Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation

Anonymous Authors
Anonymous Organization
anon@email.com

Abstract

The construction and theoretical analysis of the most popular universally consistent nonparametric density estimators hinge on one functional property: smoothness. In this paper we investigate the theoretical implications of incorporating a multi-view latent variable, a type of low-rank model, into nonparametreic density estimation. To do this we perform extensive analysis on histogram-style estimators that integrate a multi-view model. Our analysis culminates in showing that there exists a universally consistent histogram-style estimator that converges to any multi-view model with a finite number of Lipschitz continuous components at a rate of $\widetilde{O}(1/\sqrt[3]{n})$ in L^1 error, in contrast to the standard histogram estimator which can converge at a rate slower than $1/\sqrt[d]{n}$ on the same class of densities. Beyond this we also introduce a new type of nonparametric latent variable model based on the Tucker decomposition. A very rudimentary experimental implementation of the ideas in our paper already demonstrates considerable practical improvements over the standard histogram estimator. We also provide a thorough analysis of the sample complexity of our Tucker model. Thus, our paper provides solid first theoretical foundations for extending low-rank techniques to the nonparametric setting.

1 Introduction

Nonparametric density estimators are density estimators capable of estimating a density f while making few to no assumptions on f. Two of the most widely used nonparametric density estimators include the histogram estimator and the kernel density estimator (KDE). A common characteristic of these two estimators is that they make estimation tractable via a parameter that relates to smoothness: bin width and bandwidth. Large bin width or bandwidth allow for good control of estimator variance at the cost of increased bias via smoothing for KDEs and having large constant regions for the histogram. Selection of this parameter is crucial to estimator performance. In fact a recent survey on bandwidth selection for KDEs found at least 30 methods for setting this value along with a few other surveys dedicated to selection of this parameter [12].

While nonparametric density estimation has been shown to be effective for many tasks, it has been observed empirically that estimator performance typically declines as data dimensionality increases, a manifestation of the curse of dimensionality. For the histogram and KDE this phenomenon also has concrete mathematical analogs. An example of this is that these estimators are only universally consistent* if $n \to \infty$ and $h \to 0$, with $nh^d \to \infty$, where n is number of samples and h is the bin width for the histogram and bandwidth parameter for the KDE [11]. This begs the question as to whether there exists some other way to constrain model capacity so as to alleviate this exponential dependence on dimension.

*A density estimator is *universally consistent.* if it asymptotically recovers *any* density

Preprint. Under review.

In this paper we theoretically analyze the advantages of including an additional constraint that is similar to matrix/tensor rank to such models. We do so by analyzing histogram style estimators which are enforced to have a low-rank PARAFAC or Tucker decomposition. We present extensive theoretical analysis demonstrating the advantages of enforcing this constraint. It is difficult to summarize our results succinctly but we essentially find that enforcing a low rank constraint to the histogram estimator improves the estimator whenever the target density (the unknown density we are trying to estimate) is even vaguely close to low rank, and that this improvement is exponential in data dimension. We include some discussion of densities which have no low rank (we mean this in a multi-view sense) in the Appendix and find that such densities are quite weird and pathological.

The bulk of this work focuses on analysis of hypothetical estimators that, while offering large statistical advantages, are not computationally tractable, a fact that is worthwhile to keep in mind. At the end of our paper we include experiments that find low rank histograms, computed using a standard non-negative tensor factorization technique, still offer some improvement over the standard histogram estimator.

We remark that our work is distinct from a previous work that decomposes nonparametric density estimators to find a low rank density estimator [29]. The approach in [29] splits the task of finding a low rank density estimate into two components, estimation then decomposition, offering no better rate of convergence since its performance is hampered by initial estimation task.

1.1 Density Models

Here we introduce the low-rank models that will be used to constrain our estimators. The first is a *multi-view model*, which is equivalent to enforcing the marginals of a density ($\hat{f}_{i,j}$) to be independent conditioned on an unseen discrete random variable (with parameters $w_1, \ldots, w_k$). More precisely a multi-view model has the following form

$$\hat{f}\left(x_1, \ldots, x_d\right) = \sum_{i=1}^{k} w_i \hat{f}_{i,1}\left(x_1\right) \hat{f}_{i,2}\left(x_2\right) \cdots \hat{f}_{i,d}\left(x_d\right). \tag{1}$$

For the estimators we propose, we let the marginals $\hat{f}_{i,j}$ have the form of one-dimensional histograms and use number of components k to restrict estimator capacity. The histograms that we consider will have a finite number setting k equal to the number of bins our constrained space of histograms now contains all possible histograms (including the standard histogram). When $k = 1$ the model consists of Bayes models: with histogram marginals $\hat{f}\left(x_1, \ldots, x_d\right) = \hat{f}_1\left(x_1\right) \hat{f}_2\left(x_2\right) \cdots \hat{f}_d\left(x_d\right)$. Thus, as k grows, the set of potential estimates grows from naive Bayes models when $k = 1$ to all possible histogram estimates.

The models in this paper are motivated by nonnegative tensor factorizations so we term them generally *nonparametric nonnegative tensor factorization* (NNTF) *models*. The first model is related to nonnegative PARAFAC [27]. Our second model is based on the nonnegative Tucker decomposition [16]. In this model it is assumed that there are d collections of k one-dimensional densities, $\mathcal{F}_1, \ldots, \mathcal{F}_d$ with $\mathcal{F}_i = \{f_{i,1}, \ldots, f_{i,k}\}$, and some probability measure that randomly selects one density from each $\mathcal{F}_i$. This measure a can be represented by a tensor $W \in \mathbb{R}^{k^{\times d}}$ where the probability of selecting $f_{1,i_1}, \ldots, f_{d,i_d}$ is $W_{i_1,\ldots,i_d}$. To sample from this model we first randomly select the marginal distributions $f_{1,i_1}, \ldots, f_{d,i_d}$ according to W and then independently sample each feature according the randomly selected marginal distribution $X_j \sim f_{j,i_j}$. The density of this model is

$$f\left(x_1, \ldots, x_d\right) = \sum_{i_1=1}^{k} \cdots \sum_{i_d=1}^{k} W_{i_1,\ldots,i_d} f_{1,i_1}(x_1) \cdots f_{d,i_d}(x_d). \tag{2}$$

We are unaware of previous literature investigating this model so we will simply term it the *Tucker model*. Again we will investigate estimators where the marginals are one dimensional histograms and small k corresponds to reduced model capacity.

1.2 Overview of Results

The principal contribution of this paper is to analyze the advantage of incorporating rank-restriction into density estimation. In Section 2.1 we precisely introduce the multi-view histogram and a

histogram based estimator whose rank is restricted via Tucker decomposition. We are the first to introduce the latter model in the nonparametric setting, as far as we know. Given these models we then investigate how fast we can let bin width h and the estimator rank k grow in relation to the amount of data n while still being able to control estimator error[†] and selects an estimator that is nearly optimal for the allowable set. For that the multi-view histogram we can control error so long as k/h is asymptotically dominated by n and for the Tucker Histogram we need $k/h + k^d$ to be asymptotically dominated by n (both of these rates ignore logarithmic factors). This stands in stark contrast to standard space of histogram estimators that requires h^{-d} to be asymptotcially dominated by n. We furthermore show that these estimators are universally consistent and show that these rates cannot be significantly improved.

For a second style of analysis we provide finite sample bounds on the convergence of low-rank histogram estimators and show that utilizing low rank structure one can construct a class of universally consistent density estimators that converges at rate $\widetilde{O}(1/\sqrt[3]{n})$[‡] when the target distribution is a multi-view density with Lipschitz continuous marginals. We further show that there exists simple multi-view densities where the standard histogram converges at rate $\omega(1/\sqrt[d]{n})$[§] In Section 3 we show that we can use an existing low-rank nonnegative Tucker factorization algorithm to fit our model and demonstrate empirically that fitting histograms to a Tucker model outperforms the standard histogram estimator with very high statistical significance.

1.3 Previous Work

Nonparametric density estimation has been extensively studied with the histogram estimator and KDE being some of the most well-known methods. There do exist, however, alternative methods for density estimation, e.g. the *forest density estimator* [18] and *k-nearest neighbor density estimator* [19]. The L_1, L_2 and L_∞ convergence of the histogram and KDE has been studied extensively [11, 7, 31, 13]. The KDE is generally accepted as being the superior density estimator with some mathematical justification [28]. Numerous modifications and extensions of the KDE have been proposed including utilizing variable bandwidth [30], robust KDEs [14, 32, 33], and methods for enforcing support boundary constraints [26]. Finally we mention [15] that demonstrated that uniform convergence of a KDE to its population estimate suffered when the intrinsic dimension of the data was lower than the ambient dimension, a phenomenon seemingly at odds with the curse of dimensionality.

For our review of NNTF models we also include a general review of tensor/matrix factorizations since both can be viewed being low-rank models. In particular, for the multi-view model we have

$$\sum_{i=1}^{k} w_i f_{i,1}(x_1) \cdots f_{i,d}(x_d) \sim \sum_{i=1}^{k} \lambda_i \mathbf{v}_{i,1} \otimes \cdots \otimes \mathbf{v}_{i,d}. \tag{3}$$

A great deal of work has gone into leveraging low-rank assumptions to improve matrix estimation, particularly in the field of *compressed sensing* [8, 23]. The most basic version of compressed sensing is concerned with estimating a "tall" matrix x with access to a "short" vector $y \triangleq Ax$ and A is a known "short and fat" matrix. One can recover y assuming it is sparse and M satisfies a property known as the restricted isometry property (RIP) [34, 5]. These methods can be extended to the estimation of matrices where, x is now a low rank matrix[20, 21]. Additionally in this extension, A is an order-3 tensor which acts as a linear operator on x and satisfies a property called "restricted strong convexity" (RSC). These properties, RIP and RSC, arise from matrices and tensors who's entries are random. Because of this compressed sensing techniques are useful for problems where one wants to estimate x and has access to data which are random linear transforms of x. For example with matrix completion a random observation of an entry is like a vector product of x with a random matrix where one entry is set to one and A is a stack of such matrices representing the collection of all observed entries. These methods have also proven useful for multivariate regression and

[†]For an estimator V restricted space of densities $\mathcal{P}$, the *estimation error* refers the difference between $\|V - p\|$ and $\min_{q \in \mathcal{P}} \|q - p\|$, where p is the target density.

[‡]$f_n \in \widetilde{O}(g_n) \iff \exists \varepsilon > 0, k \forall N, M$ s.t. $P\left(f_n / \left(g_n + log^k g_n\right) \geq M\right) \to 0$

[§]$f_n \in \omega(g_n) \iff |f_n / g_n| \to \infty$ in probability. Basically in d dimensional space the standard histogram can converge slower than $1/\sqrt[d]{n}$ for d- dimensional space. For example, when estimating the d dimensional density on $[0,1]^d$, $p(x_1, x_2, \ldots, x_d) = 2x$ the histogram will converge slower than $1/\sqrt[d]{n}$ whereas the proposed histogram converges at around approximately $1/\sqrt[3]{n}$ rate.

autoregressive models[20, 21]. Despite our efforts, such techniques don't appear to be extensible to histogram estimation because there doesn't seem to be way to construe the data in our setting random to linear sampling in compressed sensing.

General matrix/tensor factorization, including nonnegative matrix/tensor factorizations, have been extensively studied despite their inherent difficulty due to non-convexity. The works [9, 3] present potential theoretical grounds for avoiding the computational difficulties of nonnegative matrix factorization. One notable approach to tensor factorization is to assume, in the tensor representation in (3), that $d \geq 3$ and the collections of vectors $\mathbf{v}_{1,j}, \ldots, \mathbf{v}_{k,j}$ are linearly independent for all j. Under this assumption we are guaranteed that the factorization (3) is unique [1]. In [2] the authors present a method for recovering this factorization efficiently and demonstrate its utility for a variety of tasks. This work was extended in [29] to recover a multi-view KDE satisfying an analogous linear independence assumption. This is the only work of its type of which we are aware. In this work the authors investigate the sample complexity of their estimator but do not demonstrate that their technique has potential for improving rates for nonparametric density estimation in general. Finally we note that nonparametric applications of the Tucker decomposition have been utilized in Bayesian statistics [25]. We are unaware of any literature describing the model we introduce in (2).

2 Theoretical Results

In this section we mathematically demonstrate that histogram estimators can achieve greater performance by restricting to NNTF models and using the proper procedure to select a representative of the histogram estimators from data. To simplify analysis we will only consider densities on $[0,1)^d$ and analyze number of bins per dimension b which is the inverse of the bin width, i.e. $b = 1/h$. To state the results here precisely we must introduce a fair amount of notation.

2.1 Notation

We will denote the L_1 and L_2 Lebesgue space norms via a 1 or 2 subscript. Let $\mathcal{D}_d$ be the set of all densities on $[0,1)^d$. By *density* we mean probability measures that are absolutely continuous with respect to the d-dimensional Lebesgue measure. We define a *probability vector* or *probability tensor* to simply mean a vector or tensor whose entires are nonnegative and sum to one. Let Δ_b denote the set of probability vectors in $\mathbb{R}^b$and $\mathcal{T}_{d,b}$ the set of probability tensors in $\mathbb{R}^{b^{\times d}}$. The product symbol $\prod$ will always mean the standard outer product, e.g. set product or tensor product, when the multiplicands are not real numbers. For any natural number b let $[b] = \{1, \ldots, b\}$. Finally $\mathbb{1}$ is the indicator function and Conv is the convex hull.

We will now construct the space of histograms on $[0,1)^d$. We begin with one-dimensional histograms, which will serve as the $\hat{f}_{i,j}$ terms in (1) or (2). We define $h_{1,b,i}$ with $i \in [b]$ to be the one dimensional histogram where all weight is allocated to the ith bin. Formally we define this as

$$h_{1,b,i}(x) \triangleq b\mathbb{1}\left(\frac{i-1}{b} \leq x < \frac{i}{b}\right).$$

Note that this is a valid density due to the leading b coefficient. We use these to construct higher-dimensional histograms. For a multi-index $A \in [b]^d$, let

$$h_{d,b,A} \triangleq \prod_{i=1}^{d} h_{1,b,A_i},$$

the histogram whose entire density is allocated to the bin indexed by A. Finally we define $\Lambda_{d,b,A}$ to be the support of $h_{d,b,A}$, i.e. the "bins" of a histogram estimator,

$$\Lambda_{d,b,A} \triangleq \prod_{i=1}^{d} \left[\frac{A_i - 1}{b}, \frac{A_i}{b}\right).$$

For a sequence of points in $[0,1)^d$, $\mathcal{X} = (X_1, \ldots, X_n)$, the standard histogram estimator is

$$H_{d,b}(\mathcal{X}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \sum_{A \in [b]^d} h_{d,b,A} \mathbb{1}\left(X_i \in \Lambda_{d,b,A}\right).$$

Let $\mathcal{H}_{d,b} \triangleq \text{Conv}\left(\left\{h_{d,b,A} \middle| A \in [b]^d\right\}\right)$, the set of all d-dimensional histograms with b bins per dimension. Let $\mathcal{H}_{d,b}^k$ be the set of histograms with at most k separable components, i.e.

$$\mathcal{H}_{d,b}^k \triangleq \left\{ \sum_{i=1}^k w_i \prod_{j=1}^d f_{i,j} \middle| w \in \Delta_k, f_{i,j} \in \mathcal{H}_{1,b} \right\}. \tag{4}$$

We will refer elements in this space *multi-view histograms*. Analogously we define the space of *Tucker histograms* to be

$$\widetilde{\mathcal{H}}_{d,b}^k = \left\{ \sum_{S \in [k]^{\times d}} W_S \prod_{i=1}^d f_{i,S_i} \middle| W \in \mathcal{T}_{d,k}, f_{i,j} \in \mathcal{H}_{1,b} \right\}.$$

We emphasize that the collections of densities $\mathcal{H}_{d,b}^k$ and $\widetilde{\mathcal{H}}_{d,b}^k$ the primary objects of interest in this paper as they represent NNTF histograms. The theoretical results we present are concerned with finding good density estimators restricted to these sets as k and b vary.

2.2 Estimator Theoretical Results

We present two approaches to analysis of histogram style estimators. All proofs are in the Appendix.

First we provide an asymptotic analysis of low-rank histogram estimators in terms of error control: how fast can we let b and k grow, with respect to n, while still controlling estimator error over all densities. An advantage of this analysis is that we can demonstrate that these rates are approximately optimal (up to logarithmic terms).

For our second approach we present analysis based on a finite sample bounds. For brevity, and because the results are virtually direct analogues of their multi-view histogram counterparts we reserve all results for Tucker histogram for the Appendix. In this analysis we first we present distribution-dependent bound that, for an estimator in $\mathcal{H}_{d,b}^k$, depends on n, b, k and $\min_{p^\star \in \mathcal{H}_{d,b}^k} \|p^\star - p\|_1$ where p is distribution to be estimated.

Distribution free bounds require that the target distribution belong to a well-behaved class of distributions such as Sobolev or Hölder [31]. Our distribution-free finite sample analyses are densities that satisfy the multi-view model assumption and who's multi-view components have marginals that are Lipschitz continuous. To find the bias of our estimators we first find general bounds on $\min_{h \in \mathcal{H}_{d,b}^1} \left\| h - \prod_{i=1}^d p_i \right\|_1$. Finally we find a finite sample bound on estimators that selects an element $\mathcal{H}_{d,b}^k$ to estimate a density of the form $\sum_{i=1}^m w_i \prod_{j=1}^d p_{i,j}$, where $p_{i,j}$ all lie in some Lipschitz class, converges at a rate of approximately $1/\sqrt[3]{n}$, independent of d. This is in comparison to the standard histogram estimator which can converge at rates worse than $1/\sqrt[3]{d}$

Main Technical Tools Our results rely finding L^1 ε-coverings of the spaces of histogram estimators and using Theorem 3.7 from [4] to select a good representative from that collection. The aforementioned theorem is a slight extension of Theorem 6.3 in [7] and is essentially proven in [35]. As was mentioned in [4] application of these results do not yield an computationally practical algorithm but are simply meant to highlight the potential of such models. Additional results are contained in the Appendix.

2.2.1 Asymptotic Error Control

The following theorem states that one can control the estimation error of multi-view histograms with k components and b bins per dimension so long as $n \sim bk$ (omitting logarithmic factors). Recall that the standard histogram estimator requires $n \sim b^d$, so we have removed the exponential dependence of bin rate on dimensionality. Here and elsewhere the $\sim$ symbol is not a precise mathematical statement but rather describes that the two values should be of the same order. In the following b and k are functions of n; the space of histograms which we are fitting changes as we acquire more data.

Theorem 2.1. *For any pairs of sequences $b \to \infty$ and $k \to \infty$ satisfying $n/(bk\log(b)+k\log(k)) \to \infty$, there exists an estimator $V_n \in \mathcal{H}_{d,b}^k$ such that, for all $\varepsilon > 0$*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\mathcal{H}_{d,b}^k} \|p-q\|_1 + \varepsilon\right) \to 0,$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

The sample complexity for the multi-view histogram is perhaps more accurately approximated as being on the order of dbk however the d disappears in the asymptotic analysis. The following theorem states that we can control the error of Tucker histogram estimates so long as $n \sim bk + k^d$ (omitting logarithmic factors).

Theorem 2.2. *For any pairs of sequences $b \to \infty$ and $k \to \infty$ satisfying $n/(bk\log(b) + k^d \log\left(k^d\right)) \to \infty$, there exists an estimator $V_n \in \widetilde{\mathcal{H}}_{d,b}^k$ such that, for all $\varepsilon > 0$*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\widetilde{\mathcal{H}}_{d,b}^k} \|p-q\|_1 + \varepsilon\right) \to 0,$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Allowing b to grow as aggressively as possible we achieve consistent estimation, using either the multi-view or Tucker histograms, so long as $n \sim b\log b$ regardless of dimensionality.

Corollary 2.1. *For all d, b, k fix $\mathcal{R}_{d,b}^k$ to be either $\mathcal{H}_{d,b}^k$ or $\widetilde{\mathcal{H}}_{d,b}^k$. For any sequence $b \to \infty$ with $n/\left(b\log b\right) \to \infty$, there exists a sequence $k \to \infty$ and estimator $V_n \in \mathcal{R}_{d,b}^k$ such that, for all $\varepsilon > 0$*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\mathcal{R}_{d,b}^k} \|p-q\|_1 + \varepsilon\right) \to 0,$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Replacing $b := 1/h$ allows us to arrive at the rates mentioned in Section 1. The following result shows that the bias of the estimators in Theorems 2.1 and 2.2 go to zero for all densities. Thus these estimators are universally consistent even when the NNTF model assumption is not satisfied.

Lemma 2.1. *Let $p \in \mathcal{D}_d$. If $k \to \infty$ and $b \to \infty$ then $\min_{q\in\mathcal{H}_{d,b}^k} \|p-q\|_1 \to 0$.*

A straightforward consequence of this is that the Tucker histogram bias also goes to zero.

Lemma 2.2. *Let $p \in \mathcal{D}_d$. If $k \to \infty$ and $b \to \infty$ then $\min_{q\in\widetilde{\mathcal{H}}_{d,b}^k} \|p-q\|_1 \to 0$.*

Finally we have that rate on bk in Theorem 2.1 cannot be made significantly faster.

Theorem 2.3. *Let $d \geq 2$, $b \to \infty$, and $k \to \infty$ with $b \geq k$ and $n/\left(bk\right) \to 0$. There exists no estimator $V_n \in \mathcal{H}_{d,b}^k$ such that, for all $\varepsilon > 0$, the following limit holds*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\mathcal{H}_{d,b}^k} \|p-q\|_1 + \varepsilon\right) \to 0$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

Likewise the rate on $bk + k^d$ can also not be significantly improved in Theorem 2.2.

Theorem 2.4. *Let $d \geq 2$, $b \to \infty$, and $k \to \infty$ with $b \geq k$ and $n/\left(bk + k^d\right) \to 0$. There exists no estimator $V_n \in \widetilde{\mathcal{H}}_{d,b}^k$ such that, for all $\varepsilon > 0$, the following limit holds*

$$\sup_{p\in\mathcal{D}_d} P\left(\|V_n - p\|_1 > 3 \min_{q\in\widetilde{\mathcal{H}}_{d,b}^k} \|p-q\|_1 + \varepsilon\right) \to 0$$

where V_n is a function of $X_1, \ldots, X_n \overset{iid}{\sim} p$.

2.2.2 Finite Rate Sample Bounds and Other Results

In this section we will investigate the convergence of multi-view histogram estimators to densities which satisfy the multi-view assumption. This will be done via a standard bias/variance decomposition style of argument. We begin with the following finite sample bound that can holds for *any* density p

Proposition 2.1 (Finite sample bound)**.** *Let p be a density in $[0,1]^d$ and fix a number of bins b and a maximum rank k. For any number of samples n, there exists an estimator V_n in $\mathcal{H}^k_{d,b}$ such that for all $\delta \leq 1$, with probability $\geq 1-\delta$ over the draw of the samples, the following inequality holds:*

$$\|p - V_n\|_1 \leq 3\|p - p^*\|_1 + 7\sqrt{\frac{2bdk\log(4bdkn)}{n}} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}, \tag{5}$$

where p^* is any density with

$$p^* \in \operatorname{argmin}_q(\|q-p\|_1 : q \in \mathcal{H}^k_{d,b}).$$

Here $\|p - p^\star\|_1$ plays the role of bias.

To analyze the bias term, $\|p - p^\star\|_1$, first consider the case where p has a single component with L-Lipschtiz marginals, so $p = \prod_{i=1}^d f_i$. Using the indicator function over the unit cube with Hölder's Inequality we have that $\|p - p^\star\|_1 \leq \|p - p^\star\|_2$ where its possible to show that the best L^2projection of p onto $\mathcal{H}^k_{d,b}$ is achieved by simply projecting each marginal to its best approximating one-dimensional marginal i.e.

$$\arg\min_{p^\star \in \mathcal{H}^k_{d,b}} \left\| \prod_{i=1}^d p_i - p^\star \right\|_2 = \prod_{i=1}^d \operatorname{Proj}_{\mathcal{H}_{1,b}} p_i. \tag{6}$$

Let Lip_L be the set of L-Lipschitz probability density functions $[0,1]$ and m_L be $\sup_{f \in \mathrm{Lip}_L} \|f\|_2$. We have the following

Theorem 2.5. *Let $b^2 \geq L^2/12$ and $f_1, \ldots, f_d$ be elements of Lip_L then*

$$\left\| \prod f_i - \operatorname{Proj}_{\mathcal{H}^k_{d,b}} \prod f_i \right\|_1 \leq \sqrt{m_L^{2d} - \left(m_L^2 - \frac{L^2}{12b^2}\right)^d}. \tag{7}$$

We show in the Appendix that this decays at a rate of $O\left(b^{-1}\right)$. To characterize the behavior of m_L we have the following.

Proposition 2.2.

$$m_L^2 = \begin{cases} L^2/12 + 1 & 0 \leq L \leq 2 \\ \sqrt{\frac{8L}{9}} & L \geq 2 \end{cases} \tag{8}$$

We remark that the previous proposition is tight with equality when $L \leq 2$ or when $\sqrt{2/L}$ is a constant multiple of $1/b$ (see the Appendix). The following theorem characterizes the rate of convergence of our estimators to a class of multi-view models.

Theorem 2.6. *Let $p \triangleq \sum_{i=1}^m w_i \prod_{i=1}^d p_{i,j}$ where $p_{i,j}$ are L-Lipschitz continuous[¶] densities on $[0,1]$, with $L \geq 2$, and w is in the probability simplex. Assuming we know an upper bound k on the rank m, there exists an estimator $V_n \in \mathcal{H}^k_{d,b}$ such that with probability $\geq 1-\delta$,*

$$\|V_n - p\|_1 \leq \frac{30dk^{1/3}L^{\frac{d+3}{12}}}{n^{\frac{1}{3}}}\sqrt{\log(3Ldkn)} + 7\sqrt{\frac{\log(\frac{3}{\delta})}{2n}}. \tag{9}$$

These results imply a rate of convergence of $\tilde{O}\left(\sqrt[3]{n}\right)$ of these estimators to multi-view models with Lipschits components, independent of dimension! This stand in stark contrast to the rate of convergence of the standard histogram estimator whose rate of convergence is hindered in an exponental way in dimension.

Proposition 2.3. *Let V_n be the standard histogram estimator with $n/b^d \to \infty$. There exists a density of the form in Theorem 2.6 such that $\|V_n - p\|_1 \in \omega(1/\sqrt[d]{n})$.*

[¶]Note that any L'-Lipschitz continuous function is trivially L-Lipschitz continuous for $L > L'$

2.3 Discussion

Naturally real world data likely never exactly satisfies the NNTF model assumption. Our results are meant to highlight a trade-off between model assumptions of smoothness (low b) and simple dependence between features (low k). Here we will explore this trade-off for multi-view histogram. Letting $k = b^d$ gives $\mathcal{H}^k_{d,b} = \mathcal{H}_{d,b}$ since we can allocate one component to each bin. Using the estimator from Theorem 2.1 gives a sample complexity of approximately b^{d+1}. Thus setting $k = b^d$ in the multi-view histogram gives us something which behaves similarly to the standard histogram estimator with a similar sample complexity. On the other hand setting $k = 1$ gives a naive Bayes model with a sample complexity of approximately b. The Tucker histogram can be similarly analyzed with $k = b$ corresponding to the standard histogram. Thus we have a span of k yielding different estimators with maximal k corresponding to the standard histogram and minimal k corresponding to a naive Bayes assumption. We observe in Section 3 that this trade-off is useful in practice: we virtually never want k to be maximized.

3 Experiments

The previous sections have demonstrated the existence of estimators that select low-rank histograms that offer advantages over the standard histogram. Unfortunately these estimators are not computationally tractable and only give us a glimpse of what benefits may come from nonparametric density estimation that additionally incorporate low-rank restriction. It would be nonetheless interesting to observe some of the behavior of a low-rank histogram estimators, even if it were less principled. To this end we consider an L^2 minimizing low rank histogram estimator. For all d, b, k fix $\mathcal{R}^k_{d,b}$ to be either $\mathcal{H}^k_{d,b}$ or $\widetilde{\mathcal{H}}^k_{d,b}$. We consider estimator U_n, that attempts to minimize

$$U_n = \operatorname{argmin}_{\hat{p} \in \mathcal{R}_{d,b^k}} \|\hat{p} - p\|_2^2 . \tag{10}$$

Note that

$$\|\hat{p} - p\|_2^2 = \|\hat{p}\|_2^2 - 2 \langle p, \hat{p} \rangle + \|p\|_2^2 . \tag{11}$$

Further note that $\|p\|_2^2$ in no way concerns the performance of $\hat{p}$. Additionally for data $X_1, \ldots, X_n \overset{iid}{\sim} p$ the law of large numbers gives us $\langle p, \hat{p} \rangle \approx \sum_{i=1}^m \hat{p}(X_i)$ so we may consider

$$U_n \triangleq \operatorname{argmin}_{\hat{p} \in \mathcal{R}_{d,b}} = \|\hat{p}\|_2^2 - 2 \sum_{i=1}^m \hat{p}(X_i) \tag{12}$$

as a practical version of (11) which is conveniently equivalent to nonnegative tensor factorization (see Appendix).

Now we must strongly emphasize to the reader that (12), while a very effective estimator, is not meant as a practical contribution from this paper, but simply something to compare to the standard histogram. This begs the question as to why this estimator was not theoretically analyzed. There are a few reasons for this. In the classic nonparametric density estimation book [7] the authors dedicate an entire section to why L^2 distances are to be avoided (see §6.5 L^2 Distances are to be Avoided). The (12) and similar risks have appeared in previous works on adaptations of nonparametric density estimators. Either this sort risk produces inherently slower estimators or the correct approach to analyzing them has not been found as they seem to produce slower rates [6]. This problem when seems be further exacerbated when enforcing a low-rank structure [24]

For our experiments we used the Tensorly library [17] to perform the nonnegative Tucker decomposition [16] with Tucker rank $[k, k, \ldots, k]$ which was then projected to the simplex of probability tensors using [10]. We also performed experiments with nonnegative PARAFAC decompositions using [27, 17]. These decompositions performed poorly. This is potentially because the PARAFAC optimization is more difficult or the additional flexibility of the Tucker decomposition was more appropriate for the experimental datasets.

3.1 Experimental Setup and Results

Our experiments were performed on the Scikit-learn datasets MNIST and Diabetes [22], with labels removed. We used the remaining holdout data with (12) to evaluate performance (which

Table 1: Experimental Results

Dataset	d Red.	Dim.	Hist. Perf.	Tucker Perf.	Hist. Bins	Tucker Bins	Tucker r	p-val.
MNIST	PCA	2	-1.455±0.089	-1.502±0.102	6.531±1.499	8.375±1.780	4.968±1.976	5e-4
		3	-2.040±0.196	-2.268±0.195	4.781±0.738	6.718±1.565	5.781±1.340	2e-4
		4	-3.532±0.996	-4.014±0.655	4.031±0.585	5.343±1.018	4.375±0.695	2e-3
		5	-4.673±1.026	-6.157±2.924	3.468±0.499	4.343±0.592	3.281±0.514	4e-5
	Rand.	2	-2.034±0.100	-2.099±0.102	6.062±1.197	7.562±1.657	2.062±1.784	3e-5
		3	-3.086±0.207	-3.331±0.387	4.812±0.526	6.843±1.227	2.687±1.959	1e-4
		4	-4.307±0.290	-5.731±0.435	3.500±0.559	5.656±0.642	2.593±1.497	8e-7
		5	-6.327±0.522	-9.539±1.053	3.250±0.433	4.718±0.571	2.562±1.087	8e-7
Diabetes	PCA	2	-2.079±0.122	-2.212±0.132	5.718±1.304	7.468±1.478	1.062±0.242	8e-6
		3	-3.010±0.364	-3.606±0.420	3.593±0.860	7.062±1.058	1.843±1.543	2e-6
		4	-4.002±0.415	-4.423±0.701	3.000±0.000	5.906±0.804	2.343±1.107	2e-3
		5	-6.139±0.661	-6.043±1.192	3.000±0.000	3.750±0.968	1.843±0.617	0.91
	Rand.	2	-3.074±0.224	-3.277±0.287	6.843±1.227	9.250±1.936	1.093±0.384	7e-5
		3	-4.726±0.483	-5.353±0.751	4.968±0.769	8.406±1.343	1.625±1.672	2e-5
		4	-6.017±0.873	-7.732±1.497	4.062±0.704	6.718±1.328	2.093±1.155	1e-5
		5	-8.986±1.292	-12.61±2.477	3.062±0.242	5.093±0.521	2.531±0.865	2e-6

may cause negative performance values). Our experiments considered estimating histograms in $d = 2, 3, 4, 5$ dimensional space. We consider two forms of dimensionality reduction. First we consider projecting the dataset onto its top d principle components. As an alternative we consider projecting our dataset onto a random subspace of dimension d. We have constructed our random subspace dimensionality reduction so that each additional dimension adds a new index without affecting the others. For each dataset we randomly select an orthonormal basis that remains unchanged for all experiments $v_1, v_2, \ldots$. To transform a point X to dimension d we perform the following transform: $X_{\text{reduced dim.}} = [v_1 \cdots v_d]^T X$. We consider both transforms since PCA may select dimensions where the features tend to be independent, as is the case when the distribution is a multivariate Gaussian. After dimensionality reduction we scale and translate the data to fit in the unit cube.

With our preprocessed dataset, each experiment consisted of randomly selecting 200 samples for training and using the rest to evaluate performance. For the estimators we tested all combinations using 1 to b_{max} bins per dimension and k from 1 to k_{max}. As d increased the best cross validated b and k value decreased, so we reduced b_{max} and k_{max} for larger d to reduce computational time, while still leaving a sizable gap between the best cross validated b and k across all runs of all experiment. For $d = 2, 3$ we have $b_{\text{max}} = 15$ and $k_{\text{max}} = 10$; for $d = 4$ we have $b_{\text{max}} = 12$ and $k_{\text{max}} = 8$; for $d = 5$ we have $b_{\text{max}} = 8$ and $k_{\text{max}} = 6$. For parameter fitting we used random subset cross validation repeated 80 times using 40 of the 200 samples to cross validate. Performing 80 folds of cross validation was necessary because of the variance of the histogram estimated loss. This high variance is likely due to the noncontinuous nature of the histogram estimator itself and the noncontinuity of the histogram as a function of the data, i.e. slightly moving one training sample can potentially the change histogram bin in which it lies. Each experiment was run 32 times and we report the mean and standard deviations of estimator performance as well as the best parameters found from cross validation. We additionally apply the Wilcoxon signed rank test to the 32 pairs of performance results to statistically determine if the mean performance between the standard histogram and our algorithm are different and report the corresponding p-value. Our results are in Table 3 where the Tucker histogram dominates.

Societal Impact

It seems unlikely that our method will have significant societal impact since it is simply an improvement of a standard statistcal technique.

References

[1] Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[3] Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC '12, pages 145–162, New York, NY, USA, 2012. ACM.

[4] Hassan Ashtiani, Shai Ben-David, Nicholas Harvey, Christopher Liaw, Abbas Mehrabian, and Yaniv Plan. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3412–3421. Curran Associates, Inc., 2018.

[5] Emmanuel J Candes and Terence Tao. Reflections on compressed sensing. *IEEE Information Theory Society Newsletter*, 58(4):20–23, 2008.

[6] Efrén Cruz Cortés and Clayton Scott. Consistent kernel density estimation with non-vanishing bandwidth, 2017.

[7] L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, New York, 2001.

[8] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, April 2006.

[9] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In S. Thrun, L. K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1141–1148. MIT Press, 2004.

[10] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the l₁-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 272–279, New York, NY, USA, 2008. Association for Computing Machinery.

[11] Làszlò Györfi, Luc Devroye, and Laszlo Gyorfi. *Nonparametric density estimation: the L1 view*. John Wiley & Sons, New York; Chichester, 1985.

[12] Nils-Bastian Heidenreich, Anja Schindler, and Stefan Sperlich. Bandwidth selection for kernel density estimation: a review of fully automatic selectors. *AStA Advances in Statistical Analysis*, 97(4):403–433, Oct 2013.

[13] Heinrich Jiang. Uniform convergence rates for kernel density estimation. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1694–1703, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[14] J. Kim and C. Scott. Robust kernel density estimation. *J. Machine Learning Res.*, 13:2529–2565, 2012.

[15] Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform Convergence Rate of the Kernel Density Estimator Adaptive to Intrinsic Volume Dimension. *arXiv e-prints*, page arXiv:1810.05935, Oct 2018.

[16] Yong-Deok Kim and Seungjin Choi. Nonnegative tucker decomposition. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[17] Jean Kossaifi, Yannis Panagakis, Anima Anandkumar, and Maja Pantic. TensorLy: Tensor Learning in Python. *arXiv e-prints*, page arXiv:1610.09555, Oct 2016.

[18] Han Liu, Min Xu, Haijie Gu, Anupam Gupta, John Lafferty, and Larry Wasserman. Forest density estimation. *J. Mach. Learn. Res.*, 12:907–951, July 2011.

[19] Y. P. Mack and M. Rosenblatt. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9(1):1–15, March 1979.

[20] Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.*, 39(2):1069–1097, 04 2011.

[21] Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *J. Mach. Learn. Res.*, 13:1665–1697, May 2012.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[23] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52(3):471–501, August 2010.

[24] Alexander Ritchie, Robert A Vandermeulen, and Clayton Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11676–11686. Curran Associates, Inc., 2020.

[25] Aaron Schein, Mingyuan Zhou, David Blei, and Hanna Wallach. Bayesian poisson tucker decomposition for learning the structure of international relations. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2810–2819, New York, New York, USA, 20–22 Jun 2016. PMLR.

[26] Eugene F. Schuster. Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, 14(5):1123–1136, 1985.

[27] Amnon Shashua and Tamir Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 792–799, New York, NY, USA, 2005. ACM.

[28] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[29] Le Song, Animashree Anandkumar, Bo Dai, and Bo Xie. Nonparametric estimation of multi-view latent variable models. *CoRR*, abs/1311.3287, 2013.

[30] George R. Terrell and David W. Scott. Variable kernel density estimation. *Ann. Statist.*, 20(3):1236–1265, 09 1992.

[31] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

[32] R. Vandermeulen and C. Scott. Consistency of robust kernel density estimators. *COLT*, 30:568–591, 2013.

[33] Robert A Vandermeulen and Clayton Scott. Robust kernel density estimation by scaling and projection in hilbert space. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 433–441. Curran Associates, Inc., 2014.

[34] Roman Vershynin. High-dimensional probability. 2019.

[35] Yannis G. Yatracos. Rates of Convergence of Minimum Distance Estimators and Kolmogorov's Entropy. *The Annals of Statistics*, 13(2):768 – 774, 1985.

For all authors...

1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
2. Did you describe the limitations of your work? [Yes]
3. Did you discuss any potential negative societal impacts of your work? [Yes]
4. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

If you are including theoretical results...

1. Did you state the full set of assumptions of all theoretical results? [Yes]
2. Did you include complete proofs of all theoretical results? [Yes] Complete proofs are in the Appendix, with a few notes on general methods in main text

If you ran experiments...

1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]

2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]

If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

1. If your work uses existing assets, did you cite the creators? [N/A]
2. Did you mention the license of the assets? [N/A]
3. Did you include any new assets either in the supplemental material or as a URL? [N/A]
4. Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

If you used crowdsourcing or conducted research with human subjects...

1. Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
2. Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]