Journal of Multivariate Analysis 102 (2011) 468–481

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Nonparametric estimation of the anisotropic probability density of mixed variables

Sam Efromovich

Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75083-0688, United States

ARTICLE INFO

Article history:
Received 22 February 2010
Available online 16 October 2010

AMS 2000 subject classifications:
62H12

Keywords:
Adaptation
Continuous and categorical data
Curse of dimensionality
Dimension reduction
Minimax
Shrinkage

ABSTRACT

The problem of nonparametric estimation of the joint probability density of a vector of continuous and ordinal/nominal categorical random variables with bounded support is considered. There are numerous publications devoted to the cases of either continuous or categorical variables, and the curse of dimensionality and strong regularity assumptions are the two familiar issues in the literature. Mixed variables occur in practically all applications of the statistical science and, nonetheless, the literature devoted to the joint density estimation is practically next to none. This paper develops the theory of estimation of the density of mixed variables which is on par with results known for simpler settings. Specifically, a data-driven estimator is developed that adapts to unknown anisotropic smoothness of the joint density and, whenever the density depends on a smaller number of variables, performs a dimension reduction that implies the corresponding optimal rate of the mean integrated squared error (MISE) convergence. The results hold without traditional, in the density estimation literature, minimal regularity assumptions like differentiability or continuity of the density. The procedure of estimation is based on mimicking an oracle-estimator that knows the underlying density, and the main theoretical result is the oracle inequality which relates the MISEs of the estimator and the oracle-estimator. The proof is based on a new exponential inequality for Sobolev statistics which is of interest on its own merits.

1. Introduction

Nonparametric estimation of the density is a classical statistical problem. There is enormous literature devoted to the case of the joint density of continuous variables; see reviews and discussions in [2,4,6,16,23,28,31,33]. A smaller but still impressive number of publications is devoted to the case of discrete variables (e.g., [11–13,19,21,34]).

The case of mixed variables, including continuous and ordinal/nominal categorical variables, is absolutely classical in the statistical science and it occurs in numerous applications ranging from medicine and biology to economics and actuarial science, and there is a growing understanding of the importance of developing the theory and methods of density estimation for a vector of mixed variables [7,14,15,19,20,22,24,34]. At the same time, the literature devoted to the theory of estimation of the density of mixed variables is practically next to none, with the main publications being Hall [14], Li and Racine [22,21] and Racine [27].

In this paper the case of an $(m+q)$-dimensional vector of m continuous and q categorical random variables is considered. It is established in [22] that the mean integrated squared error (MISE) of a cross-validated kernel estimator, under the assumption of 4-fold differentiability of the multivariate density with respect to each continuous variable, attains the sub-optimal rate $n^{-4/(4+m)}$. Later, Hall et al., [15] published a paper about conditional density estimation of a univariate

E-mail address: efrom@utdallas.edu.

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.10.006

random variable given a vector of mixed predictors with a bounded support. In that paper the authors estimate the conditional density as the ratio of the joint density and the design density. Thanks to this approach, the analysis of their proof reveals that it is possible to deduce the following result about estimation of the joint density with a bounded support. Suppose that a multivariate density is twice differentiable with respect to each continuous variable and, additionally, it depends only on $m_1 \leq m$ continuous variables. Then the MISE of a cross-validated kernel estimator attains the optimal (minimax) rate $n^{-4/(4+m_1)}$. This result is significant because: (i) the estimator is optimal for the classical example of twice-differentiable densities; (ii) the estimator performs the dimension reduction which relaxes the notorious curse-of-dimensionality. According to the recent review in [27], and the author's literature search, no stronger theoretical result is known for the case of mixed variables. Among related recent results, let us mention Efromovich [7] where estimation of the conditional density with the mixed vector-predictor is considered. It is established that if minimal regularities of the isotropic conditional density and the design density increase with the dimensionality of the vector-predictor, then adaptation to unknown smoothness is possible.

Let us stress that the assumption of a bounded support is critical for exploring a possible dimension reduction when the density does not depend on a subgroup of variables. Indeed, the joint density always depends on a variable with unbounded support due to the fact that the density must be integrated/summable to 1 over the support. As a result, to perform a dimension reduction in the case of an unbounded support some additional assumptions about the estimated density should be made. For instance, in [31], where the case of the vector of continuous variables supported on a real line is considered, the proposed remedy is to assume that the joint density is proportional to a normal density. This is definitely an interesting topic to consider, and the case of densities with unbounded support will be presented in a separate publication.

A statistical model, used in this paper, is as follows. It is assumed that a sample $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ of i.i.d. random vectors, distributed as $\mathbf{Z} := (\mathbf{X}, \mathbf{V})$, is available. Here $\mathbf{V} := (V_1, \ldots, V_q)$ is the vector of categorical variables, each V_ν takes on a finite number M_ν of values which are coded as $\{0, 1, \ldots, M_\nu - 1\}$. The marginal probability mass function of $\mathbf{V}$ is denoted as $p(\mathbf{v})$ and it is supported on $\mathcal{D}_q := \prod_{\nu=1}^{q}\{0, 1, \ldots, M_\nu - 1\} := \{0, 1, \ldots, M_1 - 1\} \times \cdots \times \{0, 1, \ldots, M_q - 1\}$. In what follows it is assumed that $p(\mathbf{v})$ is positive on its support. $\mathbf{X}$ denotes the vector $(X_1, \ldots, X_m)$ of continuous variables. The multivariate function of interest is the joint probability density function $f(\mathbf{z}) := f(\mathbf{x}, \mathbf{v}) := f(\mathbf{x}|\mathbf{v})p(\mathbf{v})$, with respect to the product measure of Lebesgue and counting measures, supported on $[0, 1]^m \times \mathcal{D}_q$. Here $f(\mathbf{x}|\mathbf{v})$ is the conditional density of continuous variables given the vector of categorical variables.

The aim of the paper is to expand upon the known theory of estimation of the density of mixed variables. Specifically, the case of a multivariate anisotropic density, with possibly different smoothness in each continuous variable, will be considered. This setting is motivated by the seminal paper (with discussion) of Hoffmann and Lepski [17] where the model of filtering a signal from Gaussian noise was considered under the assumption that the signal belongs to an anisotropic Sobolev function class. To shed light on anisotropic functions and the minimax MISE convergence, suppose that the density has α_r derivatives with respect to the rth continuous variable. Note that α_r describes smoothness of the density in the rth continuous variable. Then the minimax rate of the MISE convergence is $n^{-2\alpha/(2\alpha+1)}$ where the parameter

$$\alpha := \frac{1}{\sum_{r=1}^{m} \alpha_r^{-1}} \tag{1}$$

is called the effective smoothness of the multivariate density [6,17]. Note that if $\alpha_1 = \cdots = \alpha_m = 2$, then the effective rate is $\alpha = 2/m$ and this yields the classical rate $n^{-4/(4+m)}$ of the MISE convergence. Note how the effective rate sheds light on the curse-of-dimensionality.

Further, in this paper the assumption, traditionally used even in the univariate density estimation literature, that the density is differentiable, or at least continuous, is relaxed. Specifically, anisotropic Sobolev classes of any order are considered and a corresponding minimax estimator is suggested. Further, the developed minimax estimator adapts to the unknown effective smoothness of the density. Further, if the density depends on a smaller number of variables, then the proposed estimator seizes the opportunity and performs the corresponding dimension reduction. This is the main scope of the paper, and let us also note that there are many interesting theoretical opportunities for a further research motivated by, e.g., [1,3,6,8,16,18,20,22,26–30,32,34–36].

The proposed estimator uses an orthogonal series approach and the adaptation is based on a blockwise-shrinkage procedure which mimics an oracle-estimator that knows the estimated density. The main theoretical result is an oracle inequality that bounds the estimator's MISE by the oracle's MISE plus a negligibly small remainder term. For now this is a standard approach used in the literature (e.g., [4,6,9,8,7,17,35,36]).

The structure of the paper is as follows. We finish this section by introducing notation and general assumptions used in the paper. Section 2 defines the oracle-estimator. The data-driven estimator, mimicking the oracle, is defined in Section 3. Theoretical results are presented in Section 4. Proofs are placed in the Appendix.

Notation, definitions and general assumptions are presented below.

Remember that $\mathcal{D}_q := \prod_{\nu=1}^{q}\{0, \ldots, M_\nu - 1\} := \{0, \ldots, M_1 - 1\} \times \cdots \times \{0, \ldots, M_q - 1\}$ and set $\mathcal{D}_{mq} := [0, 1]^m \times \mathcal{D}_q$, $\mathbf{z} := (\mathbf{x}, \mathbf{v}) \in \mathcal{D}_{mq}$, $\mathbf{x} := (x_1, \ldots, x_m)$, $\mathbf{v} := (v_1, \ldots, v_q)$, $\mathbf{r} := (r_1, \ldots, r_m)$, $\mathbf{w} := (w_1, \ldots, w_q)$, $\mathbf{i} := (\mathbf{r}, \mathbf{w}) \in \mathcal{N}_{mq} := \{0, 1, 2, \ldots\}^m \times \mathcal{D}_q$, and $M := \prod_{\nu=1}^{q} M_\nu$. $I(\cdot)$ denotes the indicator function, $\lfloor a \rfloor$ is the largest integer which is smaller than or equal to a. Cs denote generic positive constants.