ECE598PV: Representation Learning: Algorithms and Models Fall 2017

Lecture 1: Non-parametric Mixture Models

Lecturer: Professor Pramod Viswanath *Scribe: Ali Yekkehkhany, Sept. 12, 2017*

1.1 Non-parametric Mixture Models [Was06, Loa06]

Consider data points $\{x_1, x_2, \cdots, x_N\}$ that are independent realizations of the same distribution $f(x)$, which is unknown to us. We are looking for the underlying probability law that explains the data, i.e. we are looking for an estimate of the probability density distribution $f(x)$ based on the observed data points. To this end there are two general approaches: 1- Non-parametric that is the topic of this lecture, 2- Parametric that is discussed in the next lecture. In parametric method, some assumptions are made about the probability density function, for example it is assumed that data points are drawn independently at random from same Gaussian distribution and the goal is to find the mean and variance of the Gaussian distribution. On the other hand, the goal of **non-parametric density estimation** is to have as few assumptions as possible about the underlying probability density function to estimate $f(x)$. Before presenting the common approaches for pdf estimation, the risk of an estimator is formulated in the next section.

1.1.1 Risk of the Estimator

Let $\widehat{f}_N(x)$ be the estimate of the true pdf function $f(x)$ based on the observed data $\{x_1, x_2, \cdots, x_N\}$. The following metric is broadly used to evaluate the performance of the estimator, which is the integrated mean squared error loss or the risk:

$$L = \int \left(\widehat{f}_N(x) - f(x) \right)^2 dx. \tag{1.1}$$

As we will see later, the estimator of $f(x)$, $\widehat{f}_N(x)$, depends on a smoothing parameter h that is chosen in a way to minimize the risk, then equation (1.1) can be written as

$$L(h) = \int \widehat{f}_N^2(x)\ dx - 2 \int \widehat{f}_N(x) f(x)\ dx + \int f^2(x)\ dx. \tag{1.2}$$

Since the last term of (1.2) does not depend on the smoothing parameter h, instead of minimizing the expected risk with respect to h, one can minimize the expectation of $J(h)$ that is defined as

$$J(h) = \int \widehat{f}_N^2(x)\ dx - 2 \int \widehat{f}_N(x) f(x)\ dx. \tag{1.3}$$

Unless $\mathbb{E}[J(h)]$ differs from the true risk by $\int f^2(x)\ dx$, it is also referred to as risk. **Leave-one-out cross-validation** is a method to estimate the risk which is define as

$$\widehat{J}(h) = \int \widehat{f}_N^2(x)\ dx - \frac{2}{N} \sum_{i=1}^{N} \widehat{f}_{(-i)}(x_i),$$

where $\widehat{f}_{(-i)}$ is the pdf estimator based on data points $\{x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_N\}$, where x_i is excluded. $\widehat{J}(h)$ is called **cross-validation estimator of risk** (also called the cross-validation score or simply the estimated risk). The next two sections present two popular ways of estimating the pdf.

1.1.2 Histogram Estimator

The simplest non-parametric pdf estimator is perhaps the histogram estimator. Without loss of generality, suppose the support of the pdf is $[0,1]$. Define bins of length $h = \frac{1}{m}$ as follows, where m is an integer number:

$$B_1 = \left[0, \frac{1}{m}\right), B_2 = \left[\frac{1}{m}, \frac{2}{m}\right), \cdots, B_m = \left[\frac{m-1}{m}, 1\right].$$

Define the number of observed data points in the j-th bin as Y_j and let $\widehat{p}_j = \frac{Y_j}{N}$. Then the **histogram estimator** is given by

$$\widehat{f}_N(x) = \sum_{j=1}^{m} \frac{\widehat{p}_j}{h} I(x \in B_j) \tag{1.4}$$

The intuition behind choosing the pdf function defined in (1.4) is the following. Let $p_j = \int_{B_j} f(u)\, du$, and let h to be a small number, then for $x \in B_j$ we have the following:

$$\mathbb{E}\left[\widehat{f}_N(x)\right] = \frac{\mathbb{E}\left[\widehat{p}_j\right]}{h} = \frac{p_j}{h} = \frac{\int_{B_j} f(u)\, du}{h} \approx \frac{f(x) \cdot h}{h} = f(x).$$

The important thing in the histogram estimator method is to choose the right number of bins in order to prevent under-smoothing or over-smoothing. Figure 1.1 is driven from $N = 1266$ astronomical data points. The bottom right plot shows the estimated risk versus the number of bins, so the parameter $h = \frac{1}{m}$ can be chosen to minimize the risk and to prevent under/over-smoothing. The mean and the variance of the histogram estimator for a fixed $x \in B_j$ and a fixed integer m are

$$\mathbb{E}\left[\widehat{f}_N(x)\right] = \frac{p_j}{h}, \qquad Var\left[\widehat{f}_N(x)\right] = \frac{p_j(1-p_j)}{Nh^2}.$$

The following theorem is followed, where it gives a nice intuition about the convergence rate of risk to zero and the optimum h that results in the fastest convergence of risk.

Theorem 1.1. *Let the probability density function $f(x)$ be absolutely continuous and $\int (f(u))^2\ du < \infty$, then the risk associated to the histogram estimator $\widehat{f}_N(x)$ is*

$$R\left(\widehat{f}_N, f\right) = \frac{h^2}{12} \int f^2(u)\ du + \frac{1}{Nh} + o\left(h^2\right) + o\left(\frac{1}{N}\right) \tag{1.5}$$

The optimum value for h, h^*, that minimizes the risk given in (1.5) is

$$h^* = \frac{1}{N^{\frac{1}{3}}} \left(\frac{6}{\int f^2(u)\ du}\right)^{\frac{1}{3}} \tag{1.6}$$

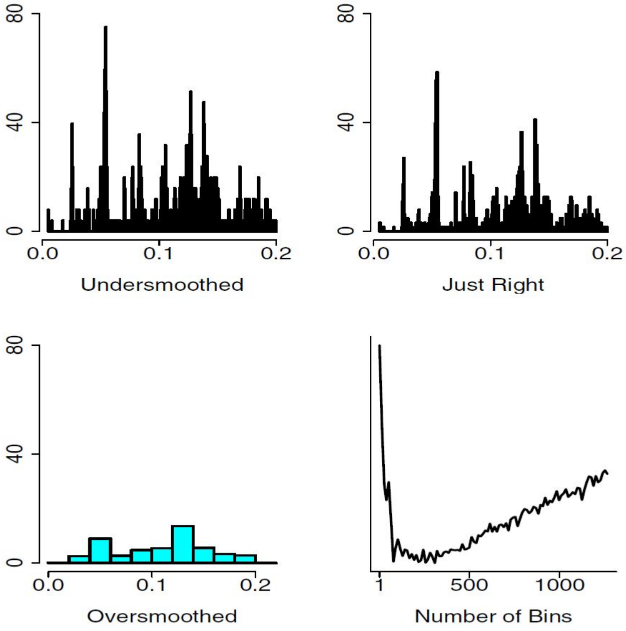

Figure 1.1: Histogram estimator for three choices of h, when it results in under-smoothing, over-smoothing, and when it is just right. The bottom right plot shows the estimated risk for different number of bins.

The risk of the histogram estimator with the choice of bin-width h in (1.6) is

$$R\left(\widehat{f}_N, f\right) \sim \frac{C}{N^{\frac{2}{3}}}, \tag{1.7}$$

where $C = \left(\frac{3}{4}\right)^{\frac{2}{3}} \left(\int f^2(u)\, du\right)^{\frac{1}{3}}$.

Equation (1.7) in theorem 1.1 gives the convergence rate of the risk as $N^{-\frac{2}{3}}$. It will be shown later that the convergence rate for kernel estimators is $N^{-\frac{4}{5}}$ and in some sense a faster rate is not feasible. Note that the optimal h given by (1.6) cannot be computed directly since $f(x)$ is not known. In practice, we use cross-validation. The cross-validation score, $\widehat{J}(h)$, is given as

$$\widehat{J}(h) = \frac{2}{h(N-1)} - \frac{N+1}{h(N-1)} \sum_{j=1}^{m} \widehat{p}_j^2.$$

The following theorem gives a confidence set for $f(x)$, but before that we define the histogramized version of $f(x)$ as

$$\overline{f}_N(x) = \mathbb{E}\left[\widehat{f}_N(x)\right] = \sum_{j=1}^{m} \frac{p_j}{h} I(x \in B_j).$$

Theorem 1.2. *The number of bins $m = m(N)$ in the histogram $\widehat{f}_N$ satisfies $m(N) \to \infty$ and $\frac{m(N)\log(N)}{N} \to 0$ as $N \to \infty$. Then*

$$P\left(l_N(x) \leq \overline{f}_N(x) \leq u_N(x) \text{ for all } x\right) \geq 1 - \alpha,$$

where

$$l_N(x) = \left(\max\left\{\sqrt{\widehat{f}_N(x)} - c, 0\right\}\right)^2, \qquad u_N(x) = \left(\sqrt{\widehat{f}_N(x)} + c\right)^2,$$

where $c = \frac{z_{\alpha/(2m)}}{2}\sqrt{\frac{m}{N}}$. *In other words,* $(l_N(x), u_N(x))$ *is an approximate* $1-\alpha$ *confidence band for* $\overline{f}_N(x)$.

1.1.3 Kernel Density Estimator (KDE)

Histograms are not smooth and their convergence rate is slower than kernel method. Consider function K that satisfies the following:

$$\int K(x)\ dx = 1, \qquad \int xK(x)\ dx = 0, \qquad \sigma_K^2 = \int x^2 K(x)\ dx > 0.$$

The box, tricube, Epanechnikov, and Gaussian kernel functions are examples of widely used ones. Given a kernel function K and **bandwidth** $h > 0$, the **kernel bandwidth estimator** is defined as

$$\widehat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h} K(\frac{x - x_i}{h}). \tag{1.8}$$

The interpretation of (1.8) is to put a smoothed function with weight $\frac{1}{N}$ over each observed data point x_i. The choice of the kernel function is not as crucial as the choice of the bandwidth h, and similar to the histogram estimator, wrong choices of h can lead to under/over-smoothing.

Theorem 1.3. *Let* $f(x)$ *be continuous at* x *and* $h_N \to 0$ *and* $Nh_N \to \infty$ *as* $N \to \infty$, *then*

$$\widehat{f}_N(x) \xrightarrow{P} f(x).$$

Theorem 1.4. *The risk at point* x *is defined as* $R_x = \mathbb{E}\left[\left(f(x) - \widehat{f}(x)\right)^2\right]$, *and the integrated risk is* $R = \int R_x\ dx$. *Assuming that* f'' *is absolutely continuous and* $\int (f'''(x))^2\, dx < \infty$, *then*

$$R_x = \frac{1}{4}\sigma_K^4 h_N^4 \left(f''(x)\right)^2 + \frac{f(x)\int K^2(x)\ dx}{Nh_N} + O\left(\frac{1}{N}\right) + O\left(h_N^6\right)$$

and

$$R = \frac{1}{4}\sigma_K^4 h_N^4 \int \left(f''(x)\right)^2 dx + \frac{\int K^2(x)\ dx}{Nh_N} + O\left(\frac{1}{N}\right) + O\left(h_N^6\right) \tag{1.9}$$

By setting the derivative of (1.9) with respect to h_N to zero to find the optimum value of h_N, we have

$$h_N^* = \left(\frac{c_1}{\sigma_K^4 A(f) N}\right)^{\frac{1}{5}}, \tag{1.10}$$

where $c_1 = \int K^2(x)\ dx$ and $A(f) = \int (f''(x))^2\, dx$. Hence, the best bandwidth decreases at rate $N^{-\frac{1}{5}}$. Plugging h_N^* back into (1.9), we see that $R = O(N^{-\frac{4}{5}})$. The following theorem shows that for the conditions on f in theorem 1.4 $\frac{1}{N^{\frac{4}{5}}}$ is the fastest rate that can be achieved.

Theorem 1.5. *Consider the set of all probability density functions denoted by $\mathcal{F}$ and denote the m-th derivative of f by $f^{(m)}$ and define*

$$\mathcal{F}_m(c) = \left\{ f \in \mathcal{F} : \int \left| f^{(m)}(x) \right|^2 dx \leq c^2 \right\}.$$

Then for any kernel estimator $\widehat{f}_N$

$$\sup_{f \in \mathcal{F}_m(c)} \mathbb{E}_f \left[\int \left(\widehat{f}_N(x) - f(x) \right)^2 dx \right] \geq b \left(\frac{1}{N} \right)^{\frac{2m}{2m+1}}, \tag{1.11}$$

where $b > 0$ only depends on m and c.

Having $m = 2$ in theorem 1.5, equation (1.11) shows that a faster convergence rate than $N^{-\frac{4}{5}}$ is impossible for kernel estimators. Similar to histogram estimators, the choice of h_N^* in (1.10) cannot be computed practically since the true pdf f is unknown. Other than the cross-validation method which will be discussed later, **Normal reference rule** can also be used which is illustrated in the following.

Normal reference rule: Under the idealized assumption that f is Normal, h_N^* in (1.10) is

$$h_N^* = \frac{1.06\widehat{\sigma}}{N^{\frac{1}{5}}},$$

where

$$\widehat{\sigma} = \min \left\{ s, \frac{Q}{1.34} \right\},$$

and s is the sample standard deviation and Q is the interquartile range (the 75-th percentile minus the 25-th percentile, and Q is divided by 1.34 to have a consistent estimate of the standard deviation of the Normal distribution). In practice, the Normal reference rule is also used for smooth densities other than the Normal distributions.

The following theorem gives the expression for the cross-validation that can be used for finding h.

Theorem 1.6. *For $\forall h > 0$,*

$$\mathbb{E}\left[\widehat{J}(h)\right] = \mathbb{E}\left[J(h)\right]$$

and

$$\widehat{J}(h) = \frac{1}{hN^2} \sum_i \sum_j K^* \left(\frac{x_i - x_j}{h} \right) + \frac{2}{Nh} K(0) + O\left(\frac{1}{N^2} \right),$$

where $K^*(x) = K^{(2)}(x) - 2K(x)$ and $K^{(2)}(z) = \int K(z-y)K(y)\, dy$.

Another approach to choose bandwidth h is **plug-in bandwidth**. The only unknown parameter in the optimum bandwidth h_N^* in equation (1.10) is $A(f) = \int (f''(x))^2 dx$. In the plug-in bandwidth approach, you first estimate f'' by $\widehat{f}''$, and use the estimate to find h_N^*, but in order to find the second derivative estimate, you need to put more assumptions on the pdf function. However, even with the more strong assumptions on f, the kernel estimator is not appropriate, where this issue is investigate in [Loa99]. There are also methods for correcting the plug-in methods [JSH99]. A generalization of the kernel method is to use different bandwidths $h(x)$ for different points x or to use different bandwidths $h(x_i)$ for different observed data points x_i, which is referred to as **adaptive kernel**. This way, it is easier to estimate with more flexibility and adapt to different regions with different smoothness [Was06]. However, the job of finding many bandwidths instead of one makes the adaptive kernel approach harder.

1.1.4 Local Polynomial Estimator

Local polynomial approach is effective to reduce boundary bias. In the following, the local likelihood density estimation (LLDE) method developed by Loader [Loa06] and Hjort [HJ96] is presented. Note that the log-likelihood is $\mathcal{L}(f) = \sum_{i=1}^{N} \log f(x_i)$, but since $\int f(u)\ du = 1$, the log-likelihood can also be written as

$$\mathcal{L}(f) = \sum_{i=1}^{N} \log f(x_i) - N\left(\int f(u)\ du - 1\right). \tag{1.12}$$

For a target value x and a given kernel function K and bandwidth h, the local version of (1.12) is

$$\mathcal{L}_x(f) = \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right) \log f(x_i) - N \int K\left(\frac{u-x}{h}\right) f(u)\ du. \tag{1.13}$$

The tailor series expansion of the term $\log f(u)$ in (1.13) is

$$\log f(u) \approx P_x(a,u) = a_0 + a_1(x-u) + \cdots + a_p \frac{(x-u)^p}{p!}. \tag{1.14}$$

By plugging (1.14) into (1.13), the local polynomial log-likelihood is

$$\mathcal{L}_x(a) = \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right) P_x(a, x_i) - N \int K\left(\frac{u-x}{h}\right) e^{P_x(a,u)}\ du.$$

Let $\widehat{a} = (\widehat{a}_0, \widehat{a}_1, \cdots, \widehat{a}_p)^T = \arg\max_a \mathcal{L}_x(a)$, then the local likelihood density estimate is

$$\widehat{f}_N(x) = e^{P_x(\widehat{a},x)},$$

where it reduces to the kernel density estimation when $p = 0$.

1.1.5 Multivariate Problem

Consider the data points are d-dimensional, i.e. $x_i = (x_{i1}, \cdots, x_{id})$ for $1 \le i \le N$. Due to curse of dimensionality though the convergence rate of the estimator decreases quickly as d increases. The product kernel extension for pdf estimator is

$$\widehat{f}_N(x) = \frac{1}{Nh_1 \cdots h_d} \sum_{i=1}^{N} \left\{ \prod_{j=1}^{d} K\left(\frac{x_j - x_{ij}}{h_j}\right) \right\}$$

and the risk associated to this estimator is

$$R \approx \frac{1}{4}\sigma_K^4 \left[\sum_{j=1}^{d} h_j^4 \int f_{jj}^2(x)\ dx + \sum_{j \neq k} h_j^2 h_k^2 \int f_{jj} f_{kk}\ dx \right] + \frac{\left(\int K^2(x)\ dx\right)^d}{Nh_1 \cdots h_d},$$

where the second derivative of f is denoted by f_{jj}. The optimal bandwidth and the corresponding risk satisfy $h_i = O\left(N^{-\frac{1}{4+d}}\right)$ and $R = O\left(N^{-\frac{4}{4+d}}\right)$, respectively. As you see, the risk increases quickly in high dimensions.

1.1.6 Entropy and Mutual Information Estimator

Since calculating the entropy and mutual information are based on pdf functions, the proposed approaches for pdf estimation can directly be used for entropy and mutual information estimation. Gao et al. [GOV16] combined the geometric nearest neighbor (NN) based approach and the kernel based approach for estimation of entropy and mutual information. They use the k-NN distances to choose a local bandwidth, where k is finite and independent of the sample size. It is known that for $p = 0$ the LLDE reduces to KDE

$$\widehat{f}_N(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \Big/ \int K\left(\frac{u - x}{h}\right) du.$$

By choosing the kernel function to be the step function, $K(x) = I(\|x\| \leq 1)$ (Gao et al. have not assumed the kernel function to integrate to one), and the local and data-dependent bandwidth to be $h(x) = \rho_{k,x}$ as the k-NN distance from x, the above estimator converts to the k-NN density estimator as

$$\widehat{f}_N(x) = \frac{\frac{1}{N}\sum_{i=1}^{N} I\left(\|x_i - x\| \leq \rho_{k,x}\right)}{Vol\left(u \in \mathbb{R}^d : \|u - x\| \leq \rho_{k,x}\right)} = \frac{k}{NC_d\rho_{k,x}^d},$$

where $C_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}$. On the other hand, the maximizer of (1.13) when using the Gaussian kernel $K(x) = e^{-\frac{\|x\|^2}{2}}$ for the cases of $p \in \{1, 2\}$ has the following closed form. For $p = 1$, $x \in \mathbb{R}^d$, and $h \in \mathbb{R}$,

$$\widehat{f}_N(x) = \frac{S_0}{N(2\pi)^{\frac{d}{2}}h^d} e^{-\frac{1}{2S_0^2}\|S_1\|^2},$$

where $S_0 \in \mathbb{R}$ and $S_1 \in \mathbb{R}^d$ are given as

$$S_0 = \sum_{j=1}^{N} e^{-\frac{\|x_j - x\|^2}{2h^2}}, \qquad S_1 = \sum_{j=1}^{N} \frac{1}{h}\left(x_j - x\right) e^{-\frac{\|x_j - x\|^2}{2h^2}}.$$

For $p = 2$ and S_0 and S_1 that are defined in the above,

$$\widehat{f}_N(x) = \frac{S_0}{N(2\pi)^{\frac{d}{2}}h^d|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2S_0^2}S_1^T\Sigma^{-1}S_1},$$

where $S_2 \in \mathbb{R}^{d\times d}$ and $\Sigma \in \mathbb{R}^{d\times d}$ are given as

$$S_2 = \sum_{j=1}^{N} \frac{1}{h^2}\left(x_j - x\right)\left(x_j - x\right)^T e^{-\frac{\|x_j - x\|^2}{2h^2}}, \qquad \Sigma = \frac{S_0S_2 - S_1S_1^T}{S_0^2}.$$

For entropy estimation, consider $\widehat{H}(X) = -\frac{1}{N}\sum_{i=1}^{N}\log\widehat{f}_N(x_i)$, where LLDE with local and adaptive choice of bandwidth is used. Specifically, the bandwidth for data point x_i, $h(x_i)$, is chosen as its distance to its k-th nearest neighor $\rho_{k,i}$. Then the k-local nearest neighbor (k-LNN) entropy estimator is given as

$$\widehat{H}_{kLLN}^{(N)}(X) = -\frac{1}{N}\sum_{i=1}^{N}\left\{\log\frac{S_{0,i}}{N(2\pi)^{\frac{d}{2}}\rho_{k,i}^d|\Sigma|^{\frac{1}{2}}} - \frac{1}{2S_{0,i}^2}S_{1,i}^T\Sigma_i^{-1}S_{1,i}\right\} - B_{k,d}, \tag{1.15}$$

where $B_{k,d}$ is a constant that depends on k and d, and defining $\mathcal{T}_{i,m} = \{j \in [N] : j \neq i \text{ and } \|x_i - x_j\| \leq \rho_{m,i}\}$, the parameters $S_{0,i}, S_{1,i}, S_{2,i}$, and Σ_i are given as

$$S_{0,i} = \sum_{j \in \mathcal{T}_{i,m}} e^{-\frac{\|x_j - x_i\|^2}{2\rho_{k,i}^2}}, \qquad S_{1,i} = \sum_{j \in \mathcal{T}_{i,m}} \frac{1}{\rho_{k,i}} (x_j - x_i)\, e^{-\frac{\|x_j - x_i\|^2}{2\rho_{k,i}^2}},$$

$$S_{2,i} = \sum_{j \in \mathcal{T}_{i,m}} \frac{1}{\rho_{k,i}^2} (x_j - x_i)(x_j - x_i)^T e^{-\frac{\|x_j - x_i\|^2}{2\rho_{k,i}^2}}, \qquad \Sigma_i = \frac{S_{0,i} S_{2,i} - S_{1,i} S_{1,i}^T}{S_{0,i}^2},$$

where $m = O\left(N^{\frac{1}{2d} - \epsilon}\right)$ for an arbitrary small ϵ. Gao et al. proved that for $k \geq 3$ and for twice continuously differentiable pdf $f(x)$ we have

$$\lim_{N \to \infty} \mathbb{E}\left[\widehat{H}_{kLNN}^{(N)}(X)\right] = H(X).$$

Furthermore, if $\mathbb{E}\left[(\log f(X))^2\right] < \infty$, then $Var\left[\widehat{H}_{kLNN}^{(N)}(X)\right] = O\left(\frac{(\log N)^2}{N}\right)$.

In order to find the mutual information between X and Y, one way is to use $\widehat{H}_{KL}$ to compute $\widehat{I}_{3KL} = \widehat{H}_{KL}(X) + \widehat{H}_{KL}(Y) - \widehat{H}_{KL}(X, Y)$. Kraskov et al. [KSG04] proposed $\widehat{I}_{KSG}(X;Y)$, where the joint entropy is computed in the usual way, but the marginal entropy is estimated by choosing the bandwidth $h(x_i) = \rho_{k,i}(X, Y)$ as the k-the nearest neighbor distance from (x_i, y_i), instead of using the k-NN distance from x_i. Consider $\widehat{I}_{3LNN}(X;Y) = \widehat{H}_{kLNN}(X) + \widehat{H}_{kLNN}(Y) - \widehat{H}_{kLNN}(X, Y)$. Inspired by [KSG04], Gao et al. defined $\widehat{I}_{LNN-KSG}(X;Y)$, where the LNN entropy estimator in (1.15) is used for the joint (X, Y) and the local bandwidth $h(x_i) = \rho_{k,i}(X, Y)$ coupled to the joint estimator is used for the marginal entropy. For the performance evaluation of these methods refer to [GOV16].

Bibliography

[GOV16] Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. In *Advances in Neural Information Processing Systems*, pages 2460–2468, 2016.

[HJ96] Nils Lid Hjort and MC Jones. Locally parametric nonparametric density estimation. *The Annals of Statistics*, pages 1619–1647, 1996.

[JSH99] MC Jones, DF Signorini, and Nils Lid Hjort. On multiplicative bias correction in kernel density estimation. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 422–430, 1999.

[KSG04] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.

[Loa99] Clive R Loader. Bandwidth selection: classical or plug-in? *Annals of Statistics*, pages 415–438, 1999.

[Loa06] Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006.

[Was06] Larry Wassermann. All of nonparametric statistics, 2006.