ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A generic approach to nonparametric function estimation with mixed data

Thomas Nagler

Department of Mathematics, Technical University of Munich, Boltzmanstraße 3, 85748 Garching, Germany

ARTICLE INFO

Article history:
Received 25 August 2017
Received in revised form 1 February 2018
Accepted 13 February 2018
Available online xxxx

Keywords:
Density
Discrete
Jitter
Mixed data
Nonparametric
Regression

ABSTRACT

Most nonparametric function estimators can only handle continuous data. We show that making discrete variables continuous by adding noise is justified under suitable conditions on the noise distribution. This principle is widely applicable, including density and regression function estimation.

1. Introduction

In applications of statistics, data containing discrete variables are omnipresent. An online retailer records information on how many purchases a customer made in the past. Social scientists typically use discrete scales on which study participants rate their satisfaction, attitude, or feelings. Another common example is where data describe unordered categories, like gender or business sectors.

Suppose that $(\boldsymbol{Z}, \boldsymbol{X})$ is a random vector with discrete component $\boldsymbol{Z} \in \mathbb{Z}^p$ and continuous component $\boldsymbol{X} \in \mathbb{R}^q$. This includes the cases $p \geq 1, q = 0$ (all variables are discrete) and $p = 0, q \geq 1$ (all variables are continuous). We consider problems where one aims at estimating a functional T of the density/probability mass function $f_{\boldsymbol{Z},\boldsymbol{X}}$ based on observations $(\boldsymbol{Z}_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. This formulation is general enough to include many common problems in nonparametric function estimation, in particular: density estimation, regression, and classification.

Some nonparametric estimation techniques have been specifically designed to allow for mixed continuous and discrete data (Ahmad and Cerrito, 1994; Li and Racine, 2003; Hall et al., 1983; Efromovich, 2011), but the number is small and the more sophisticated methods are often developed in a purely continuous framework. Examples are local polynomial methods (Fan and Gijbels, 1996; Loader, 1999) or copula-based estimators (e.g., Otneim and Tjøstheim, 2016; Nagler and Czado, 2016; Kauermann and Schellhase, 2014). These methods are no longer consistent when applied to mixed data types.

There is a popular trick among practitioners to get an approximate answer nevertheless: just make the data continuous by adding noise to each discrete variable. This trick is sometimes called *jittering* or *adding jitter*. Examples where it has been successfully applied are: avoiding overplotting in data visualization (Few, 2008), adding intentional bias to complex machine learning models (Zur et al., 2004), deriving theoretical properties of concordance measures (Denuit and Lambert, 2005), or nonparametric copula estimation for mixed data (Genest et al., 2017). An example of its misuse was pointed out by Nikoloulopoulos (2013) in the context of parametric copula models. Generally, the trick lacks theoretical justification because it can introduce bias. But we shall see that this issue is resolved under a suitable choice of noise distribution.

E-mail address: mail@tnagler.com.

https://doi.org/10.1016/j.spl.2018.02.040
0167-7152/

Please cite this article in press as: Nagler T., A generic approach to nonparametric function estimation with mixed data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.040.

This letter aims to formalize this trick and to provide a starting point for a more nuanced investigation of its properties. Some open questions and partial answers will be given at the end.

2. Jittering mixed data

2.1. Preliminaries and notation

We assume throughout that all random variables live in a space with a natural concept of ordering. Unordered categorical variables can always be coded into a set of binary dummy variables (for which $0 < 1$ gives a natural ordering). We further assume without loss of generality that any discrete random variable, say Z, is supported on a set $\Omega_Z \subseteq \mathbb{Z}$. For any continuous random vector $\boldsymbol{X}$, we write $f_{\boldsymbol{X}}$ for its joint density. In case $\boldsymbol{Z}$ is a discrete random vector, $f_{\boldsymbol{Z}}$ denotes its density with respect to the counting measure, i.e., $f_{\boldsymbol{Z}}(\boldsymbol{z}) = \Pr(\boldsymbol{Z} = \boldsymbol{z})$. A random vector with mixed types will be partitioned into $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$. Then $f_{\boldsymbol{Z},\boldsymbol{X}}$ is the density with respect to the product of the counting and Lebesgue measures,

$$f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = \frac{\partial^q}{\partial x_1 \cdots \partial x_q} \Pr\big(\boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{X} \le \boldsymbol{x}\big).$$

2.2. The density of a jittered random vector

The jittered version of a random vector is defined by adding noise to all discrete variables.

Definition 1. Let η be a bounded density function that is continuous almost everywhere on $\mathbb{R}$. The jittered version of the random vector $(\boldsymbol{Z}, \boldsymbol{X}) \in \mathbb{Z}^p \times \mathbb{R}^q$ is defined as $(\boldsymbol{Z} + \boldsymbol{\epsilon}, \boldsymbol{X})$, where $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is independent of $(\boldsymbol{Z}, \boldsymbol{X})$.

Provided that $f_{\boldsymbol{Z},\boldsymbol{X}}$ exists, the density of the jittered vector $(\boldsymbol{Z} + \boldsymbol{\epsilon}, \boldsymbol{X})$ is simply the discrete–continuous convolution of $f_{\boldsymbol{Z},\boldsymbol{X}}$ and the noise density $f_{\boldsymbol{\epsilon}}$:

$$f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = \sum_{\boldsymbol{z}' \in \mathbb{Z}^p} f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}', \boldsymbol{x}) f_{\boldsymbol{\epsilon}}(\boldsymbol{z} - \boldsymbol{z}'), \quad \text{for almost all } (\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{R}^{p+q}.$$

We observe a close relationship between the densities $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ and $f_{\boldsymbol{Z},\boldsymbol{X}}$. If we know $f_{\boldsymbol{Z},\boldsymbol{X}}$ at all values $(\boldsymbol{z}', \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, we can immediately compute $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ at all values $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{R}^{p \times q}$. The other direction is more interesting for our purposes: can we recover $f_{\boldsymbol{Z},\boldsymbol{X}}$ from known values of $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$? In general, this poses a rather challenging deconvolution problem. But we can make things easier by a suitable choice of noise density η. In fact, there is a large class of noise densities for which no deconvolution is necessary and $f_{\boldsymbol{Z},\boldsymbol{X}}$ and $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$ coincide on $\mathbb{Z}^p \times \mathbb{R}^q$.

Proposition 1. *It holds*

$$f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) = f_{\boldsymbol{Z},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x}) \tag{1}$$

for any joint density $f_{\boldsymbol{Z},\boldsymbol{X}}$ and all $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, if and only if the following two conditions are satisfied:

1. $f_{\boldsymbol{\epsilon}}(\boldsymbol{0}) = 1$,
2. *there exists $\gamma_2 \in (0, 1)$ such that $f_{\boldsymbol{\epsilon}}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathbb{R}^p \setminus [-\gamma_2, \gamma_2]^p$.*

A simple, but powerful implication is that we can estimate the discrete–continuous density $f_{\boldsymbol{Z},\boldsymbol{X}}$ by estimating the purely continuous density $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}$.

2.3. A convenient class of noise distributions

In the following we give a particularly convenient class of noise densities.

Definition 2. We say that $f_{\boldsymbol{\epsilon}} \in \mathcal{E}_{\gamma_1,\gamma_2}$ for some $0 < \gamma_1 \le 0.5 \le \gamma_2 < 1$, if $f_{\boldsymbol{\epsilon}}(\boldsymbol{x}) = \prod_{j=1}^p \eta(x_p)$ for all $\boldsymbol{x} \in \mathbb{R}^p$, where η is an absolutely continuous probability density function, $\eta(x) = 1$ for all $x \in [-\gamma_1, \gamma_1]$, and $\eta(x) = 0$ for all $x \in \mathbb{R} \setminus (-\gamma_2, \gamma_2)$.

The class $\mathcal{E}_{\gamma_1,\gamma_2}$ satisfies (1), but adds two more restrictions to the conditions given in Proposition 1: (i) the random noise is componentwise independent, (ii) it is constant in a neighborhood of zero. The first restriction is made purely for convenience and will be discussed further in Section 4.2. The second ensures that the derivatives of $f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})$ with respect to $\boldsymbol{z}$ vanish for all $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$. This property is particularly useful in nonparametric density estimation, since an estimators' bias is usually proportional to derivatives of the target density.

Proposition 2. *If $f_{\boldsymbol{\epsilon}} \in \mathcal{E}_{\gamma_1,\gamma_2}$, $(\boldsymbol{z}, \boldsymbol{x}) \in \mathbb{Z}^p \times \mathbb{R}^q$, and $\boldsymbol{m} \in \mathbb{N}^p$ such that $\sum_{k=1}^p m_k = \overline{m}$, then*

$$\frac{\partial^{\overline{m}} f_{\boldsymbol{Z}+\boldsymbol{\epsilon},\boldsymbol{X}}(\boldsymbol{z}, \boldsymbol{x})}{\partial z_1^{m_1} \cdots \partial z_p^{m_p}} = 0.$$

Please cite this article in press as: Nagler T., A generic approach to nonparametric function estimation with mixed data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.040.