High-Dimensional Kernel Density Estimation

Liu Huihang
SA18017026

Abstract

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 ($d \geq 3$). Classical 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. We reviewed some literature on nonparameter kernel density estimation method with large and high-dim data. In this essay, with a simple modifIcation of a previously developed nonparametric regression framework named rodeo (regularization of derivative expectation operator), we reviewed 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 deflned 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.

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

1. Introduction

1.1 Motivation and Model

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 (Fan and Lv, 2009) – 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.

We 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.

1.2 Literature Review

1.2.1 Classical KDE

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 (Parzen, 1962; Rosenblatt, 1956), which uses fixed bandwidth local functions (e.g. Gaussians) to interpolate the multivariate density. Hjort and Jones (1996); Hjort and Glad (1995); Loader and Others (1996) 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 Terrell et al. (1992); Sain and Scott (1996); Staniswalis (1989). 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.

Methods of data-based optimal bandwidth calculation have also high computational demands. Second order plug-in method (Sheather and Jones, 1991), which involves estimating second derivative of density function from given sample, is of $O(dn^2)$ complexity. Also least squares cross-validation method (LSCV) (Rudemo, 1982; Bowman, 1984), where selecting optimal bandwidth is based on minimizing objective function g(h), has the same polynomial time complexity. Few approximation techniques have been proposed to deal with the problem of time-consuming calculations of kernel density estimates. The first of them, proposed by Silverman (1982b), uses fast Fourier transform (FFT). The other one applies Fast Gauss Transform (FGT) as suggested by Elgammal et al. (2003). An alternative to those methods could use parallel processing for obtaining density function estimates. The pioneer paper in this subject is due to Racine (2002). It proves the usability of parallel computations for kernel density estimation but covers in detail only the estimation itself. Parallelization is done by dividing set of points where estimator is to be evaluated. Each processor obtains the density function estimation for approximately pc points (where c denotes number of CPUs involved). Zheng et al. (2013) propose randomized and deterministic algorithms with quality guarantees which are orders of magnitude more efficient than previous algorithms, due to paralleled and distributed implementation. This algorithm can run on moderate dimensional data set. Lukasik (2007) verify how parallel processing can be applied for kernel estimation, bandwidth selection and adaptation in a multicomputer environment using MPI (Message Passing Interface)(Snir et al., 1998).

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. (1984) 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 (1990) proposed the Log-spline model to estimate log $f(x)$, while Silverman (1982a) developed penalized likelihood method for density estimation. From a Bayesian perspective, Escobar and West (1995) 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.

1.2.2 Rodeo

Lafierty and Wasserman developed a nonparametric regression framework called rodeo (Wasserman and Lafferty, 2006). For the regression problem, $Y_i = m(X_i) + \epsilon_i$, $i = 1, \dot{,} n$, 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 first 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 first 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, $f(x) \propto f(x_1, \ldots, x_r) \prod_{j=1}^{d} I\{0 \leq x_j \leq 1\}$. 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

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

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 2 may look unnatural at the first 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 (2006). 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 (2003), 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.

This paper is organized as follows, Section 2 gives out the basic idea of density estimation rodeo. In Section 3, we derived the density estimation rodeo for both kernel density estimator and showed the drodeo algorithms. Section 4 specifles our main theoretical results about the asymptotic running time, selected bandwidths, and convergence rate of the risk. Section 5 uses both synthetic and real-world dataset to test our method. Finally, Section 6 summarizes the results.

2. Methodology

The key idea in our approach is as follows. Fix a point x and let $\hat{m}_h(x)$ denote an estimator of $m(x)$ based on a vector of smoothing parameters $h = (h_1, \ldots, h_d)$. If c is a scalar, then we write $h = c$ to mean $h = (c, \ldots, c)$.

Let $M(h) = \mathbb{E}(\hat{m}_h(x))$ denote the mean of $\hat{m}_h(x)$. For now, assume that $x = x_i$ is one of the observed data points and that $\hat{m}_0(x) = Y_i$. In that case, $m(x) = M(0) = \mathbb{E}(Y_i)$. If $P = (h(t) : \ 0 \leq t \leq 1)$ is a smooth path through the set of smoothing parameters with

$h(0) = 0$ and $h(1) = 1$ (or any other fixed, large bandwidth) then

$$\begin{aligned} m(x) &= M(0) = M(1) + M(0) - M(1) && \text{(4a)} \\ &= M(1) - \int_0^1 \frac{dM(h(s))}{ds}\, ds && \text{(4b)} \\ &= M(1) - \int_0^1 \big\langle D(s), \hbar(s) \big\rangle ds && \text{(4c)} \end{aligned}$$

where

$$D(h) = \nabla M(h) = \left(\frac{\partial M}{\partial h_1}, \ldots, \frac{\partial M}{\partial h_d} \right)^T \tag{5}$$

is the gradient of $M(h)$ and $\hbar(s) = \frac{dh(s)}{ds}$ is the derivative of $h(s)$ along the path. A biased, low variance estimator of $m(x)$ is $\hat{m}_1(x)$. An unbiased estimator of $D(h)$ is

$$Z(h) = \left(\frac{\partial \hat{m}_h(x)}{\partial h_1}, \ldots, \frac{\partial \hat{m}_h(x)}{\partial h_d} \right)^T . \tag{6}$$

The naive estimator

$$\hat{m}(x) = \hat{m}_1(x) - \int_0^1 \big\langle Z(s), \hbar(s) \big\rangle ds \tag{7}$$

is identically equal to $\hat{m}_0(x) = Y_i$, which has poor risk since the variance of $Z(h)$ is large for small h. However, our sparsity assumption on m suggests that there should be paths for which $D(h)$ is also sparse. Along such a path, we replace $Z(h)$ with an estimator $\hat{D}(h)$ that makes use of the sparsity assumption. Our estimate of $m(x)$ is then

$$\tilde{m}(x) = \hat{m}_1(x) - \int_0^1 \big\langle \hat{D}(s), \hbar(s) \big\rangle ds\,. \tag{8}$$

To implement this idea we need to do two things: (i) we need to find a path for which the derivative is sparse and (ii) we need to take advantage of this sparseness when estimating D along that path.

3. Algorithm

In this section, we show the detailed density rodeo algorithm for kernel density estimator which are known to be simple and have good properties. Assuming x is a d-dimensional target point at which we want to evaluate the density values, let $\hat{f}_H(x)$ represents the kernel density estimator of $f(x)$ with a bandwidth matrix H. For a specifIc point, kernel density estimators smooth out the contribution of each observed data point over a local neighborhood of that data point. Assuming that $\mathcal{K}$ is a standard symmetric kernel, s.t. $\int \mathcal{K}(u)du = 1, \int u\mathcal{K}(u)du = 0_d$ while $\mathcal{K}_H(\cdot) = \frac{1}{det(H)}\mathcal{K}(H^{-1}\cdot)$ represents the kernel with bandwidth matrix $H = diag(h_1, \ldots, h_d)$.

$$\hat{f}_H(x) = \frac{1}{n \det(H)} \sum_{i=1}^{n} \mathcal{K}\left(H^{-1} (x - X_i) \right) \tag{9}$$

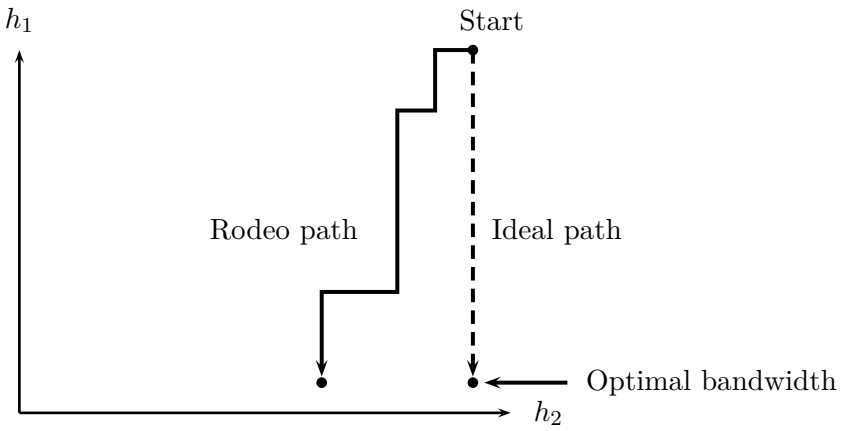

Figure 1: Conceptual illustration. The bandwidths for the relevant variables (h_1) are shrunk, while the bandwidths for the irrelevant variables (h_2) are kept relatively large. The simplest rodeo algorithm shrinks the bandwidths in discrete steps $1, \beta, \beta^2, \ldots$ for some $0 < \beta < 1$.

For the following density Rodeo algorithm, we assume that $\mathcal{K}$ is a product kernel and H to be diagonal with elements $h = (h_1, \ldots, h_d)$, therefore

$$\begin{aligned} Z_j &= \frac{\partial \hat{f}_H(x)}{\partial h_j} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{\dot{K}\left(\frac{x_j - X_{ij}}{h_j}\right)}{K\left(\frac{x_j - X_{ij}}{h_j}\right)} \prod_{k=1}^{d} K\left(\frac{x_k - X_{ik}}{h_k}\right) \\ &\equiv \frac{1}{n} \sum_{i=1}^{n} Z_{ji} \end{aligned} \tag{10}$$

For the variance term, since

$$s_j^2 = \mathrm{Var}\left(Z_j\right) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} Z_{ji}\right) = \frac{1}{n} \mathrm{Var}\left(Z_{j1}\right) \tag{11}$$

Here, we used the sample variance of the Z_{ji} to estimate $Var(Z_{j1})$. Therefore, for a Gaussian kernel, we have that

$$\begin{aligned} Z_j &= \frac{\partial \hat{f}_H(x)}{\partial h_j} \\ &= \frac{1}{nh_j^3} \prod_{k=1}^{d} \frac{1}{h_k} \sum_{i=1}^{n} \left((x_j - X_{ij})^2 - h_j^2 \right) \prod_{k=1}^{d} K\left(\frac{x_k - X_{ik}}{h_k} \right) \\ &\propto \frac{1}{n} \sum_{i=1}^{n} \left((x_j - X_{ij})^2 - h_j^2 \right) \prod_{k=1}^{d} K\left(\frac{x_k - X_{ik}}{h_k} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left((x_j - X_{ij})^2 - h_j^2 \right) \exp\left(- \sum_{k=1}^{d} \frac{(x_k - X_{ik})^2}{2h_k^2} \right) \end{aligned} \tag{12}$$

Further, for a general kernel, we have

$$Z_j = \frac{\partial \hat{f}(x)}{\partial h_j} = -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{h_j} + \frac{x_j - X_{ij}}{h_j^2} \frac{d \log K}{dx} \left(\frac{x_j - X_{ij}}{h_j} \right) \right) \prod_{k=1}^{d} \frac{1}{h_k} K\left(\frac{x_k - X_{ik}}{h_k} \right) \tag{13}$$

4. Theoretical Results

Theorem 1 *Under some moderate conditions, for every $\epsilon > 0$ the number of iterations T_n until the Rodeo stops satisfies*

$$\mathbf{P}\left(\frac{1}{4+r} \log_{1/\beta}\left(n^{1-\epsilon} a_n \right) \leq T_n \leq \frac{1}{4+r} \log_{1/\beta}\left(n^{1+\epsilon} b_n \right) \right) \longrightarrow 1 \tag{15}$$

More over, the algorithm outputs bandwidths h^* that satisfy

$$\mathbf{P}\left(h_j^* = h_j^{(0)} \text{ for all } j > r \right) \longrightarrow 1. \tag{16}$$

Also, we have

$$\mathbf{P}\left(h_j^{(0)} \left(nb_n \right)^{-1/(4+r)-\epsilon} \leq h_j^* \leq h_j^{(0)} \left(na_n \right)^{-1/(4+r)+\epsilon} \text{ for all } j \leq r \right) \longrightarrow 1 \tag{17}$$

Theorem 2 *Under the same condition of theorem 1, the risk Rh⁄ of the density Rodeo estimator satisfies*

$$\mathcal{R}_{h^*} = \tilde{O}_P\left(n^{-4/(4+r)+\epsilon} \right) \tag{18}$$

for every $\epsilon > 0$.

These theoretical properties show that rodeo algorithm guarantees convergence within a finite step in probability, outputs h^* will converge to its true value in probability and the risk of density Rodeo estimator is of order $n^{-4/(4+r)+\epsilon}$ which is the optimal order we can expect.

Rodeo: Hard thresholding version

1. *Select* $\beta_n = n^{-\alpha/\log^3 n}$ for some $0 < \alpha < 1$ and initial bandwidth

$$h_0 = \frac{c_0}{\log\log n} \tag{14}$$

   for some constant c_0. Let c_n be a sequence satisfying $c_n = O(1)$.

2. *Initialize* the bandwidths, and activate all covariates:
   (a) $h_j = h_0$, $j = 1, 2, \ldots, d$.
   (b) $\mathcal{A} = \{1, 2, \ldots, d\}$

3. *While* $\mathcal{A}$ *is nonempty*, do for each $j \in \mathcal{A}$:
   (a) Compute the estimated derivative expectation: Z_j (Eq. 12) and s_j (Eq. 11).
   (b) Compute the threshold $\lambda_j = s_j\sqrt{2\log(nc_n)}$.
   (c) If $|Z_j| > \lambda_j$, then set $h_j \leftarrow \beta h_j$; otherwise remove j from $\mathcal{A}$.

4. *Output* bandwidths $h^\star = (h_1, \ldots, h_d)$ and estimator $\tilde{m}(x) = \hat{m}_{h^\star}(x)$.

Table 1: The hard thresholding version of the rodeo, which can be applied using the derivatives Z_j of any nonparametric smoother.

5. Simulation

In this section, we applied the density rodeo on both synthetic and real data, including onedimensional, two-dimensionalexamples to investigate how the density estimation rodeo performs in various conditions. For the purpose of evaluating the algorithm performance quantitatively, we need some criterion to measure the distance between the estimated density function with the true density function.

In the following, we first use the simulated data, about which we have known the true distribution function, to investigate the algorithm performance.

5.1 One-dimensional examples

First, we illustrate the performance of the density Rodeo algorithm in one dimensional examples. We have conducted a series of comparative study on a list of 15 "test densities" proposed by Marron and Wand (1992), which are all normal mixtures representing many difierent types of challenges to density estimation. Our method achieves a comparable performance as the kernel density estimation with bandwidth selected by unbiased cross-validation (from the base library of R). Due to the space consideration, only the strongly

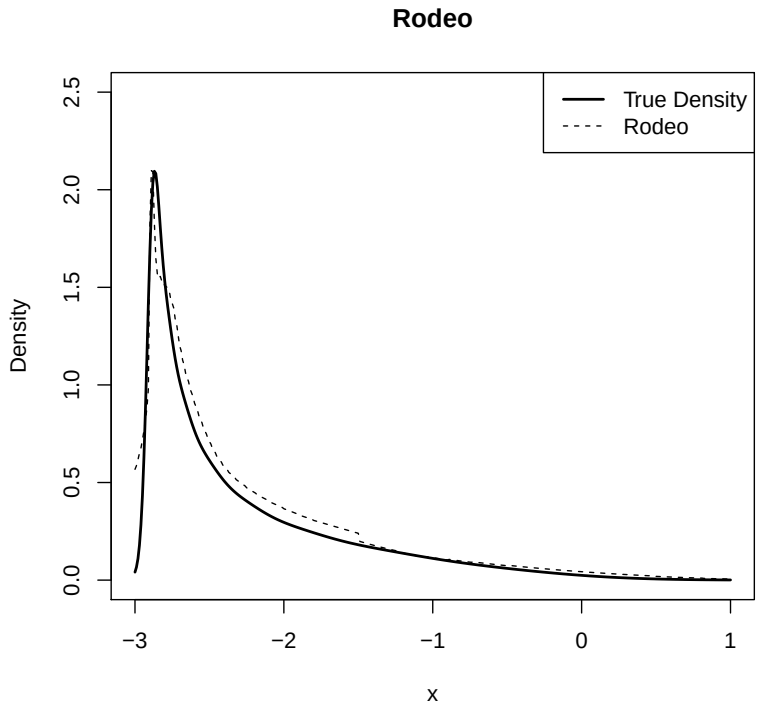

Figure 2: Local kernel density Rodeo.

skewed example is reported here, since it demonstrates the advantage of adaptive bandwidth selection for the density rodeo algorithms.

Example 1 (Strongly skewed density) *This density is chosen to resemble to lognormal distribution, it distributes as*

$$X \sim \sum_{i=0}^{7} \frac{1}{8} \mathcal{N}\left(3\left(\left(\frac{2}{3}\right)^{i} - 1\right), \left(\frac{2}{3}\right)^{2i}\right) \tag{19}$$

200 samples were generated from this distribution, The estimated density functions by the density rodeo algorithm, the density rodeo algorithm, and kernel density estimator with bandwidth chosen by unbiased cross validation are shown in Figure 3. In which, the solid line is the true density function, the dashed line illustrates the estimated densities by difierent methods. The density rodeo works the best, this is because because the true density function is strongly skewed, the flxed bandwidth density estimator can not flt the smooth tail very well. The last subplot from Figure 3 illustrates the selected bandwidth for the density rodeo method, it shows how smaller bandwidths are selected where the function is more rapidly varying. The plots show that the density rodeo works better than the remaining two methods, while the rodeo and the unbiased cross-validation methods are comparable in this one dimensional example.

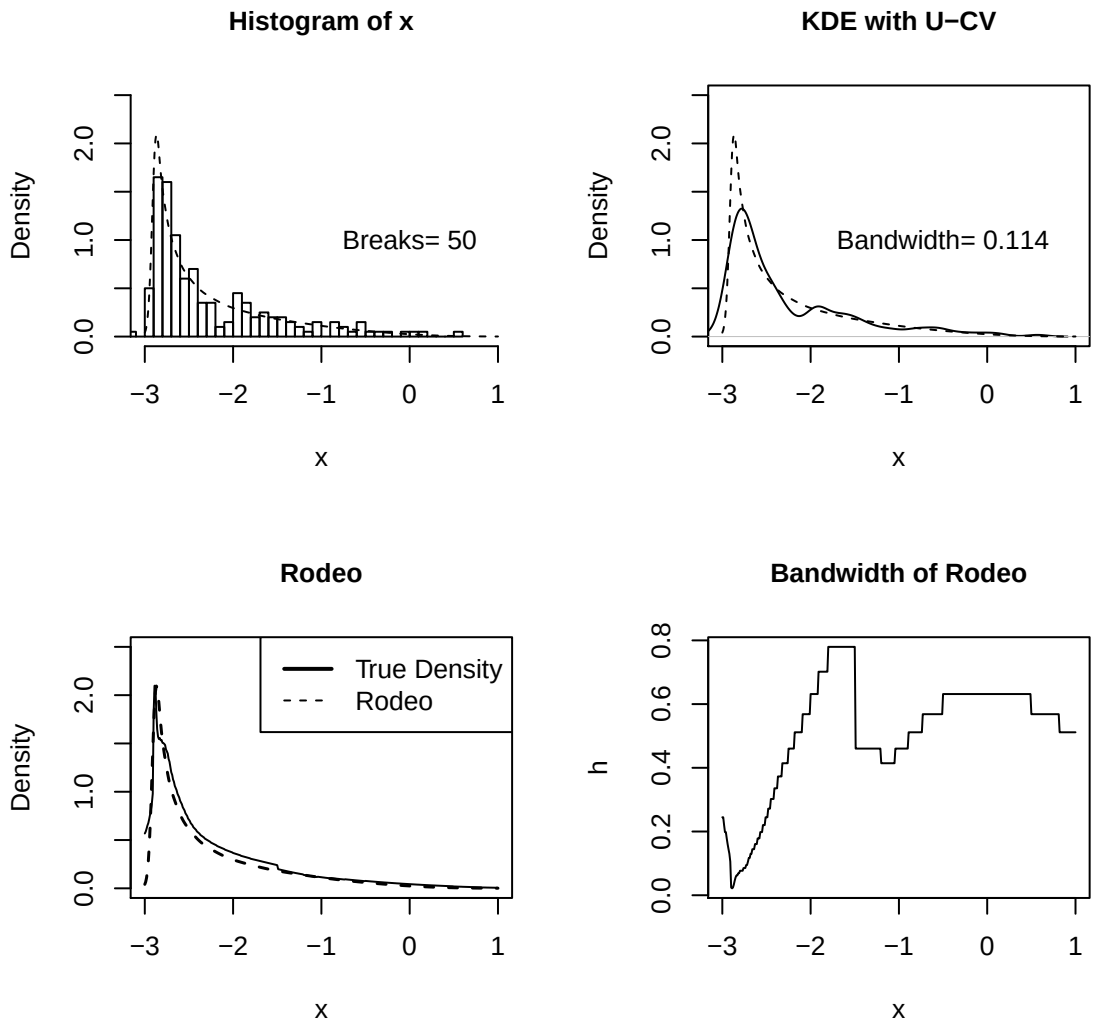

Figure 3: The difierent versions of the algorithms run on the highly skewed unimodal example. The first three plots are results for the difierent estimators, the last one is the fitted bandwidth for the density rodeo algorithm.

In stead of showing the positive cases where the density rodeo works well, we also give out a negative example. Which is a combined Beta distribution from Loader and Others (1996).

Example 2 (Combined Beta density) *The density function is a Beta mixtures on the support of [0, 1], it has a strong boundary efiect on the left side, the density function is*

$$\begin{aligned} f(x) &= \frac{2}{3} Beta(1,2) + \frac{1}{3} Beta(10,10) \\ &= \frac{2}{3} 2(1-x) + \frac{1}{3} \frac{19!}{9!^2} x^9 (1-x)^9 \end{aligned} \tag{20}$$

altogether 500 samples were generated from this distribution, the estimated density functions by difierent methods are shown in Figure 4.

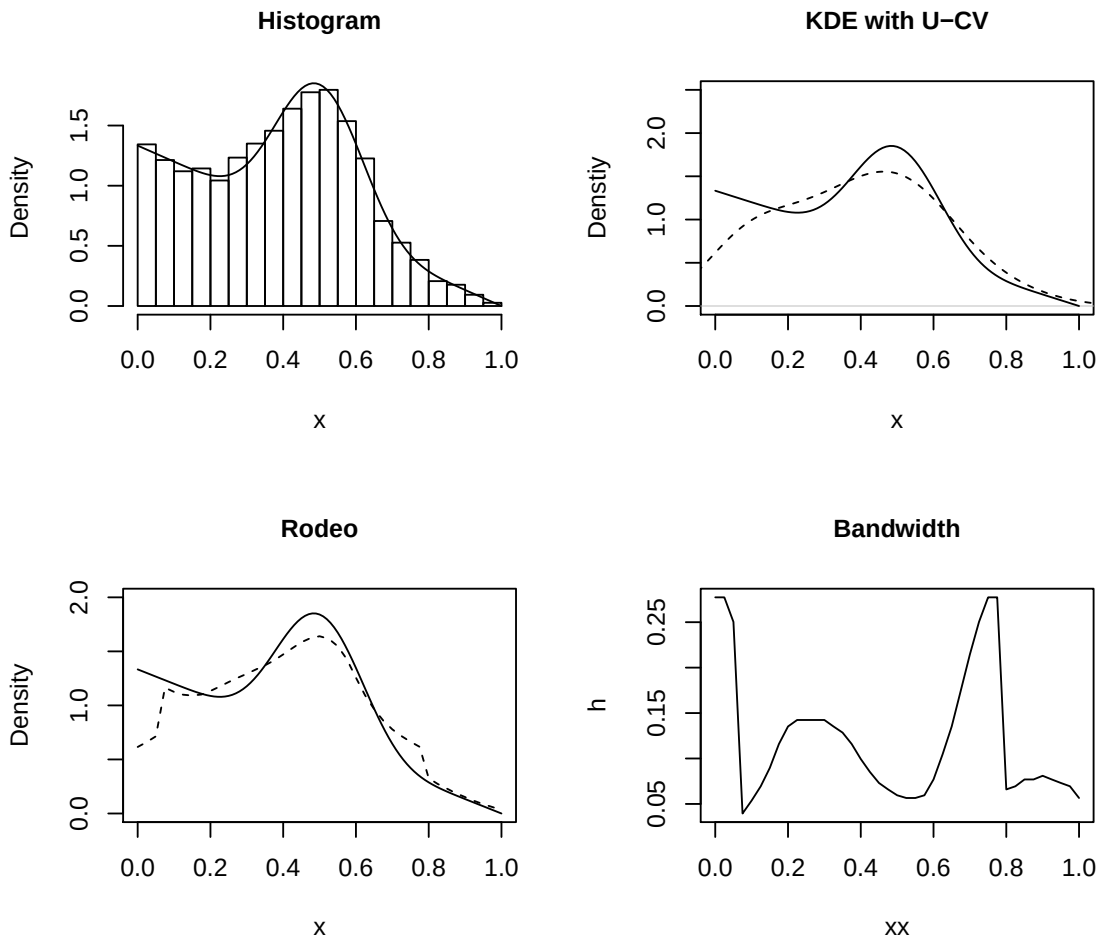

Figure 4: Results for the combined beta distribution. The first three plots are results for difierent estimators, the last one is the fitted bandwidth for the local density rodeo algorithm. In the first three plots, the solid lines represent true density functions and the dash lines represent different estimators respectively.

Similar as before, the solid line is the true density function, while the dashed line illustrates the estimated functions. In this case, the rodeo and the unbiased cross-validation estimate the density function well. However, the rodeo fails to flt the right tail. From the bandwidth plot (as shown in the last subplot in Figure 4), we see that the rodeo tends to select larger bandwidth near the right boundary, which cause the problem.

From the estimated bandwidth plot, we see that the boundary efiect problem is alleviated.

5.2 Two dimensional example

Here, two 2-dimensional examples are showed due to its ease of illustration. One example used a synthetic dataset, the other one uses some real data analyzed by the other authors. The density rodeo's performance is compared with a built-in method named KDE2d (from MASS package in R). The empirical results show that the density rodeo algorithm works better than the built-in method on the synthetic data, where we know the ground truth. For the real-world dataset, where we do not know the underling distribution, our method achieves a very similar result as those of the original authors who analyzed this data before.

Example 3 (Beta distribution with irrelevant unifrom distribution)

$$\begin{aligned} X_1 &\sim \frac{2}{3}\,\mathrm{Beta}(1,2) + \frac{1}{3}\,\mathrm{Beta}(10,10) \\ X_2 &\sim \mathit{Uniform}(0,1) \end{aligned} \tag{21}$$

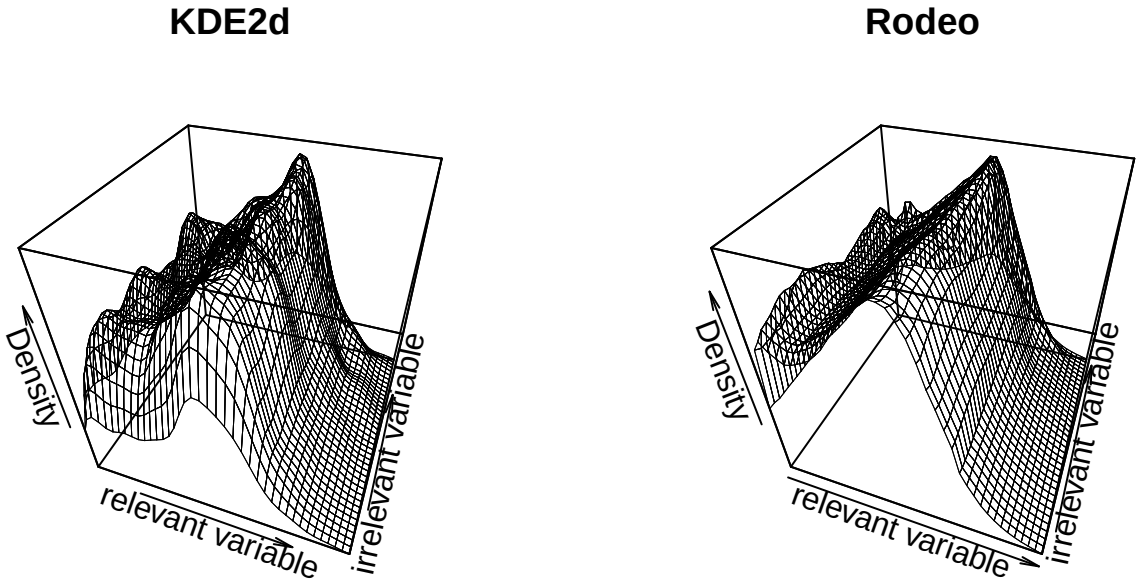

Figure 5: Results for the combined beta distribution. The first three plots are results for difierent estimators, the last one is the fitted bandwidth for the local density rodeo algorithm. In the first three plots, the solid lines represent true density functions and the dash lines represent different estimators respectively.

Figure 5 illustrates the estimated density functions by the density rodeo and the built-in method KDE2d. From which, we see that the rodeo algorithm flts the irrelevant uniform

dimension perfectly, while KDE2d fails. For a quantitative comparison, we evaluated the empirical Hellinger distance between the estimated density and the true density, the rodeo algorithm outperforms KDE2d uniformly on this example. For a qualitative study, Figure 5 illustrates the numerically integrated marginal distributions of the two estimators (not normalized), it's consistent with the previous observations.

Further, we can apply this algorithm on high dimensional data ($d \geq 3$) with spare true variables, but we lack of time and computation resource to adjust the parameters and get proper results.

6. Conclusion

This work is mainly purposed to illustrate the generality of the rodeo framework (Lafferty and Wasserman, 2005; Wasserman and Lafferty, 2006; Lafferty and Wasserman, 2008). Under some suitably-deflned sparsity conditions, the previously developed nonparametric regression framework is easily adapted to perform high-dimensional density estimation. The resulting method is both computationally e–cient and theoretically soundable. Empirical results show that our method is better than the built-in methods in some cases.

In the future, we can try to develop rodeo using paralleled or distributed computation to speed it up further. Also we can expend its variation for example boostrap sampling, using soft threshold to select h, backward method or any other statistical method.

References

Adrian W Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

Ahmed Elgammal, Ramani Duraiswami, and Larry S Davis. Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1499–1504, 2003.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Jianqing Fan and Jinchi Lv. A Selective Overview of Variable Selection in High Dimensional Feature Space (Invited Review Article). *Statistica Sinica*, 20(1):101–148, oct 2009. ISSN 10170405. URL `http://arxiv.org/abs/0910.1122`.

Jerome H Friedman, Werner Stuetzle, and Anne Schroeder. Projection pursuit density estimation. *Journal of the American Statistical Association*, 79(387):599–608, 1984.

Alexander G Gray and Andrew W Moore. Very fast multivariate kernel density estimation via computational geometry. In *Joint Stat. Meeting*, 2003.

Nils Lid Hjort and Ingrid K Glad. Nonparametric density estimation with a parametric start. *The Annals of Statistics*, pages 882–904, 1995.

Nils Lid Hjort and M Chris Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

John Lafferty and Larry Wasserman. Rodeo: Sparse nonparametric regression in high dimensions, 2005. ISSN 10495258.

John Lafferty and Larry Wasserman. Rodeo: Sparse, greedy nonparametric regression, 2008. ISSN 00905364.

Clive R Loader and Others. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

Szymon Lukasik. Parallel computing of kernel density estimates with MPI, 2007. ISSN 03029743.

J Steve Marron and Matt P Wand. Exact mean integrated squared error. *The Annals of Statistics*, pages 712–736, 1992.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Jeff Racine. Parallel distributed kernel estimation. *Computational Statistics & Data Analysis*, 40(2):293–302, 2002.

Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

Mats Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, pages 65–78, 1982.

Stephan R Sain and David W Scott. On locally adaptive density estimation. *Journal of the American Statistical Association*, 91(436):1525–1534, 1996.

Clayton D Scott and Robert D Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

Simon J Sheather and Michael C Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):683–690, 1991.

Bernard W Silverman. On the estimation of a probability density function by the maximum penalized likelihood method. *The Annals of Statistics*, pages 795–810, 1982a.

Bernhard W Silverman. Algorithm AS 176: Kernel density estimation using the fast Fourier transform. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 31(1): 93–99, 1982b.

Marc Snir, William Gropp, Steve Otto, Steven Huss-Lederman, Jack Dongarra, and David Walker. *MPI–the Complete Reference: The MPI core*, volume 1. MIT press, 1998.

Joan G Staniswalis. Local bandwidth selection for kernel estimates. *Journal of the American Statistical Association*, 84(405):284–288, 1989.

Charles J Stone. Large-sample inference for log-spline models. *The Annals of Statistics*, pages 717–741, 1990.

George R Terrell, David W Scott, and Others. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992.

Larry Wasserman and John D Lafferty. Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems*, pages 707–714, 2006.

Yan Zheng, Jeffrey Jestes, Jeff M Phillips, and Feifei Li. Quality and efficiency for kernel density estimates in large data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 433–444. ACM, 2013.