needed to cover $\mathcal{L}_n$, with $f_1^*, \ldots, f_k^*$ denoting the centers of these balls so that $\mathcal{L}_n \subset \bigcup_{i=1}^k \mathcal{L}_{n,i}$, where $\mathcal{L}_{n,i} = \{f^* : ||f^* - f_i^*|| < \delta\}$. From Lemma 4.3, it is clear we can define $\mathcal{F}_n \subset \bigcup_{i=1}^k \mathcal{F}_{n,i}$ where $\mathcal{F}_{n,i} = g(\mathcal{L}_{n,i})$ is an L_1 neighborhood around $f_i = g(f_i^*)$ of size at most δ. This defines a covering of $\mathcal{F}_n$ using k δ-sized L_1 balls, but this is not necessarily the minimal covering possible and hence $J(\delta, \mathcal{L}_n)$ provides an upper bound on $J(\delta, \mathcal{F}_n)$. □

The rest of the proof follows along almost the same lines of Ghosal et al. (1999) in showing that the sets $\mathcal{F}_n \cap \{f : ||f - f_0|| < \epsilon\}$ and $\mathcal{F}_n^C$ satisfy the conditions of an unpublished result of Barron (Ghosh and Ramamoorthi, 2003). □

We now state a theorem on the rate of convergence (contraction) of the posterior distribution. The theorem gives conditions on the prior Π^* similar to those directly required by Theorem 2.1 of Ghosal et al. (2000). This theorem implies that if an optimal rate exists for a prior Π^* the rate of the induced Π is not bigger than that.

Theorem 4.5. *Let Π be the prior on $\mathcal{F}$ induced by Π^* as described in expression (4.1) and $U = \{f : d(f, f_0) \leq M\epsilon_n\}$ with d the L_1 or Hellinger distance. Suppose that for a sequence ϵ_n, with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$, a constant $C > 0$, sets $\mathcal{L}_n \subset \mathcal{L}$ and $B_n^* = \{f^* : \int f_0^* \log(f_0^*/f^*)d\mu \leq \epsilon_n^2, \int f_0^*(\log(f_0^*/f^*))^2 d\mu \leq \epsilon_n^2\}$ defined for a given $f_0^* \in g^{-1}(f_0)$, we have*

(iii) $J(\epsilon_n, \mathcal{L}_n) < Cn\epsilon_n^2$;

(iv) $\Pi^*(\mathcal{L}_n^C) \leq \exp\{-n\epsilon_n^2(C+4)\}$;

(v) $\Pi^*(B_n^*) \geq \exp\{-Cn\epsilon_n^2\}$

then for sufficently large M, we have that $\Pi(U^C \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) \to 0$ in P_{f_0}-probability.

Proof. Let $\mathcal{F}_n = g(\mathcal{L}_n)$. From Lemma 4.4 we have $J(\delta, \mathcal{F}_n) \leq J(\delta, \mathcal{L}_n)$. Let $D(\epsilon, \mathcal{F})$ the ϵ-packing number of $\mathcal{F}$, i.e. is the maximal number of points in $\mathcal{F}$ such that the distance between every pair is at least ϵ. For every $\epsilon > \epsilon_n$, using (iii) we have

$$\log D(\epsilon/2, \mathcal{F}) < \log D(\epsilon_n, \mathcal{F}) < Cn\epsilon_n^2.$$

Therefore applying Theorem 7.1 of Ghosal et al. (2000) with $j = 1$, $D(\epsilon) = \exp(n\epsilon_n^2)$ and $\epsilon = M\epsilon_n$ with $M > 2$ there exist a sequence of tests $\{\Phi_n\}$ that satisfies

$$E_{f_0}\{\Phi_n\} \leq \exp\{-(KM^2-1)n\epsilon_n^2\}, \quad \sup_{f \in U^C \cap \mathcal{F}_n} E_f\{1-\Phi_n\} \leq C\exp\{-KnM^2\epsilon_n^2\}. \tag{4.4}$$

The posterior probability assigned to U^C can be written as

$$\begin{aligned}\Pi\left\{U^C \mid \mathbf{y}_1, \ldots, \mathbf{y}_n\right\} &= \frac{\int_{U^C \cap \mathcal{F}_n} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f) + \int_{U^C \cap \mathcal{F}_n^C} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)}{\int \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)} \\ &\leq \Phi_n + \frac{(1-\Phi_n)\int_{U^C \cap \mathcal{F}_n} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f) + \int_{U^C \cap \mathcal{F}_n^C} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)}{\int \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)}.\end{aligned}$$

Taking $KM^2 > K + 1$ the first summand $E_{f_0}\{\Phi_n\} \leq 2\exp\{-Kn\epsilon_n^2\}$ by (4.4). The rest of the proof consists in proving that the remaining equation goes to zero in P_{f_0}-probability. By Fubini's theorem and (4.4) we have

$$E_{f_0}\left\{(1-\Phi_n)\int_{U^C \cap \mathcal{F}_n} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)\right\} \leq \sup_{f \in U^C \cap \mathcal{F}_n} E_f\{1-\Phi_n\} \leq \exp\{-KnM^2\epsilon_n^2\},$$

while by (iv) we have

$$E_{f_0}\left\{\int_{U^C \cap \mathcal{F}_n^C} \prod_{i=1}^n \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f)\right\} \leq \Pi(\mathcal{F}_n^C) = \Pi^*(\mathcal{L}_n^C) \leq \exp\{-n\epsilon_n^2(C+4)\}.$$

The numerator of the second summand is hence exponentially small for $M > \sqrt{(C+4)/K}$. Finally we need to lower bound the denumerator. Clearly

$$g(B_n^*) \subseteq B_n = \left\{f : \int f_0 \log(f_0/f) d\mu \leq \epsilon_n^2, \int f_0 (\log(f_0/f))^2 d\mu \leq \epsilon_n^2\right\}$$

and then $\Pi(B_n) \geq \Pi(g(B_n^*)) = \Pi^*(B_n^*)$ and using condition (v) on $\Pi^*(B_n^*)$ we have

$$\begin{aligned}\int_{B_n} \int f_0 \log(f_0/f) d\mu d\Pi(f) &\leq \int_{B_n} \epsilon_n^2 d\Pi(f) \\ \int_{B_n} \int f_0 \left(\log(f_0/f)\right)^2 d\mu d\Pi(f) &\leq \int_{B_n} \epsilon_n^2 d\Pi(f).\end{aligned}$$

Then using Lemma 8.1 of Ghosal et al. (2000) we obtain

$$E_{P_0} \int \prod_{i=1}^{n} \frac{f(\mathbf{y}_i)}{f_0(\mathbf{y}_i)} \mathrm{d}\Pi(f) \geq \exp\{-n\epsilon_n^2(C+4)\}$$

that concludes the proof. □

Clearly, the properties of the induced prior $f = g(f^*) \sim \Pi$ depend heavily on the choice of prior $f^* \sim \Pi^*$. Our hope is to leverage the literature on models and theory for continuous density estimation in developing associated models and theory for the mixed scale case.

As already discussed, Dirichlet process mixtures are the most widely applied class of models for Bayesian density estimation, with a rich theoretical literature available in the univariate continuous case on posterior consistency (Ghosal et al., 1999; Barron et al., 1999; Tokdar, 2006; Wu and Ghosal, 2008) and rates of posterior contraction (Ghosal et al., 2000; Ghosal and van der Vaart, 2001, 2007; Walker et al., 2007; Scricciolo, 2011). DPMs of Gaussian kernels have proven successful for multivariate density estimation in challenging cases involving high-dimensional data (Chen et al., 2010).

However the only available results on asymptotic properties of Bayesian procedures for multivariate continuous density estimation are presented by Ghosal and co-authors (Wu and Ghosal, 2010; Shen and Ghosal, 2011). In both papers the models considered are quite limited in scope in focusing on DP location mixtures of Gaussian kernels. Posterior consistency is studied in Wu and Ghosal (2010) assuming a truncated inverse-Wishart prior for the Gaussian covariance. In Shen and Ghosal (2011) near minimax optimal rates of posterior contraction are shown under some conditions on the true density assuming a diagonal covariance in the Gaussian kernel with independent truncated inverse-gamma priors on the diagonal elements. In practice, it is well known that using a diagonal covariance may lead to less efficient results in small to moderate samples. In addition, it is preferable to avoid arbitrary truncations and allow broader priors than inverse gammas and inverse Wisharts. For example, for high-dimensional data it is well known that inverse Wisharts provide a poor choice and alternatives based on factor analytic and other factorizations are commonly used.

The generalization of Wu and Ghosal (2010) and Shen and Ghosal (2011) to more flexible mixtures that enable scaling to higher dimensions deserve further studies both from the pure continuous multivariate density estimation point of view and since it is the starting point of the procedure discussed in this chapter.

Chapter 5

Count stochastic processes modeling

In this chapter we focus again only in count observations. Motivated by the need of modeling count customer specific mobile phone traffic, we discuss a class of count stochastic process models that rely on mapping a real-valued stochastic process $y^* : \mathcal{D} \to \mathbb{R}$. Under some regularity conditions we show that the introduced methodology leads to L_1 posterior consistency. In a functional data analysis context we analyze the data introduced in Section 2.1.1 with the goal of predicting outgoing churners.

5.1 Model formulation

Let $y \in \mathcal{C}$ denote a count-valued stochastic process, with $\mathcal{D} \subset \mathbb{R}^p$ compact and $\mathcal{C}$ the set of all $\mathcal{D} \to \mathcal{N}$ functions. In Chapter 3 we used the notation $\mathcal{C}$ to denote the set of probability mass functions over $\mathbb{N}$, while we refer here to the set of step functions from a domain space $\mathcal{D}$ with values in $\mathbb{N}$. Also, in Chapter 3, y was a real scalar while here y is a function from $\mathcal{D}$ to $\mathbb{N}$. Since in this chapter we do not use any result of Chapter 3 we are confortable that this abuse of notation will be free of any misunderstanding.

We choose a prior $y \sim \Pi$, where Π is a probability measure over $(\mathcal{C}, \mathcal{B})$, with $\mathcal{B}(\mathcal{C})$ the Borel σ-algebra of subsets of $\mathcal{C}$. The measure Π induces the marginal probability mass functions

$$\text{pr}\{y(s) = j\} = \Pi\{y : y(s) = j\} = \pi_j(s), \quad j \in \mathcal{N}, \quad s \in \mathcal{D}, \tag{5.1}$$

and the joint probability mass functions

$$\begin{aligned} \text{pr}\{y(s_1) = j_1, \ldots, y(s_k) = j_k\} &= \Pi\{y : y(s_1) = j_1, \ldots, y(s_k) = j_k\} \\ &= \pi_{j_1 \ldots j_k}(s_1, \ldots, s_k), \end{aligned} \tag{5.2}$$

for $j_h \in \mathcal{N}$ and $s_h \in \mathcal{D}$, $h = 1, \ldots, k$, and any $k \geq 1$.

We specify a prior Π with large support using the rounding idea described in Chapter 3. To our knowledge, there is no previously defined stochastic process that satisfies a large support condition. In the absence of prior knowledge that allows one to assume y belongs to a pre-specified subset of $\mathcal{C}$ with probability one, priors must satisfy the large support property to be coherently Bayesian.

We propose to induce a prior $y \sim \Pi$ through

$$y = h(y^*), \quad y^* \sim \Pi^*, \tag{5.3}$$

where with similar notation as before $y^* : \mathcal{D} \to \mathbb{R}$ is a real-valued stochastic process, h is a rounding operator, Π^* is a probability measure over $(\mathcal{Y}, \mathcal{B})$, $\mathcal{Y}$ is the set of all $\mathcal{D} \to \mathbb{R}$ continuous functions and $\mathcal{B}(\mathcal{Y})$ are the corresponding Borel sets. Unlike count-valued stochastic processes, there is a rich literature on real-valued stochastic processes. For example, Π^* could be chosen to correspond to a GP or could be induced through various basis or kernel expansions of y^*.

Also here there are various ways in which the rounding operator h can be defined. We will focus for simplicity on the case in which $y(s) = h\{y^*(s)\} = j$ if $y^*(s) \in A_j = [a_j, a_{j+1})$ for $j \in \mathcal{N}$, with $\{A_j\}_{j=1}^{\infty}$ and pre-specify the thresholds $a_0 < \cdots < a_\infty$ as in Chapter 3.

Under this setting $\pi_j(s)$ and $\pi_{j_1 \ldots j_k}(s_1, \ldots, s_k)$ of equations (5.1) – (5.2) become

$$\begin{aligned} \pi_j(s) = g\{f_s(y)\}[j] &= \int_{A_j} f_s(y^*) \mathrm{d}y^* && (5.4) \\ \pi_{j_1 \ldots j_k}(s_1, \ldots, s_k) &= g(f_{s_1,\ldots,s_k})[J] \\ &= \int_{A_{j_1} \times \cdots \times A_{j_k}} f_{s_1,\ldots,s_k}(y_1^*, \ldots, y_k^*) \mathrm{d}y_1^* \ldots \mathrm{d}y_k^* && (5.5) \end{aligned}$$

where $f_s\{y^*(s)\}$ and $f_{s_1,\ldots,s_k}(y_1^*, \ldots, y_k^*)$ are the marginal and joint density functions of the underlying process.

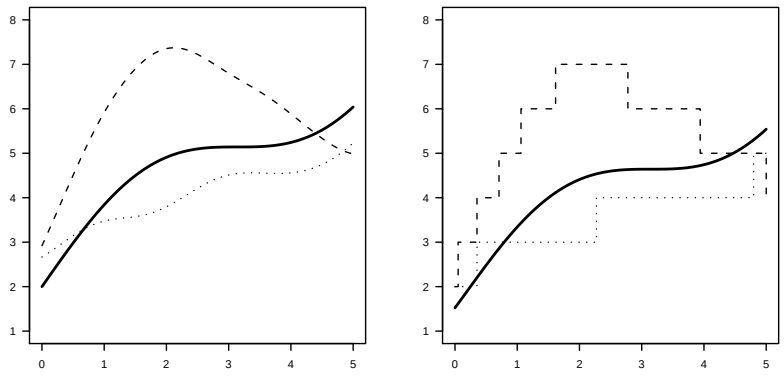
Figure 5.1 illustrates the prior through showing realizations of the underlying stochastic process (a) and resulting count processes after applying the rounding operator (b). The thick lines represents the mean functions of the real valued process and of the induced process. The latter is

$$E\{y(s)\} = \sum_{j=0}^{\infty} j\{F_s(a_{j+1}) - F_s(a_j)\},$$

where $F_s(x) = \int_{-\infty}^{x} f_s(y^*) \mathrm{d}y^*$.

5.2 Asymptotic properties

We first introduce a regularity condition on y useful later to prove some asymptotic property. Assumption 5.1 rules out infinitely many jumps in y



(a) (b)

Figure 5.1: Panel (a) represents samples from a Gaussian process with mean function $\mu(s) = 2+\sin(s)+s$ (bold line) and squared exponential covariance function. Panel (b) shows how the mapping function (5.3) works with $a_0 = -\infty$ and $a_j = j$ for $j = 1, 2, \ldots$. Dotted and dashed lines are the rounded version of panel (a) realizations while the bold line is the induced mean function.

across the compact domain.

Assumption 5.1. *y is a piecewise constant function such that the domain $\mathcal{D}$ can be expressed as a countable union of mutually disjoint sets $\mathcal{D} = \bigcup_l \mathcal{D}_l(y)$, with $y(s)$ constant within the interior of each set $\mathcal{D}_l(y)$ and with unit increments at the boundaries $\mathcal{B}(y)$. As a convention, the points $s \in \mathcal{B}(y)$ at the boundaries are defined to fall within the set having the higher $y(s)$ value.*

The mapping function $h(\cdot)$ in (5.3) is surjective and hence the inverse mapping $h^{-1}(y)$ will correspond to an uncountable set of infinitely many continuous stochastic processes y^* such that $y = h(y^*)$. As an important step in characterizing the support of the induced prior $y \sim \Pi$, Lemma 5.1 ensures the existence of at least one continuous stochastic process for each count process.

Lemma 5.1. *For every count stochastic process $y_0 \in \mathcal{C}$, there exists at least one continuous $y^* : \mathcal{D} \to \mathbb{R}$ such that $y_0 = h(y^*)$.*

Proof. For any count stochastic process y_0 satisfying Assumption 5.1, we can partition the domain $\mathcal{D}$ into mutually disjoint sets $\mathcal{D}_l(y_0)$, with $y_0(s)$ constant within the interior of each $\mathcal{D}_l(y_0)$ and with unit increments at the boundaries. There are clearly infinitely many continuous functions $y^* : \mathcal{D} \to \mathbb{R}$ satisfying the constraints (i) $y^*(s) \in [a_{y_0(s)}, a_{y_0(s)+1})$ for all $s \in \mathcal{D}$ and (ii) $y^*(s) = a_{y_0(s)}$ for $s \in \mathcal{B}(y_0)$. For all such y^*, we have $y_0 = h(y^*)$. □

It is also important, in the flavor of Lemma 3.2, to assess how neighborhoods are maintained in applying the mapping. An L_∞ neighborhood of y_0 of size ϵ is defined as

$$\eta_{\epsilon,\infty}(y_0) = \Big\{ y : d_\infty(y_0, y) = \sup_{s\in\mathcal{D}} |y_0(s) - y(s)| < \epsilon \Big\},$$

while an L_1 neighborhood is

$$\eta_\epsilon(y_0) = \left\{ y : d_1(y_0, y) = \int |y_0(s) - y(s)| \mathrm{d}s < \epsilon \right\}.$$

Lemma 5.2. *Suppose y^* and y_0^* are continuous and bounded by $M \in \mathbb{R}$ such that $d_1(y^*, y_0^*) = \epsilon^*$, $y = h(y^*)$, $y_0 = h(y_0^*)$. Then, $y \in \eta_\epsilon(y_0)$ for all $\epsilon > \zeta(\epsilon^*; y_0^*)$, where $\zeta(\epsilon^*; y_0^*)$ is non decreasing in ϵ^* having $\lim_{\epsilon^*\to 0} \zeta(\epsilon^*; y_0^*) = 0$.*

Proof. Take $\mathcal{D} = [0,1]^p$ without loss of generality. Let $\{\mathcal{D}_l(y_0, y)\}_{l=1}^m$ the partition of $\mathcal{D}$ induced by $\{\mathcal{D}_l(y_0)\}_{l=1}^{m_0}$ and $\{\mathcal{D}_l(y)\}_{l=1}^{m_1}$ such that $y(s) = j_l$ and $y_0(s) = k_l$ for all $s \in \mathcal{D}_l(y_0, y)$ and some $j_l, k_l \in \mathcal{N}$. Let $\delta_l(y_0, y) = |j_l - k_l|$, for $l = 1, \ldots, m$ and $\lambda(\cdot)$ be the Lebesgue measure. Define

$$\zeta(\epsilon^*; y_0^*) = \sup_{y^*\in\eta_{\epsilon^*}(y_0^*)} \left\{ \max_{l=1,2,\ldots} [\delta_l\{y_0, h(y^*)\}] \sum_{l:\delta_l\neq 0} \lambda[\mathcal{D}_l\{y_0, h(y^*)\}] \right\}.$$

Clearly $y \in \eta_\epsilon(y_0)$ for all $\epsilon > \zeta(\epsilon^*; y_0^*)$ since

$$d_1(y_0, y) = \sum_{l=1}^m \delta_l(y_0, y)\lambda\{\mathcal{D}_l(y_0, y)\} \leq \zeta(\epsilon^*; y_0^*).$$

We show first that $\lim_{\epsilon^*\to 0} \zeta(\epsilon^*; y_0) = 0$. What follows holds for all $y^* \in \eta_{\epsilon^*}(y_0^*)$. Consider the general $y^* \in \eta_{\epsilon^*}(y_0^*)$. Since $\sum_{l:\delta_l\neq 0} \lambda[\mathcal{D}_l\{y_0, h(y^*)\}]$ is finite, $\zeta(\epsilon^*; y_0)$ goes to zero if $\max \delta_l\{y_0, h(y^*)\}$ goes to zero. Define $M_{\epsilon^*} = \max |y^*(s) - y_0^*(s)|$ and let $s_M = \arg\max |y^*(s) - y_0^*(s)|$ with s_M belonging to a given $\mathcal{D}_l$ where $y^*(s) \leq a_{j_l+1}$ and $y_0^*(s) \leq a_{k_l+1}$. For construction $|a_{l_l+1} - a_{k_l+1}| \leq M_{\epsilon^*}$ and so for $M_{\epsilon^*} \to 0$ we have $a_{l_l+1} = a_{k_l+1}$. Considering that $\max |y^*(s) - y_0^*(s)| \to 0$ then $|y^*(s) - y_0^*(s)| \to 0$ for all $s \in \mathcal{D}$ leading also to $\max \delta_l \to 0$. Whereas the absolute value of the difference $|y^*(s) - y_0^*(s)|$ is bounded and continuous we have that if $\int_\mathcal{D} |y^*(s) - y_0^*(s)| \mathrm{d}s$ goes to zero, also $\limsup_\mathcal{D} |y_0^*(s) - y^*(s)|$ goes to zero and hence also M_{ϵ^*}.

The fact that $\zeta(\cdot; y_0)$ is non decreasing follows directly from its definition. □

Lemma 5.2 ensures that the mapping h maintains L_1 neighborhoods, immediately implying that a prior Π^* that assigns positive probability to arbitrarily small L_1 neighborhoods of y_0^* will assign positive probability to

arbitrarily small neighborhoods of $y_0 = h(y_0^*)$. Theorem 5.3, which is critical to showing large support for the prior Π, is an immediate consequence of the first two lemmas.

Theorem 5.3. *Assuming the prior Π^* assigns positive probability to L_1 neighborhoods of any continuous function $y_0^* : \mathcal{D} \to \mathbb{R}$, the prior Π induced through (5.3) assigns positive probability to L_1 neighborhoods of any $y_0 \in \mathcal{C}$.*

Proof. By Lemma 5.2 with suitable ϵ^* we have

$$\Pi\{\eta_\epsilon(y)\} = \Pi[h\{\eta_{\epsilon^*}(y^*)\}] = \Pi^*\{\eta_{\epsilon^*}(y^*)\} > 0.$$

□

In addition to showing large support of the prior, it is important to verify that the posterior distribution for y concentrates increasingly around the true process y_0 as the sample size increases. Theorem 5.4 provides sufficient conditions under which L_1 posterior consistency is obtained.

Assumption 5.2. *Let $\mathcal{D} = [0,1]^p$ and assume the n values of s_i arise with an in-fill design such that we can cover $\mathcal{D}$ with n L_∞ balls centered around $s_1, \ldots, s_n$ of size δ with $2\delta \in \left(n^{-1/p}, \lfloor n^{1/p} \rfloor^{-1}\right)$.*

Theorem 5.4. *Let $y \in \mathcal{C}$ be a count stochastic process with $y_i = y(s_i)$, for $i = 1, \ldots, n$ and $(s_1, \ldots, s_n)$ following Assumption 5.2. Letting $y_0 \in \mathcal{C}$ denote the true stochastic process and $y \sim \Pi$, then if $\Pi\{\eta_\epsilon(y_0)\} > 0$ for any ϵ and there exist sets $\{\mathcal{C}_n\}_{n=1}^\infty$ with $\mathcal{C}_n \in \mathcal{C}$ and $\mathcal{C}_n^C$ the complement of $\mathcal{C}_n$, where $\Pi\{\mathcal{C}_n^C\} < c_1 e^{-c_2 n}$, and c_1, c_2 positive constants, then*

$$\Pi\left\{\eta_\epsilon^C(y_0) \,|\, y_1, \ldots, y_n\right\} \to 0. \tag{5.6}$$

Proof. Since $y_0(s_i)$ is equal to the observed y_i for all i, we can rewrite the posterior (5.6) as

$$\begin{aligned}
&\Pi\left\{y \in \eta_\epsilon^C(y_0) \mid y_1, \ldots, y_n\right\} = \\
&= \frac{\int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y) + \int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n^C} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y)}{\int_{\mathcal{C}} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y)} \\
&\leq \Phi_n + \frac{(1-\Phi_n)\int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y) + \int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n^C} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y)}{\int_{\mathcal{C}} \prod_{i=1}^n \delta_{y_i}(y_i)\mathrm{d}\Pi(y)} \\
&= \Phi_n + \frac{I_{1,n}(y_1, \ldots, y_n) + I_{2,n}(y_1, \ldots, y_n)}{I_{3,n}(y_1, \ldots, y_n)},
\end{aligned}$$

where δ_a is a delta mass at a, Φ_n is a test function and $\mathcal{C}_n$ is a sieve that grows eventually to the whole space $\mathcal{C}$. It suffices to show that

$$\begin{aligned} \Phi_n &\to 0 && (5.7)\\ e^{\beta_1 n} I_{1,n}(y_1,\ldots,y_n) &\to 0 && (5.8)\\ e^{\beta_2 n} I_{2,n}(y_1,\ldots,y_n) &\to 0 && (5.9)\\ e^{\beta n} I_{3,n}(y_1,\ldots,y_n) &\to \infty && (5.10) \end{aligned}$$

with $\beta < \min\{\beta_1, \beta_2\}$.

Denote $\lfloor a \rfloor$ the integer part of a and let $\mathcal{D} = \bigcup_{j=1}^{\lfloor n^{1/p} \rfloor^p} \mathcal{G}_j$ with $\mathcal{G}_j$ an L^∞ ball of size $0.5(\lfloor n^{1/p} \rfloor)^{-1}$ and center s'_j, where the centers are chosen on a grid so that $\lfloor n^{1/p} \rfloor^p$ balls cover $\mathcal{D}$ and each $\mathcal{G}_j$ contains at least one element of $(s_1,\ldots,s_n)^T$ under Assumption 5.2. Define $X_i = 1\{y(s_i) = y_0(s'_j)\}$ with s'_j being the centroid of the $\mathcal{G}_j$ in which s_i is contained. Let $\Phi_n = 1\{\sum_{i=1}^n X_i < n\}$ the test on the set

$$\mathcal{C}_n = \left\{ y : y \text{ is constant in } \mathcal{G}_j, \text{for all } j = 1,\ldots,\lfloor n^{1/p} \rfloor^p, ||y||_\infty < M_n \right\} \tag{5.11}$$

with $M_n = \mathcal{O}(n^\alpha)$ and $1/2 < \alpha < 1$. The first condition on the sieve governs the regularity of the process while the second gives an upper bound for the infinity norm as in Choi and Schervish (2007). The true y_0 belongs to $\mathcal{C}_n$ for a given n and hence for n sufficiently large the test functions have exactly zero type I and type II probability. From this (5.7) is directly verified. We continue to prove (5.8). By Fubini's theorem we have

$$\begin{aligned} E_{y_0}\{I_{1,n}(y_1,\ldots,y_n)\} &= E_{y_0}\left\{ (1-\Phi_n) \int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n^C} \prod_{i=1}^n \delta_{y_i}(y_i) \mathrm{d}\Pi(y) \right\} \\ &= \int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n^C} E_y\{(1-\Phi_n)\} = 0 \end{aligned}$$

where the final equality is directly verified by the test construction. Next we prove (5.9). Again by Fubini's theorem we have

$$\begin{aligned} E_{y_0}\{I_{2,n}(y_1,\ldots,y_n)\} &= E_{y_0}\left\{ \int_{\eta_\epsilon^C(y_0)\cap\mathcal{C}_n^C} \prod_{i=1}^n \delta_{y_i}(y_i) \mathrm{d}\Pi(y) \right\} \\ &\leq \Pi(\mathcal{C}_n^C) \\ &\leq c_1 e^{-c_2 n}. \end{aligned}$$

An implication of the first Borel-Cantelli Lemma yields to

$$e^{cn} I_{2,n}(y_1,\ldots,y_n) \to 0.$$

Finally the prior positivity of Π makes $I_{3,n}(y_1,\ldots,y_n)$ to be positive. This proves also (5.10) and concludes the proof. □

From Theorems 5.3 and 5.4, it follows that the prior proposed in equation (5.3) will lead to L_1 posterior consistency under Assumption 5.2 as long as Π^* assigns positive probability to L_1 neighborhoods of any continuous function and negligible probability to $\mathcal{Y}_n^C = h^{-1}(\mathcal{C}_n^C)$ for n increasing. Choi and Schervish (2007) shown that this condition holds, if $\mathcal{Y}_n^C$ has a particular form, for Π^* corresponding to suitably chosen Gaussian process and orthogonal basis expansion priors. Gaussian process priors and infinite basis expansions lead to well known computational bottlenecks, so as a practical alternative we instead rely on finite basis expansions using Bayesian P-splines (Eilers and Marx, 1996; Lang and Brezger, 2004).

5.3 Posterior computation

Suppose we observe a count process y at n different locations $s_i \in \mathcal{D}$, for $i = 1, \ldots, n$ and $y_i = y(s_i)$. We estimate the whole process y given the realizations at the observed locations using the model proposed in Section 5.1. Defining the thresholds as $a_0 = -\infty$ and $a_j = j$ for $j = 1, 2, \ldots$ our rounded Bayesian P-splines model has $y_i^* \sim N\{b(s_i)^{\mathrm{T}}\theta, \tau^{-1}\}$ where $b(x)$ is the B-spline basis at x with priors $p(\tau) \propto \tau^{-1}$, $p(\theta \mid \lambda) \propto \exp(-1/2\lambda\theta^{\mathrm{T}}P\theta)$, where $P = D^{\mathrm{T}}D$ is a penalty matrix with D the rth order difference matrix, e.g. for $r = 2$

$$D = \begin{pmatrix} 1 & -2 & 1 & 0 & \ldots & 0 \\ 0 & 1 & -2 & 1 & \ldots & 0 \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & -2 & 1 \end{pmatrix},$$

and Gamma priors for λ and ρ, precisely $\lambda \sim \mathrm{Ga}(\nu/2, \rho\nu/2)$ and $\rho \sim \mathrm{Ga}(a, b)$. The prior for the basis coefficients induces a penalty on finite differences of the coeffients of adjacent B-splines. The parameter λ is a roughness penalty. The hyperparameter ρ controls dispersion of the prior. By choosing a hyperprior with small a, b values, one induces a prior with heavy tails and good performance in a variety of settings (Jullion and Lambert, 2007).

Under these settings the Gibbs sampler in Algorithm 4 trivially follows from Lang & Brezger (2004).

Algorithm 4 Gibbs sampling algorithm: rounded P-spline

Step 1: Generate each y_i^* from $N\{b(s_i)^{\mathrm{T}}\theta, \tau^{-1}\}$ under the constraints $a_{y_i} \leq y_i^* < a_{y_i+1}$

Step 2: Sample θ, τ and λ from their full conditional posterior distributions

5.4 Simulation study

A simulation study is performed to assess the performance of the proposed approach. Two different approaches for estimating the trajectory of the stochastic process were compared with our rounded P-splines. The first is a Poisson regression with mean parameter $\lambda(s)$ estimated nonparametrically with a frequentist spline smoother. The second is a simple interpolating step function defined as

$$f(s) = y_1 \mathbb{I}_{s<s_2}(s) + \sum_{j=2}^{n} y_j \mathbb{I}_{s_j \leq s < s_{j+1}}(s).$$

For our method, we considered both the posterior median of $y(s)$ and the median of the posterior predictive distribution.

Several simulations have been run under different simulation settings leading to qualitatively similar results. We report the results for four scenarios. The first scenario generates count stochastic processes from a Poisson distribution with domain varying mean parameter equal to $2 + s/5 + \sin(s)$ while in the second the stochastic process is generated rounding the realization of a Gaussian process plus an error term,

$$y = h(y^*), \quad y^* \sim \mathrm{GP}(\mu, k) + \epsilon$$

with mean function $\mu(s) = 2 + \exp(s/5)$, covariance function $k(s, s^{'})$ taken to be squared exponential and $\epsilon(s)$ independent draws from $N(0, 2)$. Under the third scenario we generate from a Poisson count process with rate parameter $1/2$ and in the fourth from the same Gaussian process of the second scenario without adding the error term. For each case, we generated data on a equispaced grid of 200 points between 0 and 20. Taking equispaced subsamples for different level of sparsity, namely of sizes $n = 10$, $n = 25$ and $n = 50$ we estimate the trajectory on a fine grid for 1 000 replicates for each scenario and each method. Methods are compared based on averaging the mean absolute deviation between the estimate and the true process across the replicates and grid points. Table 5.1 summarizes the results. Our rounded P-splines methodology always lead to a lower mean absolute deviation. Note that the first two scenarios do not satisfy Assumption 5.1 since in both cases for each point of the domain we are generating independent random variables that can lead to infinitely many discontinuity points. Nonetheless in these two cases we do not achieve the asymptotic properties studied in Section 5.2, in finite samples the methods still have good performance.

In implementing the blocked Gibbs sampler for the rounded P-splines, the first 3,000 iterations were discarded as a burn-in and the next 5,000 samples were used to calculate the posterior median of $y(s)$. For the hyperparameters we chose $a = b = 1/2$, $\nu = 1$ and D to be the second order

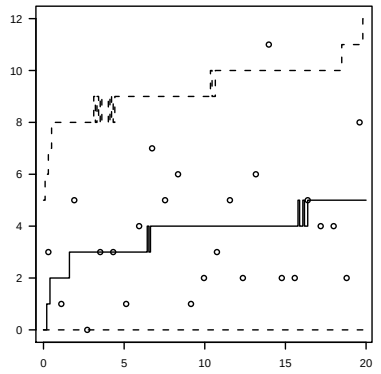
Table 5.1: Mean absolute deviation in simulation study of Section 5.3. RPS, rounded P-splines; PRPS, predictive rounded P-splines; NPP, nonparametric Poisson model, E empirical interpolating step function

	Scenario 1			Scenario 2		
	$n = 10$	$n = 25$	$n = 50$	$n = 10$	$n = 25$	$n = 50$
RPS	1.79	1.53	1.27	2.05	1.63	1.32
PRPS	1.85	1.72	1.67	2.08	1.81	1.70
NPP	2.25	2.08	1.94	14.21	13.74	12.85
E	2.26	2.20	2.19	3.58	2.62	2.35
	Scenario 3			Scenario 4		
RPS	0.22	0.12	0.07	0.51	0.27	0.18
PRPS	0.22	0.13	0.08	0.51	0.28	0.22
NPP	3.14	3.01	2.82	14.06	13.57	12.68
E	0.44	0.20	0.11	2.53	1.22	0.70

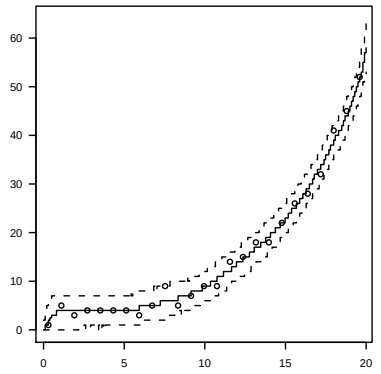
difference matrix. The number of knots is prespecified to be equal to 20 leading to a coarse equispaced grid of the domain. Similar results were obtained increasing the number of knots. The trace plots of the parameters showed excellent mixing and the Geweke (1992) diagnostic indicated rapid convergence. Figure 5.2 shows the posterior predictive median, of the process along with 95% credible bands for representative simulations under $n = 25$.

5.5 Count functional data

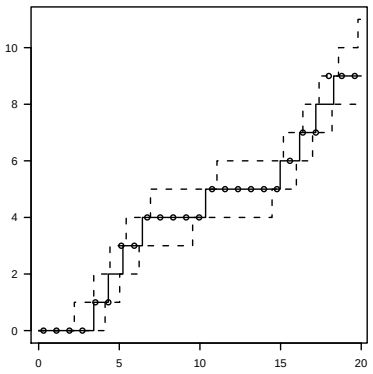
We have focused on the case in which there is a single count process y that is observed at different locations $s = (s_1, \ldots, s_n)^T$. However, in many applications, there are count processes $\{y_i, i = 1, \ldots, n\}$ independently observed from n individuals, with the ith process observed at locations $s_i = (s_{i1}, \ldots, s_{in_i})^T$. We refer to such data as count functional data. As in other functional data settings, it is of interest to borrow information across the individual functions through use of a hierarchical model. This can be accomplished easily within our rounded stochastic processes framework by first defining a functional data model for a collection of underlying continuous functions $\{y_i^*, i = 1, \ldots, n\}$, and then letting $y_i = h(y_i^*)$, for $i = 1, \ldots, n$. There is a rich literature on appropriate models for $\{y_i^*, i = 1, \ldots, n\}$ such as hierarchical Gaussian processes (Behseta et al., 2005), wavelet-based functional mixed models (Morris and Carroll, 2006) and multivariate kernel partition process mixtures (Dunson, 2010).



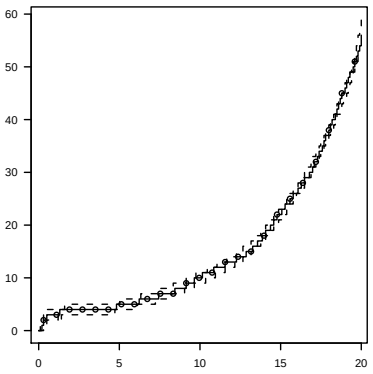
(a) Fitted model under scenario 1



(b) Fitted model under scenario 2



(c) Fitted model under scenario 3



(d) Fitted model under scenario 4

Figure 5.2: Posterior medians and 95% credible intervals for one representative simulation under each scenario for $n = 25$.

5.5.1 Outgoing churn prevision

As motivated in Section 2.1, churn prevision is one of the main goal of telecommunications companies. We tackle here a simplified problem in which we use the dataset introduced in Section 2.1.1 and we want to predict churn as a function of outgoing video calls traffic which, as mentioned before, is relevant variables for the company. To perform this analysis we use the proposed rounded P-spline methodology to estimate the traffic count trajectories of deactivated and active customers.

Let $y_i(s)$ denote the number of outgoing video calls for customer i at time s, $y_{it} = y_i(s_{it})$ denote the number at the tth observation time, and $g_i \in \{0, 1\}$ index if the customer deactiveted his/her contract in the 19th month or not. Then, we assume that

$$y_{it} = \xi_i + b(s_{it})^T \theta_{g_i} + \epsilon_{it}, \quad \xi_i \sim Q, \quad \epsilon_{it} \sim N(0, \tau^{-1}), \tag{5.12}$$

where $b(x)$ is the B-spline basis at x, θ_g represents the basis coefficients specific to group g, ξ_i is a customer-specific random effect, and ϵ_{it} is an error term. To allow the random effect distribution to be unknown, we choose a DP prior, with $Q \sim \text{DP}(\alpha Q_0)$, $\alpha = 1$ and $Q_0 = N(0, \psi)$. This hierarchical structure allows different mean traffic trajectories within each group, while allowing certain customers to have greater or lower usage behaviour.

As prior distributions we let $p(\theta \mid \lambda) \propto \exp(-1/2\lambda\theta^T P\theta)$ with $P = D^{\mathrm{T}}D$, and $\lambda \sim \text{Ga}(\nu/2, \rho\nu/2)$ as in Section 5.3. We additionally choose hyperpriors $p(\tau) \propto \tau^{-1}$, $\rho \sim \text{Ga}(a_\rho, b_\rho)$, and $\psi \sim \text{Ga}(a_\psi, b_\psi)$.

The full conditional posterior distributions are

$$\begin{aligned}
\theta_j \mid \lambda, \tau &\sim N\left\{\tau S_j V_j^{-1}, V_j\right\}, \\
\tau \mid y, \theta &\sim \text{Ga}\left(nt/2, R/2\right), \\
\lambda \mid \nu, \rho &\sim \text{Ga}\left(\nu/2 + K\text{rank}\{P\}/2, 1/2\rho\nu + U/2\right), \\
\rho \mid a_\rho, b_\rho, \nu, \lambda &\sim \text{Ga}(a_\rho + \nu/2, b_\rho + \nu\lambda/2),
\end{aligned}$$

where $S_j = \sum_{i:g_i=j}\{B^{\mathrm{T}}(y_i - \xi_i)\}$, $V_j = (\tau n_j B^{\mathrm{T}}B + \lambda P)^{-1}$, $R = \sum_{i=1}^{n}(y_i - B\theta_i - \xi_i)^{\mathrm{T}}(y_i - B\theta_i - \xi_i)$ and $U = \sum_{g=1}^{K}\theta_g^{\mathrm{T}}P\theta_g$. Updating of $\{\xi_i\}$ and ψ follows along standard lines for Dirichlet process mixture models and hence details are excluded. In this applications we use the blocked Gibbs sampler of Ishwaran and James (2001).

To evaluate the accuracy of the model we split the dataset into two parts, performing the analysis on one subset (training set), and validating the analysis on the other subset (test set). As described in Section 2.1.1, the proportion of deactivations is very low (4.07%) as usual in this kind of applications. A common approach, when the event we want to predict is rare, is to balance the training set in order to have the response variable outcomes equally represented. With this in mind we fit the model (5.12) to a

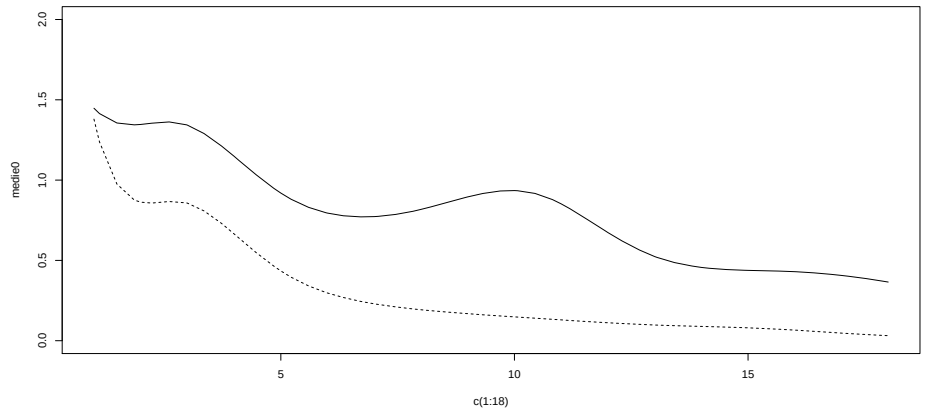


Figure 5.3: Estimated posterior mean trajectory of outgoing video calls (lines) and montly sample means (points) for active customers (solid line and circles) and deacticated customers (dashed line and crosses).

subsample of 2,000 observations equally representing active and deactivated customers. The posterior mean trajectories $\hat{\mu}_0$, $\hat{\mu}_1$ of the number of video calls for the two groups are reported in Figure 5.3.

We predict the deactivations for the test subset of 6,500 customers containing a proportion between active and deactivates equal to that of the original sample. To do so, for each customer-specific observation, we compute the distance d_i^h, with $h = \{0, 1\}$ from the two estimated curves as

$$d_i^h = \sum_s \left(y_i(s) - \hat{\mu}_h(s)\right)^2 . \tag{5.13}$$

We compute the distances in (5.13) and for each of them, based on a 0-1 loss function that classified a trajectory with respect to the minimum distance from the estimated mean trajectories, we compute the classification error rate for the test set. We compared our result with two logistic regressions, the first with only the last month as explanatory variable and the second considering all the 18 months and a logistic generalized additive model with splines smoothing function. The results are reported in Table 5.2.

5.5.2 Transgenic mouse bioassay

To show other scientific contexts in which we can applyed the proposed methodology, in this section we analyze data from a Tg.AC mouse bioassay study of pentaerythritol triacrylate, a chemical used in many industrial processes. Animals are randomized to a control or one of five dose groups

Table 5.2: Classification error rates for churn prediction; RPS, rounded P-splines classification; Logistic, logistic model with one explanatory variable; Logistic (18 var), logistic model with 18 explanatory variables; Logistic GAM (18 var), logistic generalized additive model with 18 explanatory variables

Model	Global error	False positive	False negative
RPS	0.43	0.45	0.39
Logistic	0.76	0.92	0.02
Logistic (18 var)	0.54	0.63	0.15
Logistic GAM (18 var)	0.45	0.50	0.23

each of size 30. The five dose groups are 0.75, 1.5, 3, 6, or 12 mg/kg. The number of skin papillomas on the back of each mouse is counted weekly for 26 weeks and it is of interest to compare the groups to see if there is a significant increase in tumorigenicity relative to control, while trying to find the lowest dose at which there is a significant increase in tumorigenicity and looking for a dose response trend. Dunson and Herring (2005) jointly studied the latency time prior to the first tumor, the increase of papilloma burden and occurrence of internal tumors at the end of the study accommodating dependencies among the outcomes through a Poisson-gamma frailty model. As motivated in Section 5.1-5.2, Poisson hierarchical models are quite restrictive and our focus here is on using the proposed rounded hierarchical P-spline model to improve robustness in estimating the tumor count trajectories for each dose group and assessing dose response trends.

Let $y_i(s)$ denote the number of tumors on mouse i at time s, $y_{it} = y_i(s_{it})$ denote the number at the tth observation time, and $g_i \in \{1, \ldots, G\}$ index the treatment group for mouse i. Then, we assume the same model of (5.12) where ξ_i is a subject-specific random effect. As before the hierarchical structure allows different mean tumor trajectories within each dose group, while allowing certain mice to have greater susceptibility to tumors. Given data sparsity, we avoid allowing the shape of the tumor trajectory to vary across mice within groups.

As a global measure of toxicity of the chemical we consider the average papilloma burden per group. Both the two lower dose groups showed no significant difference from the control group with the posterior mean of the average tumor burden <0.001 and the 95% credible intervals concentrated near zero. In the higher groups the average tumor burden grows with the dose level. Mean tumor burden and 95% credible intervals are 0.18 [0.06,0.39], 9.51 [9.21,9.80] and 12.33 [11.90,12.72] for the 3, 6 and 12 mg/kg dose group respectively. Cumulative tumor burdens along with the dose group-specific empirical means for each week are reported in Figure 5.4.

As a measure of time varying increase in papilloma burden, we computed the mean burden per dose group per week substracting the average number

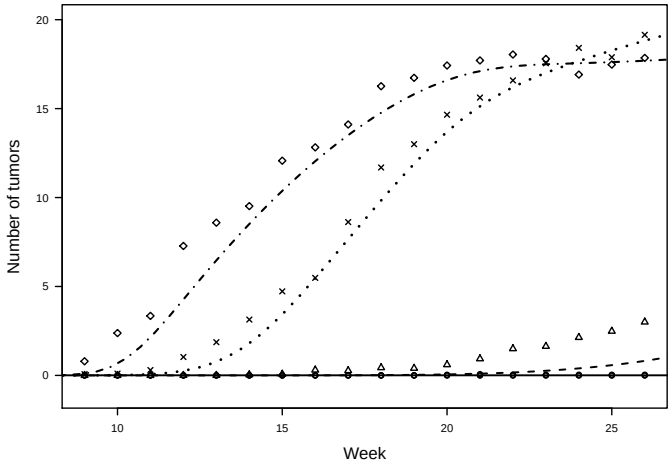


Figure 5.4: Estimated cumulative mean tumor burden (lines) and weekly sample means (points) for the control group, 0.75 mg/kg and 1.5 mg/kg (solid line and circles), 3 mg/kg (dashed line and triangles), 6 mg/kg (dotted line and crosses) and 12 mg/kg (dash-dotted line and squares) dose groups.

for the control group. Posterior means and 95% credible bands are reported in Figure 5.5. The two lower dose groups are indistinguishable from control while the 3, 6 and 12 mg/kg dose groups exhibit clear increases relative to control starting from the 17th, 9th and 8th week, respectively.

Higher dosages lead to higher numbers of skin papillomas, and earlier onset of the first tumor. Our modeling approach allows us to estimate the average time of onset of first tumor, which occurs on the 27th, 14th and 11th week, for the three higher dose groups. In other groups, the typical mouse did not develop tumors prior to the end of the study.

Our overall conclusions agree with those of Dunson and Herring (2005) in terms of differences between groups in latency time and tumor burden, though the estimates differed somewhat. The group comparison results were also consistent with results from a simple frequentist generalized linear model analysis. Similar results are obtained considering the time of development of the first tumor as a summary of the tumor trajectory. As partly illustrated in Figure 5.4, which shows the empirical and estimated mean tumor burdens in each group, the model has a good fit to the data.

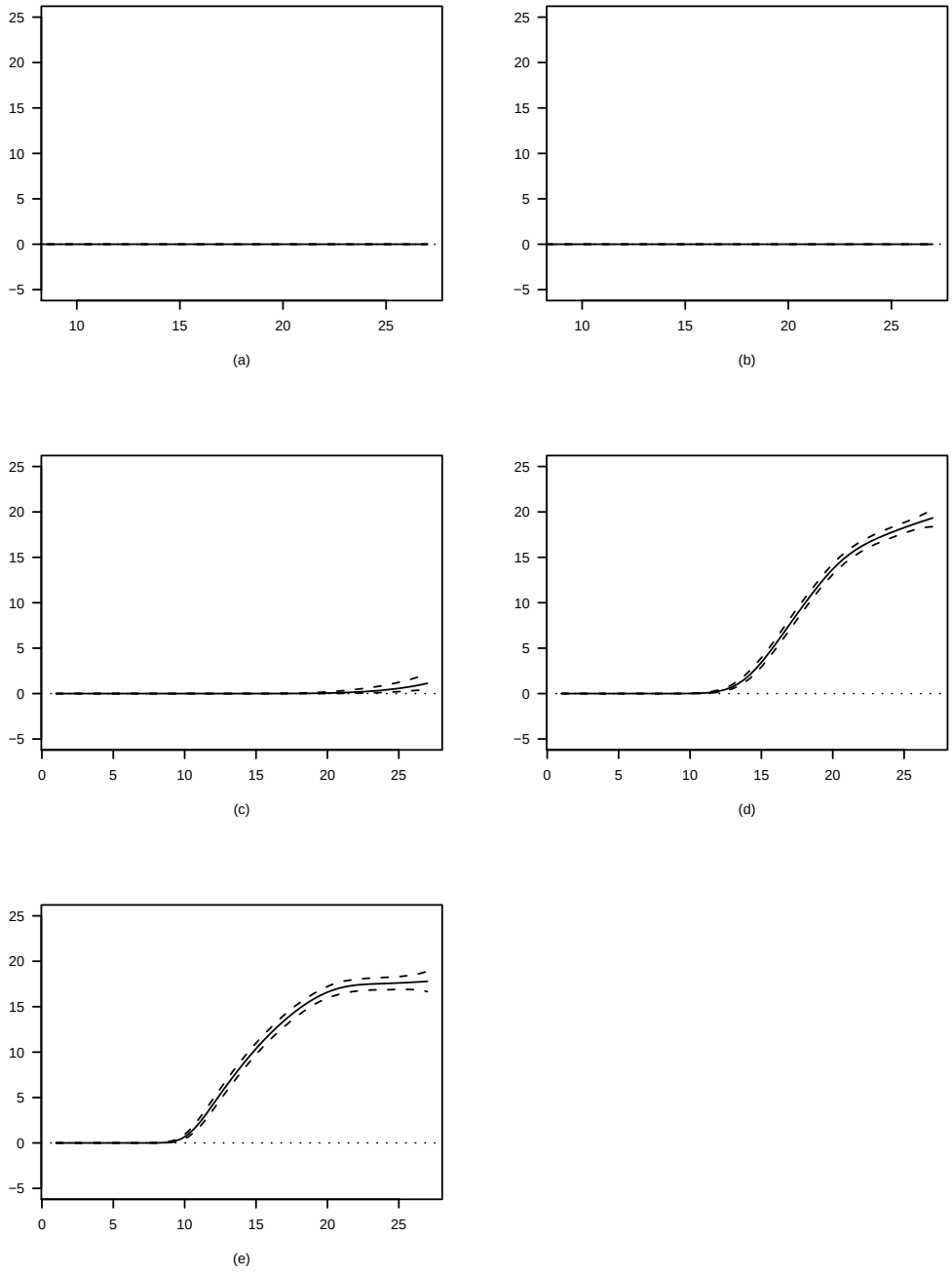


Figure 5.5: Time varying chemical exposure posterior mean effect on tumorigenicity (continuous line) with 95% credible bands (dashed lines) for (a) 0.75 mg/kg, (b) 1.5 mg/kg, (c) 3 mg/kg, (d) 6 mg/kg and (e) 12 mg/kg dose groups. Dotted line at zero corresponds to no effect of the chemical.

Conclusions

Usual parametric models for count data lack flexibility in several key ways, and nonparametric alternatives are scarce or have clear disadvantages. In this work, we introduced and studied a general framework that relates theory and methods developed for continuous data to the count context.

The main contribution of this thesis consists in introducing the idea of rounding continuous objects to induce probability measures on spaces related to count data. This idea has been applied in several contexts and for each case it has been widely studied showing several advantages and ease of implementation in a broad variety of applications.

To give a convincing and comprehensive presentation of the methodology we followed Ferguson (1973) showing that his three desiderata for a Bayesian nonparametric procedure are satisfied. First for each new prior distribution we showed its easy interpretation. Then, we described the size of the prior support while also studying Bayesian asymptotic properties such as posterior consistency. Finally we showed that practical implementation is easy and reliable, supporting every method with simulation studies and applications to real data.

From the applied point of view, we showed that the methodology can be used in usual customer base management problems such as prediction of traffic variables, or churn forecast. For this latter setting the idea of having a model that take in consideration the longitudinal nature of the measurements of traffic usage is relatively new. As an extension of the model proposed in Chapter 5, one could jointly model the traffic variables and the probability of churn, for example with logit or probit models with count functional predictors. Similar ideas are used in biostatistics but deal with continuous functional predictors (Dunson et al., 2008; Bigelow and Dunson, 2009). A further extension consists in jointly consider multivariate mixed scale functional data. Even if our methodology could be directly applied in those settings, the computational cost will be heavy and hence future research in this direction could focus on fast and approximate methods for posterior computation.

In addition to the customer base management setting we showed that the methods can be applied in other scientific contexts, focusing in toxicity and developmental toxicity. Particularly when usual parametric assumptions are not satisfied our methodology has clear advantages.

Bibliography

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, *88*(422), 669–679.

Arellano-Valle, R. B. and Azzalini, A. (2006). On the Unification of Families of Skew-normal Distributions. *Scandinavian Journal of Statistics*, *33*(3), 561–574.

Arellano-Valle, R. B., Genton, M., and Loschi, R. H. (2009). Shape mixtures of multivariate skew-normal distributions. *Journal of Multivariate Analysis*, 100, 97–101.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.*, 12, 171–178.

Azzalini, A. (1986). Further results on a class of distributions which includes the normal ones. *Statistica*, 46, 199–208.

Azzalini, A. (2005). The Skew-normal Distribution and Related Multivariate Families. *Scand. J. Statist.*, *32*(2), 159–188.

Azzalini, A. and Scarpa, B. (2012). *Data Analysis and Data Mining: An Introduction*. Oxford: Oxford University Press.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian Predictive Process Models for Large Spatial Data Sets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *70*(4), 825–848.

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The Consistency of Posterior Distributions in Nonparametric Problems. *The Annals of Statistics*, *27*(2), 536–561.

Behseta, S., Kass, R. E., and Wallstrom, G. L. (2005). Hierarchical models for assessing variability among functions. *Biometrika*, *92*(2), 419–434.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35, 99–109.

Bigelow, J. and Dunson, D. (2007). Bayesian Adaptive Regression Splines for Hierarchical Data. *Biometrics*, 63, 724–732.

Bigelow, J. and Dunson, D. (2009). Bayesian Semiparametric Joint Models for Functional Predictors. *Journal of the American Statistical Association*, 104, 26–36.

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

Canale, A. and Dunson, D. B. (2011). Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, *106*(496), 1528–1539.

Carota, C. and Parmigiani, G. (2002). Semiparametric Regression for Count Data. *Biometrika*, *89*(2), 265–281.

Cavatti Vieira, C. (2011). 'Estimação de Densidades via Mistura de Distribuições skew-normal por processos de Dirichlet'. Master's thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D. B., and Carin, L. (2010). Compressive Sensing on Manifolds Using a Nonparametric Mixture of Factor Analyzers: Algorithm and Performance Bounds. *IEEE Transaction in Signal Processes*, *58*(12), 6140–6155.

Choi, T. and Schervish, M. J. (2007). On Posterior Consistency in Nonparametric Regression Problems. *Journal of Multivariate Analysis*, *98*(10), 1969–1987.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *The Annals of Statistics*, 14, 1–67.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56, 363–375.

Doob, J. L. (1949). *Application of the theory of martingales*. Technical report, Coll. Int. du C. N. R. S., Paris. pages 2327.

Dunson, D. B. (2000). Bayesian Latent Variable Models for Clustered Mixed Outcomes. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *62*(2), 355–366.

Dunson, D. B. (2003). Dynamic Latent Trait Models for Multidimensional Longitudinal Data. *Journal of the American Statistical Association*, *98*(463), 555–563.

Dunson, D. B. (2010). Multivariate Kernel Partition Process Mixtures. *Statistica Sinica*, *20*(4), 1395–1422.

Dunson, D. B. and Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics*, *6*(1), 11–25.

Dunson, D. B., Herring, A. H., and Siega-Ritz, A. (2008). Bayesian inference on changes in response densities over predictor clusters. *Journal of the American Statistical Association*, 103, 1508–1517.

Efromovich, S. (2011). Nonparametric estimation of the anisotropic probability density of mixed variables. *Journal of Multivariate Analysis*, *102*(3), 468–481.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11, 89–121.

Escobar, M. D. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association*, 90, 577–588.

Eurostat (2011). Telecommunication statistics. Statistics Explained, 2011/4/3. Retrived 13/4/2011 from `epp.eurostat.ec.europa.eu/statistics_explained/index.php/Telecommunication_statistics`.

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1, 209–230.

Ferguson, T. S. (1974). Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics*, 2, 615–629.

Frühwirth-Schnatter, S. and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modeling. *Biometrika*, *93*(4), 827–841.

Frühwirth-Shnatter, S. and Pyne, S. (2010). Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11, 317–336.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-based Approaches to the Calculation of Posterior Moments. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4*. Oxford: Oxford University Press.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior Consistency of Dirichlet Mixtures in Density Estimation. *The Annals of Statistics*, *27*(1), 143–158.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, *28*(2), 500–531.

Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, *34*(5), 2413–2429.

Ghosal, S. and van der Vaart, A. W. (2001). Entropies and Rates of Convergence for Maximum Likelihood and Bayes Estimation for Mixtures of Normal Densities. *The Annals of Statistics*, *29*(5), 1233–1263.

Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, *35*(5), 697–723.

Ghosh, J. K. and Ramamoorthi, R. (2003). *Bayesian Nonparametric*. Springer, New York.

Ghurye, S. G. (1968). Information and sufficient sub-fields. *The Annals of Mathematical Statistics*, *39*(6), 2056–2066.

Gill, J. and Casella, G. (2009). Nonparametric Priors for Ordinal Bayesian Social Science Models: Specification and Estimation. *Journal of the American Statistical Association*, *104*(486), 453–454.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of American Statistical Association*, *99*(468), 1015–1026.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag Inc.

Hjort, N. L. e., Holmes, C. e., Müller, P. e., and Walker, S. G. e. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *Ann. Appl. Statist.*, *1*(1), 265–283.

Ishwaran, H. and James, Lancelot, F. (2001). Gibbs Sampling Methods for Stick Breaking Priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

Jara, A., Garcia-Zattera, M., and Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis*, *51*(11), 5402–5415.

Jasra, A., Holmes, C., and Stephens, D. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20, 50–67.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis*, *51*(5), 2542 – 2558.

Karlis, D. and Xekalaki, E. (2005). Mixed Poisson Distributions. *International Statistical Review*, *73*(1), 35–58.

Kottas, A., Müller, P., and Quintana, F. (2005). Nonparametric Bayesian Modeling for Multivariate Ordinal Data. *Journal of Computational and Graphical Statistics*, *14*(3), 610–625.

Krnjajic, M., Kottas, A., and Draper, D. (2008). Parametric and nonparametric Bayesian model specification: A case study involving models for count data. *Computational Statistics & Data Analysis*, *52*(4), 2110 – 2128.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13, 183–212.

Lehmann, D., Gupta, S., and Steckel, J. (1998). *Marketing research*. Addison-Wesley.

Li, E., Zhang, D., and Davidian, M. (2004). Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics*, 60, 1–7.

Li, Q. and Racine, J. (2003). Nonparametric estimation of distributions with categorical and continuous data. *Journal of Multivariate Analysis*, *86*(2), 266–292.

Li, Q. and Racine, J. (2008). Nonparametric Estimation of Conditional CDF and Quantile Functions With Mixed Categorical and Continuous Data. *Journal of Business and Economic Statistics*, *26*(4), 423–434.

Lin, T. I., Lee, J. C., and Yen, S. Y. (2007). Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17, 909–927.

Liseo, B. and Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical planning and inference*, 136, 373–389.

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates: I. Density Estimates. *The Annals of Statistics*, 12, 351–357.

Medvedovic, M. and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics*, *18*(9), 1194–1206.

Meligkotsidou, L. (2007). Bayesian Multivariate Poisson Mixtures with an Unknown Number of Components. *Statistics and Computing*, *17*(2), 93–107.

MIIT (2009). Statistical communique of China on the 2008 development of the telecom industry. Retrived 15/4/2011 from `http://www.miit.gov.cn/n11293472/n11295057/n11298508/11979497.html`.

Morris, J. and Carroll, R. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *68*(Part 2), 179–199.

Moustaki, I. and Knott, M. (2000). Generalized Latent Trait Models. *Psychometrika*, *65*(3), 391–411.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian Curve Fitting Using Multivariate Normal Mixtures. *Biometrika*, 83, 67–79.

Murray, I. and Adams, R. (2010). Slice sampling covariance hyperparameters in latent Gaussian models. *Advances in Neural Information Processing Systems*, to appear.

Nath, S. and Behara, R. (2003). Customer churn Analysis in the Wireless Industry: A Data Mining Approach. In *Proceedings of the 34th meeting of the Decision Sciences Institute*.

Neal, R. M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

Nikoloulopoulos, A. K. and Karlis, D. (2010). Modeling multivariate count data using copulas. *Communications in Statistics, Simulation and Computation*, 393, 172–187.

O'Hagan, A. (1978). Curve Fitting and Optimal Design for Prediction (with discussion). *Journal of the Royal Statistical Society, Series B: Methodological*, 40, 1–42.

Ouyang, D., Li, Q., and Racine, J. (2006). Cross-validation and the estimation of probability distributions with categorical data. *Journal of Nonparametric Statistics*, *18*(1), 69–100.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985). The developemental toxicity of ethylene glycol in rats and mice. *Toxicological and Applied Pharmacology*, 81, 113–127.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts.

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). Latent stick-breaking processes. *Journal of the American Statistical Association*, *105*(490), 647–659.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *71*(2), 319–392.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent Variable Models for Mixed Discrete and Continuous Outcomes. *Journal of the Royal Statistical Society. Series B (Methodological)*, *59*(3), pp. 667–678.

Scarpa, B. and Dunson, D. B. (2009). Bayesian Hierarchical Functional Data Analysis Via Contaminated Informative Priors. *Biometrics*, 65, 772–780.

Scarpa, B. and Dunson, D. B. (2011). Enriched Stick Breaking Processes for Functional Data. submitted.

Schwartz, L. (1965). On Bayes Procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4, 10–26.

Scricciolo, C. (2011). Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics*, 5, 270–308.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4, 639–650.

Shen, W. and Ghosal, S. (2011). *Adaptive Bayesian multivariate density estimation with Dirichlet mixtures*. Technical report, arXiv:1109.6406v1.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal Of The Royal Statistical Society Series C*, *54*(1), 127–142.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 62, 795–809.

Strouse, K. G. (2004). *Customer-centered telecommunications services marketing*. Artech House Inc.

Thompson, W. and Rosen, O. (2007). A Bayesian Model for Sparse Functional Data. *Biometrics*, 64, 54–63.

Tokdar, S. T. (2006). Posterior Consistency of Dirichlet Location-scale Mixture of Normals in Density Estimation and Regression. *Sankhya*, *68*(1), 90–110.

van der Vaart, A. W. and van Zanten, J. H. (2009). RAdaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, *37*(5B), 2655–2675.

Wahba, G. (1978). Imroper priors, spline smoothing and the problems of guarding agains model errors in regression. *Journal of the Royal Statistical Society, Series B: Methodological*, 40, 364–372.

Walker, S. G. (2007). Sampling the Dirichlet Mixture Model with Slices. *Communications in Statistics, Simulation and Computation*, 34, 45–54.

Walker, S. G., Lijoi, A., and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *The Annals of Statistics*, 35, 738–746.

Wang, C., Wang, N., and Wang, S. (2000). Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56, 487–495.

Wedel, M., Böckenholt, U., and Kamakura, W. A. (2003). Factor Models for Multivariate Count Data. *Journal of Multivariate Analysis*, *87*(2), 356–369.

Wilson, A. and Ghahramani, Z. (2010). Copula Processes. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds.), *Advances in Neural Information Processing Systems 23*, pages 2460–2468.

Wu, Y. and Ghosal, S. (2008). Kullback Leibler Property of Kernel Mixture Priors in Bayesian Density Estimation. *Electronic Journal of Statistics*, 2, 298–331.

Wu, Y. and Ghosal, S. (2010). The L_1-consistency of Dirichlet mixtures in multivariate B ayesian density estimation. *Journal of Multivariate Analysis*, 101, 2411–2419.

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2010). Bayesian non parametric Hidden Markov Models with applications in genomics. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, to appear.

Zhu, H., Williams, C. K. I., Rohwer, R. J., and Morciniec, M. (1998). Gaussian Regression and Optimal Finite Dimensional Linear Models. In Bishop, C. (ed.), *Neural Networks and Machine Learning*. Berlin: Springer-Verlag.

Antonio Canale

CURRICULUM VITAE

Personal Details

Date of Birth: February 27, 1984
Place of Birth: Padova, Italy
Nationality: Italian

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4111
e-mail: canale@stat.unipd.it

Current Position

Since January 2009; (expected completion: January 2012)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Bayesian nonparametric models for count data with application to customer base management
Supervisor: Bruno Scarpa
Co-supervisor: David B. Dunson

Research interests

- Bayesian nonparametric;
- flexible distributions;
- data mining;
- business statistics;

Education

September 2007 – July 2008
Master (*laurea specialistica*) degree in Statistics and Computer Science.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Statistical aspects in the extended skew normal distribution" (in italian)
Supervisor: Prof. Adelchi Azzalini
Final mark: 110 *cum laude*

September 2004 – February 2007
Bachelor degree (*laurea triennale*) in Statistics and Management.
University of Padova, Faculty of Statistical Sciences
Title of dissertation: "Branding strategies in italian safety footware market: USS safety system casestudy." (in italian)
Supervisor: Prof. Rita Zillo
Final mark: 110 *cum laude*.

Visiting periods

July 2009 – August 2009
McMaster University
Hamilton, Ontario, Canada.
Supervisor: Prof. N. Balakrishnan

February 2010 – December 2010
Duke University
Durham, North Carolina, USA.
Supervisor: Prof. David B. Dunson

July 2011 – August 2011
Duke University
Durham, North Carolina, USA.
Supervisor: Prof. David B. Dunson

Further education

June 2010
Course name: ABS10 (Applied Bayesian statistics Summer school)
Organizing Institution: CNR-IMATI, EURAC
Organizer: Guido Consonni, Fabrizio Ruggeri
Instructor: David B. Dunson (Duke University)

Work experience

Januaruy 2007 – now
Freelance Statistical consultant.

Partners: CUOA business school, Altavilla vicentina, Vicenza (Lecturer), Ecomatica, Padova (Analyst, BI developer), Nordest business consulting, Vicenza (Analyst, BI developer, CRM expert).

Awards and Scholarship

2009
PhD scholarship (University of Padova)

July 2011
Young researcher travel award for the 8th BNP workshop, Vercruz, Mexico. (International Society for Bayesian Analysis).

September 2007
BEST2007 (Bologna Experience for superior talents). Granted summer school with the 50 best italian graduates (Alma graduate school).

Computer skills

- Operative System: Linux, Windows, OSX
- Programming: C, R, Matlab
- Markup Languages: LaTeX, HTML
- Other skills: Parellel computing

Language skills

Italian: native; English: fluent;

Publications

Articles in peer reviewed journals

Canale, A. and Dunson, D. B. (2011), *Bayesian Kernel Mixtures for Counts*, Journal of American Statistical Association, 106, 1528–1539

Canale, A., *Statistical aspects in the extended skew-normal model*, Metron, in press

Articles submitted

Canale, A. and Dunson, D. B., *Nonparametric Bayes modeling of count processes*, 2011

Canale, A. and Dunson, D. B., *Bayesian multivariate mixed-scale density estimation*, 2011

Canale, A. and Scarpa, B., *Dirichlet process mixtures of rounded skew-normal kernels*, 2011

Conference presentations

Canale, A. (2011), Bayesian rounded stochastic count processes; *Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction*, Padova, Italy, September 19 -21, 2011 (contributed talk)

Canale, A. (2011), Bayesian Kernel mixtures for counts; *SAHD workshop*, Duke University, Durham, NC, USA, July 26-28, 2011 (contributed poster)

Canale, A. (2011), Bayesian Kernel mixtures for counts; *8th workshop on Bayesian nonparametric*, Veracruz, Mexico, June 26 -30, 2011 (contributed poster)

Canale, A. (2011), Dirichlet process mixtures of rounded skew normal distributions; *IV Skew Workshop*, Santiago del Chile, Chile, May 16 -19, 2011 (invited talk)

Canale, A. (2010),Bayesian nonparametric mixtures of rounded kernels; *Applied Bayesian Summer School*, Bolzano, Italy, June 11-15 2010 (participants talk)

Canale, A. (2009), Some Inferential results on the extended skew-normal

distribution in the scalar case; *Summer mini conference in Statistic and probability*, McMaster University, Hamilton, Ontario, Canada, July 31, 2009 (invited talk)

Canale, A. (2009), Statistical aspects in the extended skew normal model; *Workshop on Skew Symmetric Distributions*, Benevento, Italy, March 16 - 20 2009 (contributed talk)

Canale, A. (2008), Some preliminar results on the information matrix of the extended skew normal model; *Workshop on Skew Symmetric Probability Distributions*, Bertinoro (FC), Italy, April 6 - 10 2008 (contributed talk)

Teaching experience

June 2011
Course name Analisi dei dati (Data Mining)
Teaching task (lab), total number of hours: 4
Institution: Facoltà di Scienze Statistiche, University of Padua
Instructor: Prof. Bruno Scarpa

Other Interests

Music, travelling, hiking, home brewing;

References

Prof. N. Balakrishnan
McMaster University
Department of Mathematics and Statistics, Hamilton, Ontario
Phone: (905)525-9140
e-mail: bala@mcmaster.ca

Prof. David B. Dunson
Duke University
Department of statistical science
Durham, NC 27708-0251
Phone: (919)684-8025
e-mail: dunson@stat.duke.edu

Prof. Bruno Scarpa
Università degli studi di Padova
Dipartimento di scienze statistiche,
Padua, Italy
Phone: (0039) 049 8274193
e-mail: scarpa@stat.unipd.it