High-Dimensional Kernel Density Estimation

Liu Huihang
SA18017026
QQ: 184050473
MAIL: huihang@mail.ustc.edu.cn

Abstract

Theory and Practice Casua

Keywords: Nonparametric density estimation, Sparsity, Adaptive bandwidth selection, High-dimensionality

1. Introduction

1.1 Motivation

Modern data acquisition routinely produces massive amounts of high dimensional and highly complex datasets, including interactive logs from search engines, traffic records from network routing, chip data from high throughput genomic experiments, and image data from functional Magnetic Resonance Imaging (fMRI). Driven by the complexity of these new types of data, highly adaptive and reliable data analysis procedures are crucially needed.

Older high dimensional theories and learning algorithms rely heavily on parametric models, which assume the data come from an underlying distribution that can be characterized by a finite number of parameters. If these assumptions are correct, orcal property – accurate estimates, precise preditcions and consistent variable selections (**?**) – can be expected. However, given the increasing complexity of modern data, conclusions inferred under these restrictive assumptions can be misleading. To handle this challenge, we focus on nonparametric methods, which directly conduct inference in infinite-dimensional spaces and thus are powerful enough to capture the subtleties in most modern applications.

2. Density-rodeo: High-dimensional Nonparametric Density Estimation

We consider the problem of estimating the joint density of a d-dimensional random vector $X = (X_1, X_2, \ldots, X_d)$ when d is large ($d \gg 3$). Current nonparametric density estimation methods, such as kernel density estimation or local likelihood methods, fail in this case due to the exponentially increasing amount of data required and intractable computational cost of bandwidth selection. In this paper, with a simple modiflcation of a previously developed nonparametric regression framework named rodeo (regularization of derivative expectation operator), we propose a computationally attractive alternative to perform high-dimensional density estimation. We empirically show that the density rodeo works well even for very high-dimensional problems in terms of both accuracy and efficiency. When the unknown density function satisfles some suitably defined sparsity condition, our approach avoids the curse of dimensionality and achieves an optimal converge rate of the risk. Theoretical guarantees are provided even when the dimension is allowed to increase with sample size.

2.1 Inro

Consider the problem of estimating the joint density of a continuous d-dimensional random vector

$$X = (X_1, \ldots, X_d) \sim \mathcal{F} \tag{1}$$

where F is the unknown distribution with the density function $f(x)$. The objective is to estimate a function $\hat{f}(x)$ that best approximates $f(x)$ according to some criterion. If the parametric form of the distribution is known, parametric density estimation methods can be applied. However, in most real applications, it's unlikely that the underlying distribution can be characterized by just a few parameters. In these cases, nonparametric density estimation is preferred, as it makes fewer assumptions about the true density.

The nonparametric density estimation problem has been the focus of a large body of research. From a frequentist perspective, the most popular technique is Parzen and Rosenblatt's kernel based method [1, 2], which uses flxed bandwidth local functions (e.g. Gaussians) to interpolate the multivariate density. Hjort et al. and Loader [3, 4, 5] independently developed the local likelihood method, which corrects the boundary bias for standard kernel density estimators. Difierent adaptive bandwidth kernel density estimators were introduced by Scott et al. and Staniswalis [6, 7, 8]. Bandwidth selection approaches for these density estimators include crossvalidation or some heuristic techniques. These methods work very well for low-dimensional problems ($d \leq 3$) but are not efiective for high-dimensional problems. The major difficulty is due to the intractable computational cost of cross validation, when bandwidths need to be selected for each dimension, and the lack of theoretical guarantees for the other heuristics. What's more, in a d-dimensional space, minimax theory shows that the best convergence rate for the mean squared error is $\mathcal{R}_{opt} = O(n^{-2k/(2k+d)})$. For example, in a Sobolev space of order k, which represents the "curse of dimensionality" when d is large. Instead of using local kernels, Friedman et al. [9] developed an exploratory projection pursuit approach which looks for "interesting" (e.g. non-Gaussian) low dimensional data projections to reveal the distribution pattern of the original data. Stone et al. [10] proposed the Log-spline model to estimate $\log f(x)$, while Silverman et al [11] developed penalized likelihood method for density estimation. From a Bayesian perspective, Escobar et al. [12] proposed a Bayesian density estimation method for normal mixtures in the framework of mixtures of Dirichlet processes. When the base distribution of the Dirichlet process is conjugate to the data likelihood, MCMC sampling can be derived for posterior inference. However, for very high-dimensional problems, the computation is slow, representing another form of the curse of dimensionality in the Bayesian setting. Therefore, for a very high-dimensional density estimation problem, it is desirable to somehow exploit low dimensional structure or sparsity in order to combat the curse of dimensionality and develop methods that are both computationally tractable and amenable to theoretical analysis.

Lafierty and Wasserman developed a nonparametric regression framework called rodeo [13]. For the regression problem, $Y_i = m(X_i) + \epsilon_i$, $i = 1, ;n$, where $X_i = (X_{i1}, \ldots, X_{id}) \in R^d$ is a d-dimensional covariate. Assuming that the true function only depends on r covariates $r \ll d$, the rodeo can simultaneously perform bandwidth selection and (implicitly) variable selection to achieve a better minimax convergence rate of $O(n^{-4/(4+r)})$, as if the r relevant variables were explicitly isolated in advance. The purpose of this paper is to extend the idea of rodeo to the nonparametric density estimation setting. Toward this goal, we need to flrst

deflne an appropriate "sparsity" condition in the density estimation setting. Assume that the variables are numbered such that the "relevant" dimensions correspond to $1 \leq j \leq r$ and the "irrelevant" dimensions correspond to $r+1 \leq j \leq d$, one may flrst want to write $f(x) \propto f(x_1, \ldots, x_r)$. However, this is not a proper distribution on the irrelevant dimensions. To make it well-deflned, we can assume, without loss of generality, that all dimensions have compact support [0; 1]d, that is, $f(x) \propto f(x1, \ldots, x_r) \prod_{j=1}^{d} I\{0 \leq xj \leq 1\}$. In this case, we deem the uniform distribution as irrelevant (or, "uninteresting") dimensions, while the non-uniform distributions are relevant. In fact, we can further generalize this deflnition. Our sparsity speciflcation is characterized by

$$f(x) \propto g\,(x_1, \ldots, x_r)\, h(x) \quad \text{where } h_{jj}(x) = o(1) \text{ for } j = 1, \ldots, d. \tag{2}$$

Thus, we assume that the density function $f(\cdot)$ can be factored into two parts: the relevant components $g(\cdot)$ and the irrelevant components $h(\cdot)$; $h_{jj}(x)$ is the second partial derivative of h on the j-th dimension. The constraint in expression **??** may look unnatural at the flrst sight, but it simply imposes a condition that $h(\cdot)$ belongs to a family of very smooth functions (e.g. the uniform distribution). We adopt a standard setup where the function f and dimension d are allowed to vary with sample size n. The motivation for such a deflnition comes from both realworld scenarios and the need for theoretical analysis. Empirically, many problems have such a sparsity property; later in this paper, we will show an image processing example. Theoretically, this definition is strong enough for us to prove our main theoretical results, showing that minimax rates in the efiective dimension r are achievable. In fact, we can even generalize $h(\cdot)$ to other other distributions (e.g. Gaussian) to build a more general framework. Later in this paper, we use Gaussian as a special case to illustrate this possibility.

In this paper, based on the above deflned sparsity assumptions, we adapt the regression *rodeo* framework to density estimation problems, referring to the resulting adaptive bandwidth density estimation method as the density rodeo. Similar to rodeo for regression, the *density rodeo* is built upon relatively simple and theoretically well understood nonparametric density estimation techniques, such as the kernel density estimator or local likelihood estimator, leading to methods that are simple to implement and can be used to solve very high dimensional problems. As we present below, for any $\epsilon > 0$, the density rodeo achieves the near optimal convergence rate in the risk

$$\mathcal{R}_{h^*} = O\left(n^{-4/(4+r)+\epsilon}\right) \tag{3}$$

Thus, it avoids the curse of apparent dimensionality d by zeroing in on the efiective dimension $r \ll d$ by bandwidth selection. Theoretical guarantees are provided even when d is allowed to increase with n. This work also illustrate the generality of the rodeo framework, showed that it is adaptable to a wide range of nonparametric inference problems. Other recent work that achieves risk bounds for density estimation with a smaller efiective dimension is the minimum volume set learning method of Scott and Nowak [14]. Based on a similar sparsity assumption, they provide theoretical guarantees for a technique that uses dyadic trees. Another related work is done by Gray and Moore [15], from a computational perspective, their dual-tree algorithm could also deal with large sample size and high dimensional problems ($d = 20 \sim 30$). It would be a very interesting future work to compare the performance of density rodeo with their algorithm.