# A Bayesian nonparametric model for data on different scales of measure; an application to customer base management of telecommunications companies.

# A Bayesian nonparametric model for data on different scale of measure; an application to customer base management of telecommunications companies.

Antonio Canale* and David B. Dunson

**Abstract** To analyze telecommunications marketing data which are usually made of discrete and continuous observations we consider a general framework to jointly model continuous, count and categorical variables under a nonparametric prior, which is induced through rounding latent variables having an unknown density with respect to Lesbesgue measure. For the proposed class of priors large support, strong consistency and rates of posterior contraction can be proved. The approach is applied to model the joint density of traffic data for a portion of customers of a European mobile phone operator.

**Key words:** Mixed discrete and continuous; Nonparametric regression; Zero-inflated models

## 1 Introduction

Telecommunications companies store plenty of informations about their customer behaviors. Such variables include continuous, counts, and binary variables. The most common approach to model such data is to link each observed variable to latent Gaussian variables. Relationships among the underlying Gaussian variables are typically characterized through latent factor or structural equation models as in Muthén (1984). Although the underlying Gaussian class is computationally convenient and mathematical tractable, the flexibility is limited in implying Gaussian distributions for continuous variables, and probit models for categorical variables. In addition, issues arise in modeling counts and categorical variables having very many levels due to the need to introduce and do computation for very many threshold parameters. A different class of models can be obtained defining a separate generalized linear

* Antonio Canale
University of Turin and Collegio Carlo Alberto, e-mail: antonio.canale@unito.it

model for each variable, with shared latent variables to induce dependence (Sammel et al., 1997; Moustaki and Knott, 2000; Dunson, 2000, 2003). This framework assumes that observed variables are independently drawn from distributions in the exponential family conditionally on latent variables.

In this paper we discuss classes of Bayesian models for mixed scale densities, which are computationally convenient and can be shown to have appealing theoretical properties, such as large support, posterior consistency and near optimal rates of convergence. Particularly, we focus on a multivariate mixed scale generalization of the rounding framework of Canale and Dunson (2011a) already discussed in Canale and Dunson (2011b).

## 2 Mixed-scale densities

Let $y = (y_1^T, y_2^T)^T$, where $y_1 = (y_{1,1}, \dots, y_{1,p_1})$ is a $p_1 \times 1$ vector of continuous observations and $y_2 = (y_{2,p_1+1}, \dots, y_{2,p})$ is a $p_2 \times 1$ vector of discrete variables, be is a $p \times 1$ vector of variables having mixed measurement scales. We let $y \sim f$, with $f$ denoting the joint density with respect to an appropriate product measure $\mu$.

To induce a prior $f \sim \mathit{\Pi}$ for the density of the mixed scale variables, we let

$$y = h(y^*), \quad y^* \sim f^*, \quad f^* \sim \mathit{\Pi}^*, \tag{1}$$

where $h : \mathbb{R}^p \to \mathit{\Omega}$, $y^* = (y_1^*, \dots, y_p^*)^T \in \mathbb{R}^p$, $f \in \mathcal{F}^*$, $\mathcal{F}^*$ is the set of densities with respect to Lesbesgue measure over $\mathbb{R}^p$, and $\mathit{\Pi}^*$ is a prior over $\mathcal{F}^*$. The mapping function $h$ is defined as $h(y^*) = \left\{h_1(y_1^*)^T, h_2(y_2^*)^T\right\}^T$, where $h_1(y_1^*) = y_1^*$ is the identity function and $h_2$ are thresholding functions. Let $A^{(j)} = \{A_1^{(j)}, \dots, A_{q_j}^{(j)}\}$ denote a prespecified partition of $\mathbb{R}$ into $q_j$ mutually exclusive subsets, for $j = 1, \dots, p_2$, with the subsets ordered so that $A_h^{(j)}$ is placed before $A_l^{(j)}$ for all $h < l$. Then, letting $A_{y_2} = \{y_2^* : y_{2,j}^* \in A_{y_{2j}}^{(j)}, j = 1, \dots, p_2\}$, the mixed scale density $f$ is defined as

$$f(y) = g(f^*) = \int_{A_{y_2}} f^*(y^*) dy^*. \tag{2}$$

The function $g : \mathcal{F}^* \to \mathcal{F}$ defined in (2) is a mapping from the space of densities with respect to Lesbesgue measure on $\mathbb{R}^p$ to the space of mixed-scale densities $\mathcal{F}$.

This framework generalizes Canale and Dunson (2011a), which focused only on count variables. The theory is substantially more challenging in the mixed scale case when there are continuous variables involved. Clearly the properties of the induced prior $f \sim \mathit{\Pi}$ will be driven largely by the properties of $f^* \sim \mathit{\Pi}^*$. The study of asymptotic properties of the induced prior are omitted here but can be found in Canale and Dunson (2011b). Such properties are in terms of large support, strong posterior consistency and rate of posterior contraction and are based in studying the topology induced by the mapping functions $g$ and $h$ in the space $\mathcal{F}$.

## 3 Marketing application

### *3.1 Customer base management of telecommunications companies*

A key aspect in customer base management of telecommunications companies is to design contractual plans that match customer needs. For example, a mobile phone contractual plan can have or not a connection charge, can include or not some minutes free of charge for each day, while the time slot can be of minutes, half-minutes or actual seconds used. Furthermore, the text messages can be included in the subscription or can be payed one by one.

In order to understand the usual customers' usage behaviors, it can be of interest to study the joint distribution of traffic variables, including number and duration of outgoing and incoming calls and number of text messages sent. With such information available, marketing managers can conceive suitable contractual plans that merge customers' needs and expectations with marketing and sales goals.

In the next section we analyze a real dataset using the model described in Section 2 to estimate the joint probability distribution of some traffic variables.

### *3.2 A real application*

Consider the data on 1,000 SIM (Subscriber Identification Module) cards of a prepayed European mobile phone operator in a given month. For each SIM card, we have several monthly traffic variables including total number and duration of outgoing and incoming calls, video-calls, and the total number of text messages sent. In the following we try to model the joint distribution of two of the most interesting variables: the outgoing duration of calls, after a log transformation ($y_1$) and the total number of text messages sent ($y_2$). This joint distribution is made of continuous and count marginals and has $p_1 = 1$ and $p_2 = 1$.

To model the latter distribution we use the model described in Section 2, assuming that $\Pi^*$ is a Dirichlet process mixtures of Gaussian kernels (Lo, 1984; Escobar and West, 1995). Prior hyperparameters elicitation follows usual empirical Bayes approach. The partition $A_{y_2}$ in the case $p_2 = 1$ is simply induced by a sequence of thresholds, which are assumed to be $(-\infty, 1, 2, \dots)$.

We run a Markov Chain Monte Carlo algorithm for 4,000 iteration, discarding the first 2,000 as burn in. In addition to the values of the latent parameters we collect, for a fine grid of points, the values of the joint density and of the marginals. Their posterior means are reported in Figure 1. Our method naturally accounts for the zero inflation typically present in this kind of dataset, as can be seen from panel (c) of Figure 1. Indeed, the portion of SIM cards with $y_2 = 0$ is of 73.7%. While giving a full characterization of the joint distribution of the data, from which we can obtain the marginals or any conditional distribution, the model can be used also as a tool for nonparametric regression in the spirit of Müller et al. (1996).

## References

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

Canale, A. and Dunson, D. B. (2011b). Bayesian multivariate mixed-scale density estimation. Technical report, arXiv:1110.1265.

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.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of Amer. Stat. Association*, 90:577–588.

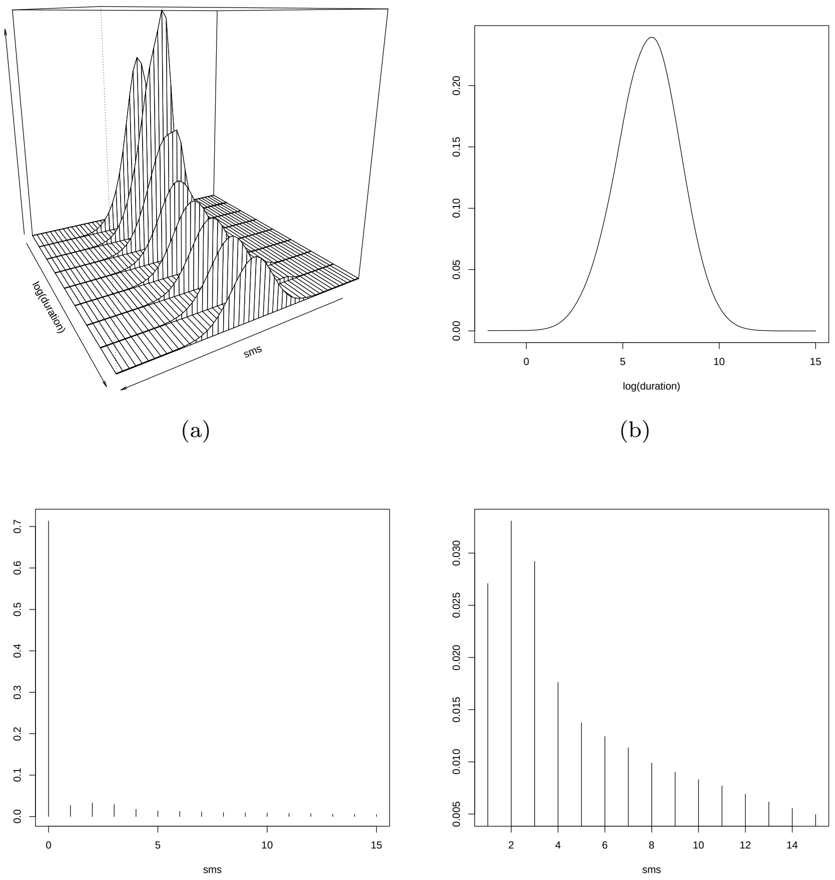
Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The Annals of Statistics*, 12(1):351–357.

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.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, 49(1):115–132.

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.



(a)

(b)

(c)

(d)

**Fig. 1** Joint mixed-scale density plotted for $y_2 > 0$ (a), marginal density for $y_1$ (b), marginal probability mass function for $y_2$ (c), and marginal probability mass function for $y_2 > 0$ (d).