Sparse Nonparametric Density Estimation in High Dimensions Using the Rodeo

Han Liu\(^1, 2\)  John Lafferty\(^2, 3\)  Larry Wasserman\(^1, 2\)

\(^1\)Statistics Department, \(^2\)Machine Learning Department, \(^3\)Computer Science Department, Carnegie Mellon University

July 1st, 2006
Rodeo is a general strategy for nonparametric inference. It has been successfully applied to solve sparse nonparametric regression problems in high dimensions by Lafferty & Wasserman, 2005.

Our goal

Trying to adapt the rodeo framework to nonparametric density estimation problems. So that we have a unified framework for both density estimation and regression problems which is computationally efficient and theoretically soundable.
1 Background
   - Nonparametric density estimation in high dimensions
   - Sparsity assumptions for density estimation

2 Methodology and Algorithms
   - The main idea
   - The local rodeo algorithm for the kernel density estimator

3 Asymptotic Properties
   - The asymptotic running time and minimax risk

4 Extension and Variations
   - The global density rodeo and the reverse density rodeo
   - Using other distributions as irrelevant dimensions

5 Experimental Results
   - Empirical results on both synthetic and real-world datasets
Problem statement

Problem

To estimate the joint density of a continuous $d$-dimensional random vector

$$X = (X_1, X_2, ..., X_d) \sim \mathcal{F}, \quad d \geq 3$$

where $\mathcal{F}$ is the unknown distribution with density function $f(x)$.

This problem is essentially hard, since the high dimensionality causes both computational and theoretical problems.
From a frequentist perspective
- Kernel density estimation and the local likelihood method
- Projection pursuit method
- Log-spline models and the penalized likelihood method

From a Bayesian perspective
- Mixture of normals with Dirichlet processes as prior

Difficulties of current approaches
- Some methods only work well for low-dimensional problems
- Some heuristics lack the theoretical guarantees
- More importantly, they suffer from the curse of dimensionality
The curse of dimensionality

Characterizing the curse

In a Sobolev space of order $k$, minimax theory shows that the best convergence rate for the mean squared error is

$$R_{opt} = O \left( n^{-2k/(2k+d)} \right)$$

which is practically slow when the dimension $d$ is large.

Combating the curse by some sparsity assumptions

If the high-dimensional data has a low dimensional structure or a sparsity condition, we expect that some methods could be developed to combat the curse of dimensionality. This motivates the development of the rodeo framework.
Rodeo (regularization of derivative expectation operator) is a general strategy for nonparametric inference. Which has been used for nonparametric regression.

For a regression problem

\[ Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n \]

where \( X_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d \) is a \( d \)-dimensional covariate. If \( m \) is in a \( d \)-dimensional Sobolev space of order 2, the best convergence rate for the risk is

\[ R^* = O \left( n^{-4/(4+d)} \right) \]

Which shows the curse of dimensionality in a regression setting.
Rodeo for nonparametric regression (II)

Assume the true function only depends on \( r \) covariates \( (r \ll d) \)

\[
m(x) = m(x_1, \ldots, x_r)
\]

for any \( \epsilon > 0 \), the rodeo can simultaneously perform bandwidth selection and (implicitly) variable selection to achieve a better minimax convergence rate of

\[
R_{\text{rodeo}} = O \left( n^{-4/(4+r)+\epsilon} \right)
\]

as if the \( r \) relevant variables were explicitly isolated in advance.

Rodeo beats the curse of dimensionality in this sense. We expect to apply the same idea to solve density estimation problems.
Sparse density estimation

For many applications, the true density function can be characterized by some low dimensional structure.

**Sparsity assumption for density estimation problems**

Assume $h_{jj}(x)$ is the second partial derivative of $h$ on the $j$-th variable, there exists some $r \ll d$, such that

$$f(x) \propto g(x_1, ..., x_r)h(x) \text{ where } h_{jj}(x) = 0 \text{ for } j = 1, ..., d.$$  

Where $x_R = \{x_1, ..., x_r\}$ are the **relevant** dimensions. This condition imposes that $h(\cdot)$ belongs to a family of very smooth functions (e.g. the uniform distribution).

$h(\cdot)$ can be generalized to be any parametric distribution!
Generalized sparse density estimation

We can generalize $h(\cdot)$ to other distributions (e.g. Gaussian).

General sparsity assumption for density estimation problems

Assume $h(\cdot)$ is any distribution (e.g. Gaussian) that we are not interested in

$$f(x) \propto g(x_1, \ldots, x_r) h(x) \quad \text{where} \quad r \ll d.$$  

Thus, the density function $f(\cdot)$ can be factored into two parts: the relevant components $g(\cdot)$ and the irrelevant components $h(\cdot)$. Where $x_R = \{x_1, \ldots, x_r\}$ are the relevant dimensions.

Under this framework, we can hope to achieve a better minimax rate

$$\mathcal{R}_{\text{rodeo}}^* = O\left(n^{-4/(4+r)}\right)$$
Related work

Recent work that addressed this problem

- Minimum volume set (Scott & Nowak, JMLR 06)
- Nongaussian component analysis; (Blanchard et al. JMLR 06)
- Log-ANOVA model; (Lin & Joen, Statistical Sinica 2006)

Advantages of our approach:

- Rodeo can utilize well-established nonparametric estimators
- A unified framework for different kinds of problems
- Easy to implement and is amenable to theoretical analysis
The main idea

Density rodeo: the main idea

The key intuition: if a dimension is irrelevant, then changing the smoothing parameter of that dimension should only result in a small change in the whole estimator.

- Basically, Rodeo is just a regularization strategy.
- Use a kernel density estimator start with large bandwidths.
- Calculate the gradient of the estimator w.r.t. the bandwidth.
- Sequentially decrease the bandwidths in a greedy way, and try to freeze this decay process by some thresholding strategy to achieve a sparse solution.
Density rodeo: the main idea

Assuming a fixed point $x$ and let $\hat{f}_H(x)$ denote an estimator of $f(x)$ based on smoothing parameter matrix $H = \text{diag}(h_1, \ldots, h_d)$

Let $M(h) = \mathbf{E}(\hat{f}_h(x))$ denote the mean of $\hat{f}_h(x)$, therefore, $f(x) = M(0) = \mathbf{E}(\hat{f}_0(x))$. Assuming $P = \{h(t) : 0 \leq t \leq 1\}$ is a smooth path through the set of smoothing parameters with $h(0) = 0$ and $h(1) = 1$, then

$$f(x) = M(1) - (M(1) - M(0)) = M(1) - \int_0^1 \frac{dM(h(s))}{ds} ds$$

$$= M(1) - \int_0^1 \langle D(s), \dot{h}(s) \rangle ds$$

where $D(h) = \nabla M(h) = \left( \frac{\partial M}{\partial h_1}, \ldots, \frac{\partial M}{\partial h_d} \right)^T$ is the gradient of $M(h)$ and $\dot{h}(s) = \frac{dh(s)}{ds}$ is the derivative of $h(s)$ along the path.
Density rodeo: the main idea

A biased, low variance estimator of $M(1)$ is $\hat{f}_1(x)$. An unbiased estimator of $D(h)$ is

$$Z(h) = \left( \frac{\partial \hat{f}_H(x)}{\partial h_1}, \ldots, \frac{\partial \hat{f}_H(x)}{\partial h_d} \right)^T$$

This naive estimator has poor risk due to the large variance of $Z(h)$ for small bandwidth. However, the sparsity assumption on $f$ suggests that there should be paths for which $D(h)$ is also sparse. Along such a path, $Z(h)$ could be replaced with an estimator $\hat{D}(h)$ that makes use of the sparsity assumption. Then, the estimate of $f(x)$ becomes

$$\tilde{f}_H(x) = \hat{f}_1(x) - \int_0^1 \langle \hat{D}(s), \dot{h}(s) \rangle ds$$
Kernel density estimator

Let $\hat{f}_H(x)$ represents the kernel density estimator of $f(x)$ with a bandwidth matrix $H$. Assuming that $K$ is a standard symmetric kernel, s.t.

$$\int K(u)du = 1, \quad \int uK(u)du = 0_d$$

while $K_H(\cdot) = \frac{1}{\det(H)}K(H^{-1}\cdot)$ represents the kernel with bandwidth matrix $H = diag(h_1, \ldots, h_d)$.

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

$$= \frac{1}{n} \sum_{i=i}^{n} \prod_{j=1}^{d} \frac{1}{h_j} K \left( \frac{x_j - X_{ij}}{h_j} \right)$$

Here, we assume that $K$ is a product kernel.
The main idea

The local rodeo algorithm for the kernel density estimator

The rodeo statistics for the kernel density estimator

The density estimation Rodeo is based on the statistics

\[ Z_j = \frac{\partial \hat{f}_H(x)}{\partial h_j} \]

\[ = \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{K}\left(\frac{x_j - X_{ij}}{h_j}\right)}{K\left(\frac{x_j - X_{ij}}{h_j}\right)} \prod_{k=1}^{d} K\left(\frac{x_k - X_{ik}}{h_k}\right) \equiv \frac{1}{n} \sum_{i=1}^{n} Z_{ji} \]

For the conditional variance term

\[ s_j^2 = \text{Var}(Z_j | X_1, ..., X_n) \]

\[ = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} Z_{ji} | X_1, ..., X_n \right) = \frac{1}{n} \text{Var}(Z_{j1} | X_1, ..., X_n) \]

Here, we used the sample variance of the \( Z_{ji} \) to estimate \( \text{Var}(Z_{j1}) \).
Density rodeo algorithms

**Density Rodeo: Hard thresholding version**

1. *Select* parameter $0 < \beta < 1$ and initial bandwidth $h_0$, where
   
   
   $h_0 = c_0 / \log \log n$

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

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

3. *While* $A$ is nonempty, do for each $j \in A$
   
   (a) Compute the derivative and variance: $Z_j$ and $s_j$.
   
   (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 $A$.

4. *Output* bandwidths $h^*$ and the estimator $\tilde{f}(x) = \hat{f}_{H^*}(x)$
The purpose of the analysis

The analysis is trying to characterize the asymptotic aspects of

- the selected bandwidths
- the asymptotic running time (efficiency)
- the convergence rate of the risk (accuracy)

To make the theoretical results more realistic, a key aspect of our analysis is that we allow the dimension $d$ to increase with the sample size $n$! For this, we need to make some assumptions about the unknown density function to take the increasing dimension into account.

Liu, Lafferty, Wasserman

Sparse Nonparametric Density Estimation
Assumptions

- (A1) Kernel assumption:
  \[ \int uu^T \mathcal{K}(u) du = v_2 I_d \text{ and } v_2 < \infty \text{ and } \int \mathcal{K}^2(u) du = R(\mathcal{K}) < \infty \]

- (A2) Dimension assumption:
  \[ d \log d = O(\log n) \]

- (A3) Initial bandwidth assumption:
  \[ h_j^{(0)} = c_0 / \log \log n \text{ for } (j = 1, \ldots, d) \]
  Combing with A2, this implies that \( \lim_{n \to \infty} n \prod_{j=1}^d h_j^{(0)} = \infty \)

- (A4) Sparsity assumption:
  \[ f(x) \propto g(x_1, \ldots, x_r) h(x) \text{ where } h_{jj}(x) = 0, \text{ and satisfies } r = O(1) \]

- (A5) Hessian assumption:
  \[ \int tr(\mathcal{H}_R^T(u)\mathcal{H}_R(u)) du < \infty, \text{ and } f_{jj}(x) \neq 0 \text{ for } j = 1, 2, \ldots, r \]
  where \( \mathcal{H}_R(u) \) represents the Hessian for the relevant dimensions
Derivatives of both relevant and irrelevant dimensions

**Key Lemma:** Under assumptions A1 − A5, suppose that \( x \) is interior to the support of \( f \). Suppose that \( K \) is a product kernel with bandwidth matrix \( H^{(s)} = \text{diag}(h_1^{(s)}, \ldots, h_d^{(s)}) \). Then

\[
\mu^{(s)}_j = \frac{\partial}{\partial h_j^{(s)}} \mathbb{E}[\hat{f}_{H^{(s)}}(x) - f(x)] = o_P(h_j^{(s)}) \text{ for all } j \in R^c
\]

For \( j \in R \) we have

\[
\mu^{(s)}_j = \frac{\partial}{\partial h_j^{(s)}} \mathbb{E}[\hat{f}_{H^{(s)}}(x) - f(x)] = h_j^{(s)} v_2 f_{jj}(x) + o_P(h_j^{(s)}).
\]

Thus, for any integer \( s > 0 \), \( h_s = h_0 \beta^s \), each \( j > r \) satisfies

\[
\mu^{(s)}_j = o_P(h_j^{(s)}) = o_P(h_j^{(0)}).
\]
Main Theorem: Suppose that \( r = O(1) \) and (A1)-(A5) hold. In addition, suppose that \( A_{min} = \min_{j \leq r} |f_{jj}(x)| = \tilde{\Omega}(1) \) and \( A_{max} = \max_{j \leq r} |f_{jj}(x)| = \tilde{O}(1) \). Then, for every \( \epsilon > 0 \), the number of iterations \( T_n \) until the Rodeo stops satisfies

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

where \( a_n = \tilde{\Omega}(1) \) and \( b_n = \tilde{O}(1) \). Moreover, the algorithm outputs bandwidths \( h^* \) that satisfy

\[
P \left( h^*_j = h_j^{(0)} \text{ for all } j > r \right) \longrightarrow 1
\]

Also, we have

\[
P \left( h_j^{(0)}(nb_n)^{-1/(4+r)-\epsilon} \leq h^*_j \leq h_j^{(0)}(na_n)^{-1/(4+r)+\epsilon} \text{ for all } j \leq r \right) \longrightarrow 1
\]

assuming that \( h_j^{(0)} \) is defined as in A3.
Theorem 2 Under the same condition of the main theorem, the risk $R_{h^*}$ of the density Rodeo estimator satisfies

$$R_{h^*} = \tilde{O}_P \left( n^{-4/(4+r)} + \epsilon \right)$$

for every $\epsilon > 0$.

We write $Y_n = \tilde{O}_P(a_n)$ to mean that $Y_n = O(b_n a_n)$ where $b_n$ is logarithmic in $n$. As noted earlier, we write $a_n = \Omega(b_n)$ if

$$\lim \inf_n \left| \frac{a_n}{b_n} \right| > 0;$$

similarly $a_n = \tilde{\Omega}(b_n)$ if $a_n = \Omega(b_n c_n)$ where $c_n$ is logarithmic in $n$. 

---

**Background**

Methodology and Algorithms

**Asymptotic Properties**

Extension and Variations

Experimental Results

Liu, Lafferty, Wasserman

Sparse Nonparametric Density Estimation
The density rodeo algorithm could be extended in many ways

- the soft-thresholding version
- the global version
- the reverse version
- the bootstrapping version
- the rodeo algorithm for local linear density estimator
- the greedier version
- using other distributions as irrelevant
Global density rodeo

The idea is by averaging the test statistics for multiple evaluation points $x_1, \ldots, x_k$ sampled from the empirical distribution.

To avoid the cancellation problem, the statistic is squared, let $Z_j(x_i)$ represents the derivative for the $i$-th evaluation point with respect to the bandwidth $h_j$. We define the test statistic

$$T_j = \frac{1}{m} \sum_{k=1}^{m} Z_j^2(x_k), \quad j = 1, \ldots, d$$

while

$$s_j = \sqrt{\text{Var}(T_j)} = \frac{1}{m} \sqrt{2\text{tr}(C_j^2) + 4\hat{\mu}_j^T C_j \hat{\mu}_j}$$

where $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} Z_j(x_i)$. The threshold is

$$\lambda_j = s_j^2 + 2s_j \sqrt{\log(nc_n)}$$
The bootstrapping version

Instead of derive the expression explicitly, bootstrapping can be used to evaluate the variance of $Z_j$.

**Bootstrapping Method to calculate the $s_j^2$**

1. Draw a sample $X_1^*, ..., X_n^*$ of size $n$, with replacement:
   - Loop for $i = 1, ..., B$,
     - Compute the estimate $Z^*_{ji}$ from data $X_1^*, ..., X_n^*$

2. Compute the bootstrapped variance
   \[ s_j^2 = \frac{1}{B} \sum_{b=1}^{B} (Z^*_{ji} - \bar{Z}_j)^2. \]
   where \[ \bar{Z}_j = \frac{1}{B} \sum_{b=1}^{B} \hat{Z}_j \]

3. Output the resulted $s_j^2$. 

The reverse and greedier version

- **Reverse density Rodeo:** Instead of using a sequence of decreasing bandwidths. On the contrary, we could begin from a very small bandwidth, and use a sequence of increasing bandwidths to estimate the optimal value.

- **Greedier density rodeo:** Instead of decaying all the bandwidths, only the bandwidths associated the largest $Z_j/\lambda_j$ quantities is decayed.

- **Hybrid density Rodeo:** Different variations could be combined arbitrarily.
Using other distributions as irrelevant dimensions

We can use a general parametric distribution $h(x)$ as irrelevant dimensions. The key trick is that a new semi-parametric density estimator will be used

$$
\hat{f}_H(x) = \frac{\hat{h}(x) \sum_{i=1}^{n} K_H(X_i - x)}{n \int K_H(u - x) \hat{h}(u) du}
$$

where $\hat{h}(x)$ is a parametric density estimator at point $x$. The motivation of this estimator comes from local likelihood method (Loader 1996).

We see that, for one dimensional case, starting from a large bandwidth, if the true function is $h(x)$, the algorithm will tend to freeze the bandwidth decaying process immediately.
Experimental design

- The density rodeo algorithms were applied on both synthetic and real data, including one-dimensional, two-dimensional, high-dimensional and very high-dimensional examples
- to measure the distance between the estimated density function with the true density function. The Hellinger distance are used

\[
D(\hat{f} \| f) = \int \left( \sqrt{\hat{f}(x)} - \sqrt{f(x)} \right)^2 dx = 2 - 2 \int f(x) \sqrt{\frac{\hat{f}(x)}{f(x)}} dx
\]

Assuming that we have altogether \(m\) evaluation points, then the hellinger distance could be numerically calculated by the Monte Carlo integration

\[
D(\hat{f} \| f) \approx 2 - \frac{2}{m} \sum_{i=1}^{m} \sqrt{\frac{\hat{f}_H(X_i)}{f(X_i)}}
\]
Strongly skewed density: We simulated 200 samples from the skewed distribution. The boxplot of the Hellinger distance is produced based on 100 simulations.

This density is chosen to resemble to lognormal distribution, it distributes as

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

The true density is plotted.
Strongly skewed density: result

Unbiased Cross Validation

Local kernel density Rodeo

Global kernel density Rodeo

Bandwidth estimated

Liu, Lafferty, Wasserman
Sparse Nonparametric Density Estimation
Combined Beta distribution with uniform distribution as irrelevant dimension. We simulate a 2-dimensional dataset with $n = 500$ points.

The two dimensions are independently generated as

$$X_1 \sim \frac{2}{3} \text{Beta}(1, 2) + \frac{1}{3} \text{Beta}(10, 10)$$
$$X_2 \sim \text{Uniform}(0, 1)$$

The true density for the relevant dimension is
Combined Beta and Uniform: result

The first plot is the rodeo result, the second plot is the result fitted by the built-in function KDE2d (MASS package in R)
Combined Beta and Uniform: marginal distribution

Numerically integrated marginal distributions based on the perspective plots of the two estimators (not normalized).
Two-dimensional example: geyser data

**Geyser Data:** A version of the eruptions data from the “Old Faithful” geyser in Yellowstone National Park, Wyoming. (Azzalini and Bowman 1990) and is of continuous measurement from August 1 to August 15, 1985.

- There are altogether \( n = 299 \) samples and two variables.
- “Duration” = the numeric eruption time in minutes.
- “waiting” = the waiting time to next eruption.
Geyser data: result

The first plot is the rodeo result, the second plot is the result fitted by the built-in function KDE2d (MASS package in R)
The first plot is the contour plot fitted by the built-in function KDE2d (MASS package in R), the second one is fitted by the rodeo algorithm.
High dimensional example

**30-dimensional example:** We generate 30-dimensional synthetic dataset with \( r = 5 \) relevant dimensions \((n = 100, \text{ with } 30 \text{ trials})\). The relevant dimensions are generated as

\[ X_i \sim \mathcal{N}(0.5, (0.02i)^2), \quad \text{for} \quad i = 1, \ldots, 5 \]

while the irrelevant dimensions are generated as

\[ X_i \sim \text{Uniform}(0, 1), \quad \text{for} \quad i = 6, \ldots, 30 \]

The test point is \( x = (\frac{1}{2}, \ldots, \frac{1}{2}) \).
30-dimensional example: result

The Rodeo path for the 30-dimensional synthetic dataset and the boxplot of the selected bandwidths for 30 trials.
Very high dimensional example: image processing

The algorithm was run on 2200 grayscale images of 1s and 2s, each with $256 = 16 \times 16$ pixels with some unknown background noise; thus this is a 256-dimensional density estimation problem. A test point and the output bandwidth plot are shown here.
Image processing example: evolution plot

The output bandwidth plots sampled at the Rodeo step 10, 20, 30, 40, 50, 60, 70, 80, 90, 100. Which visualizes the evolution of the bandwidths and could be viewed as a dynamic process for feature selection — the earlier a dimension’s bandwidth decays, the more informative it is.
A example using Gaussian as irrelevant

Using Gaussian as irrelevant dimensions: We apply the semiparametric rodeo algorithm on both 15-dimensional and 20-dimensional synthetic datasets with \( r = 5 \) relevant dimensions \((n = 1000)\). When using gaussian distributions as irrelevant dimensions, the relevant dimensions are generated as

\[
X_i \sim \text{Uniform}(0, 1), \quad \text{for} \quad i = 1, \ldots, 5
\]

while the irrelevant dimensions are generated as

\[
X_i \sim \mathcal{N}(0.5, (0.05i)^2), \quad \text{for} \quad i = 6, \ldots, d
\]

The test point is \( x = (\frac{1}{2}, \ldots, \frac{1}{2}) \).
Using Gaussian as irrelevant: result

The Rodeo path for the 15-dimensional synthetic data (Left) and for the 20-dimensional data (Right)
Using Gaussian as irrelevant: one-dimensional case (I)

1000 one-dimensional data points with $x_i \sim \text{Uniform}(0, 1)$.
Using Gaussian as irrelevant: one-dimensional case (II)

1000 one-dimensional data points with $x_i \sim \mathcal{N}(0, 1)$. 

Liu, Lafferty, Wasserman

Sparse Nonparametric Density Estimation
Summary

- This work adapts the general rodeo framework to solve density estimation problems.
- The sparsity assumption is crucial to guarantee the success of the density rodeo algorithm.
- The density rodeo algorithm is efficient on high-dimensional problems both theoretically and empirically.
- The rodeo framework can utilize current available density estimators, the implementation is simple.
- Future work: develop the rodeo algorithms for the other kinds of problems, e.g. classification.