International Journal of Statistics and Probability; Vol. 5, No. 3; May 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Properties of Transmetric Density Estimation

Sigve Hovda[1]

[1] Department of Petroleum Engineering and Applied Geophysics, Norwegian University of Science and Technology, 7491 Trondheim, Norway

Correspondence: Sigve Hovda, E-mail: sigve.hovda@ntnu.no

Received: December 7, 2015 Accepted: December 21, 2016 Online Published: April 14, 2016

doi:10.5539/ijsp.v5n3p63 URL: http://dx.doi.org/10.5539/ijsp.v5n3p63

Abstract

Transmetric density estimation is a generalization of kernel density estimation that is proposed in Hovda (2014) and Hovda (2016), This framework involves the possibility of making assumptions on the kernel of the distribution to improve convergence orders and to reduce the number of dimensions in the graphical display. In this paper we show that several state-of-the-art nonparametric, semiparametric and even parametric methods are special cases of this formulation, meaning that there is a unified approach.

Moreover, it is shown that parameters can be trained using unbiased cross-validation. When parameter estimation is included, the mean integrated squared error of the transmetric density estimator is lower than for the common kernel density estimator, when the number of dimensions is larger than two.

Keywords: kernel density estimation, semiparametric models, pseudometrics, nonparametric functional data analysis, cross-validation

1. Introduction

The common kernel density estimator with a fixed bandwidth is given by

$$\hat{f}_{\mathbf{H}}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\boldsymbol{x} - \boldsymbol{X}_i), \tag{1}$$

where $K_{\mathbf{H}}(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}$ is a single modal, symmetric, nonnegative and zero-mean function that integrates to 1 (Fix and Hodges, 1951; Rosenblatt, 1956; Parzen, 1962; Cacoullos, 1966; Epanechnikov, 1969; Deheuvels, 1977).

A generalization is the transmetric density estimatior, which is described in Hovda (2016) This framework is motivated by the work in Hovda (2014), which involved using only one transmetric. In Hovda (2014), the bias and the variance are described using Monte Carlo simulations, while asymptotic arguments are given in Hovda (2016) In the last paper, it is shown that the convergence order of the mean integrated squared error can be as high as 4/5. The important point is that the convergence order is independent of the number of dimensions.

For comparison, the convergence orders of common kernel density estimators are $4/(4 + m)$, where m is the number of dimensions (Epanechnikov, 1969). This improvement in accuracy over common kernel density estimation is motivating the path for finding useful and practical applications of this theory.

This paper contains two contributions in this direction. The first contribution involves a dedicated chapter that shows how a number of state-of-the-art problems are special cases of transmetric density estimation. The examples includes distributions with elliptic level sets (Liescher, 2005), linear regression, partial linear models (Härdle et al., 2000), projection pursuit models (Friedman et al., 1984) and also an example of nonparametric functional analysis (Ferraty and Vieu, 2006). Based on these examples, it should be straightforward to see that other methods can be formalized in this way as well. This clarifies the relationship between methods and opens up for new ways to combine.

The second contribution is a description of how unbiased cross-validation can be used for parameter selection. The parameters in the model are the bandwidth matrix and the parameters related to the transmetric. In normal kernel density estimation, there are generally two methods that have received some attention, namely plug-in methods (Duong and Hazelton, 2003). and cross-validation methods (Duong and Hazelton, 2005). The plug-in methods require an analytic expression of the asymptotic mean square error. The special cases that are outlined in Hovda(2016) are complex and it is probably impossible in the general case.

In the case of cross-validation, there are more opportunities. In Duong and Hazelton (2005), the unbiased, the biased and the smoothed cross-validation methods are described and compared. The biased cross-validation methods requires

Table 1. List of important symbols

Symbol	Explanation and properties
n	Number of elements in data matrix.
$\boldsymbol{x}$	A point in the sample space.
x_i	The i'th component of $\boldsymbol{x}$.
$\boldsymbol{X_i}$	The i'th data point of a data set.
$X_{i,j}$	The j'th component of $\boldsymbol{X_i}$
m	Number of dimensions.
q	Number of transmetrics.
$\{P_1, P_2, ...P_q\}$	Partition of $\{1, 2, ...m\}$.
m_j	Size of P_j, i.e. $\sum_{j=1}^{q} m_j = m$.
$\mathbf{H}$	Symmetric bandwidth matrix of size $m \times m$.
$\boldsymbol{x_j}$	The P_j elements of $\mathbf{H}^{-1/2}\boldsymbol{x}$.
$\boldsymbol{y_j}$	The P_j elements of $\mathbf{H}^{-1/2}\boldsymbol{y}$.
$\boldsymbol{X_{i,j}}$	The P_j elements of $\mathbf{H}^{-1/2}\boldsymbol{X_i}$.
$\boldsymbol{c_j}$	Center point parameter in the jth transmetric (not the center of the balls).
$\boldsymbol{p_j}$	Vector of positive exponents in the j'th transmetric.
$\tilde{p}_j$	Harmonic average of the elements in $\boldsymbol{p_j}$.
$V_{\boldsymbol{p_j}}$	Volume of the generalized unit ball with parameters $\boldsymbol{p_j}$.
$V_{t,j}$	Volume of the unit ball in the jth transmetric space.
$\boldsymbol{c}$	Concatenated vector of all the $\boldsymbol{c_j}$s. Size is m.
c_j	The jth component of $\boldsymbol{c}$, not any of the $\boldsymbol{c_j}$s.
$\boldsymbol{p}$	Concatenated vector of all the $\boldsymbol{p_j}$s. Size is m.
p_j	The jth component of $\boldsymbol{p}$, not any of the $\boldsymbol{p_j}$s.
$d_{type,j,par}$	The jth premetric of type $type$ with parameters par.
$\boldsymbol{d_{type}}$	Tuple of q premetrics of type $\boldsymbol{type}$.

an expression for asymptotic mean square error and is not suitable for transmetric density estimators. The unbiased cross-validation methods is called unbiased since it is designed to improve the mean integrated squared error and not the asymptotic version. The smoothed cross-validation method is similar, but more complex as it involves finding a pilot bandwidth matrix. The main conclusion of the paper is that the smoothed cross-validation method is the most reliable. However, it also points out that the unbiased cross-validation method has reliable performance on several distributions.

In this paper, we have chosen to develop an expression for the unbiased cross-validation method. This is because of its algorithmic simplicity and computational efficiency.

In section 2, the definition of transmetric density estimation and relevant theorems from Hovda (2016) are repeated. Relationship to other methods is described in section 3, while the unbiased cross-validation method is outlined and discussed using Monte Carlo simulations in section 4. The paper is concluded in section 6. Table 1 contains a list of symbols that are frequently used in this paper.

2. Transmetric Density Estimation

In Hovda (2016) the transmetric is defined as:

Definition 2.1. *A* transmetric *on a set* $\mathbb{R}^m$ *is a function (distance function)* $d_t : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+$*, which for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$ *and* $\epsilon \in \mathbb{R}^+$*, satisfies the following conditions:*

1. $d_t(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ *(nonnegativity)*
2. $d_t(\boldsymbol{x}, \boldsymbol{x}) = 0$ *(mild identity)*
3. $(\mathbb{R}^m, \{\mathbb{B}_{d_t}(\boldsymbol{x}, \epsilon) \,|\, \epsilon > 0 \;\&\; \boldsymbol{x} \in \mathbb{R}^m\})$ *is a topological space* *(topology)*
4. $V_{\mathbb{B}_{d_t}(\boldsymbol{x},\epsilon)} = V_{\mathbb{B}_{d_t}(\boldsymbol{y},\epsilon)}$ *(translation invariance of ball volumes)*

The two first criteria is the definition of a premetric. A transmetric is therefore a premetric with the additional criteria of topology and translation invariance of ball volumes. All metrics that induces a norm are transmetrics, which is practically all metrics that are used in common kernel density estimation.

The key idea of this definition is to generalize the metric that induces a norm as much as possible, but still make sense for kernel density estimation. The first two criteria ensures that *close* points (in the sense of the transmetric) are weighted higher than points that are *further* away. The third criterion implies that if the kernel is continuous, then the estimator is also. The last criterion in definition 2.1, ensures that the variance of the estimator is everywhere proportional to the density (Hovda, 2016). It is worth emphasizing that relaxing the identity criterion yield some opportunities in smoothing along level sets of the distribution.

Morevover, we also need this definition:

Definition 2.2. *A kernel K is said to be* associated to a transmetric space $(\mathbb{R}^m, d_t)$*, or* associated to a transmetric d_t*, if K is nonnegative, monotonically decreasing with compact support and satisfy*

$$\int_{\mathbb{R}^m} K(d_t(\boldsymbol{x},\boldsymbol{y}))d\boldsymbol{y} = 1,$$

in the case when the unit ball $V_{\mathbb{B}_{d_t}(\boldsymbol{x},1)}$ is finite and in the infinite case

$$\int_{\mathbb{R}^m} K(d_t(\boldsymbol{x},\boldsymbol{y}))d\boldsymbol{y} \propto V_{\mathbb{B}_{d_t}(\boldsymbol{x},1)}.$$

In the case when $V_{\mathbb{B}_{d_t}(\boldsymbol{x},1)}$ is infinite, it is not meaningful to set the integral equal to one, since the kernel would clearly approach zero everywhere. To elaborate on this, we define the equivalence relation $\sim_{d_t}$ as for all $\boldsymbol{x},\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{x} \sim_{d_t} \boldsymbol{y}$ iff $d_t(\boldsymbol{x},\boldsymbol{y}) = 0$. It is clear that all equivalence classes in $\mathbb{R}^m$ with respect to $\sim_{d_t}$ are unbounded. To compensate for this unboundedness, we have decided to let the integral be proportional to $V_{\mathbb{B}_{d_t}(\boldsymbol{x},1)}$. Another way to handle this, would be to impose boundaries on the transmetric space (i.e let it be a subset of $\mathbb{R}^m$), such that all $\mathbb{B}_{d_t}(\boldsymbol{x},1)$s would be bounded.

Two useful examples of kernels are when the transmetric is either Euclidean or Chebyshev distance

$$K(\epsilon) = \frac{1}{(2\pi)^{\frac{m}{2}}} \exp\left(-\frac{1}{2}\epsilon^2\right). \tag{2}$$

$$K(\epsilon) = \begin{cases} \frac{1}{3^{\frac{m}{2}} 2^m} & \text{for} \quad |\epsilon| \leq \sqrt{3} \\ 0 & \text{for} \quad |\epsilon| > \sqrt{3}. \end{cases} \tag{3}$$

Transmetric density estimation is defined here.

Definition 2.3. *Let $\{P_1, P_2, ...P_q\}$ be a partition of $\{1,2,...m\}$, where $m_j = |P_j|$. Let $\boldsymbol{x_j}$, $\boldsymbol{y_j}$ and $\boldsymbol{X_{i,j}}$ be vectors that, respectively, consist of the P_j elements of $\mathbf{H}^{-1/2}\boldsymbol{x}$, $\mathbf{H}^{-1/2}\boldsymbol{y}$ and $\mathbf{H}^{-1/2}\boldsymbol{X_i}$. If we choose a tuple of transmetrics $\boldsymbol{d_t}$, such that for every P_j, there is a K_j associated with the transmetric space $(\mathbb{R}^{m_j}, d_{t,j}(\boldsymbol{x_j},\boldsymbol{y_j}))$, then the* transmetric density estimator *is defined as*

$$\begin{aligned} \hat{f}_{\mathbf{H},\boldsymbol{d_t}}(\boldsymbol{x}) &= |\mathbf{H}|^{-\frac{1}{2}} \int_{\mathbb{R}^m} \prod_{j=1}^{q} K_j(d_{t,j}(\boldsymbol{x_j},\boldsymbol{y_j}))\hat{f}(\boldsymbol{y})d\boldsymbol{y} \\ &= n^{-1}|\mathbf{H}|^{-\frac{1}{2}} \sum_{i=1}^{n} \prod_{j=1}^{q} K_j(d_{t,j}(\boldsymbol{x_j},\boldsymbol{X_{i,j}})), \end{aligned}$$

where $\hat{f}(\boldsymbol{y})$ is equal to $1/n$ at the $\boldsymbol{X_i}$s and 0 elsewhere.

The idea is to weight the contributions of the $\boldsymbol{X_{i,j}}$s according to the distances in the transmetric spaces. It is very important to note that in the case when any of the transmetrics has an unbounded unit ball, $\hat{f}_{\mathbf{H},\boldsymbol{d_t}}(\boldsymbol{x})$ is only proportional to $f(\boldsymbol{x})$. As said before, this is because the whole $\mathbb{R}^m$ is considered, rather than a subset where all transmetric spaces have bounded unit balls.

In the trivial case, all P_j are equal to $\{j\}$ and for all j, $d_{t,j}(\boldsymbol{x_j},\boldsymbol{y_j}) = d_j = |\boldsymbol{x_j} - \boldsymbol{y_j}|$. In this case $\hat{f}_{\mathbf{H},\boldsymbol{d_t}}(\boldsymbol{x}) = \hat{f}_{\mathbf{H},\boldsymbol{d}}(\boldsymbol{x})$, where $\boldsymbol{d}$ is the tuple of d_js. If we let $\prod_{j=1}^{m} K_j(u_j) = K(\boldsymbol{u})$, then we arrive at the estimator defined in equation (1).

It is easy to see how any metric that defines a normed vector space can be used, but the interesting question is what other options do we have. What transmetrics make sense to use and which kernels can be associated to them? It is of particular interest to investigate the opportunities of relaxing the identity criterion. This motivates the definition of *associated distributions*:

Definition 2.4. *A probability density function is said to be an* associated distribution of $\hat{f}_{\mathbf{H},d_t}$, *denoted as* $f_{\sim_{\mathbf{H},d_t}} : \mathbb{R}^m \to \mathbb{R}^+$, *when all* $\boldsymbol{x_j} \sim_{d_{t,j}} \boldsymbol{y_j}$, *imply that* $f_{\sim_{\mathbf{H},d_t}}(\boldsymbol{x}) = f_{\sim_{\mathbf{H},d_t}}(\boldsymbol{y})$ *for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^m$. *Here, the equivalence relation* $\sim_{d_{t,j}}$ *is defined as for all* $\boldsymbol{x_j}, \boldsymbol{y_j} \in \mathbb{R}^{m_j}$, $\boldsymbol{x_j} \sim_{d_{t,j}} \boldsymbol{y_j}$ *iff* $d_{t,j}(\boldsymbol{x_j}, \boldsymbol{y_j}) = 0$.

While discussing the usefulness of a family of transmetrics, it makes sense to ask which density functions are invariant to the equivalence relations that are implied by the transmetrics.

Another point to note is that the graphical display of the estimated density can be reduced to a q-dimensional display. This is because the estimated density at a certain point $\boldsymbol{x}$ is constant on $\{\boldsymbol{y} \in \mathbb{R}^m \,|\, \boldsymbol{y_j} \in [\boldsymbol{x_j}]_{\sim_{d_{t,j}}}\}$.

All the transmetric spaces that will be discussed in this paper, have the following property regarding how the volumes of the balls vary as a function of the radius.

Definition 2.5. *A transmetric* d_t *or a transmetric space* $(\mathbb{R}^m, d_t)$, *is said to be* of order u, *if the volume of the ball is on the form* $V_{\mathbb{B}_{d_t}(\boldsymbol{x},\epsilon)} = V_{\mathbb{B}_{d_t}(\boldsymbol{x},1)}\epsilon^u = V_t\epsilon^u$, *where* $\boldsymbol{x} \in \mathbb{R}^m$ *and* u *is a positive integer. Such transmetrics are denoted* $d_{t,V_t,u}$.

This property is useful as it makes it easy to find associated kernels.

Theorem 2.1. *Given a transmetric space* $(\mathbb{R}^m, d_{t,V_t,u})$ *of order* u, *then for any function* $g : \mathbb{R}^+ \to X$, *where* $X \subset \mathbb{R}$, *that is bounded and continuous almost everywhere, the following property is valid*

$$\int_{\mathbb{B}_{d_{t,V_t,u}}(\boldsymbol{x},U)} g(d_{t,V_t,u}(\boldsymbol{x},\boldsymbol{y}))d\boldsymbol{y} = uV_t \int_0^U g(\epsilon)\epsilon^{u-1}d\epsilon,$$

where U *is finite valued. If* g *also has compact support then*

$$\int_{\mathbb{R}^m} g(d_{t,V_t,u}(\boldsymbol{x},\boldsymbol{y}))d\boldsymbol{y} = uV_t \int_0^\infty g(\epsilon)\epsilon^{u-1}d\epsilon.$$

A direct consequence of this is that the normal and the uniform associated kernels of $d_{t,V_t,u}$ are:

$$K(\epsilon) = \frac{1}{2^{\frac{u}{2}}\Gamma(\frac{u}{2})V_t} \exp\left(-\frac{1}{2}\epsilon^2\right) \tag{4}$$

and

$$K(\epsilon) = \begin{cases} \frac{1}{3^{\frac{u}{2}}V_t} & \text{for} \quad |\epsilon| \leq \sqrt{3} \\ 0 & \text{for} \quad |\epsilon| > \sqrt{3}, \end{cases} \tag{5}$$

when V_t is finite and in the infinite case, equation (2) and (3) can be used.

An important source for finding transmetrics is using pseudometrics. In general the volumes of the balls in pseudometric spaces with constant radius are dependent on the locations of the centers. This means that they can not be used directly. However, the following theorem shows how transmetrics can be designed from pseudometrics that are inducing a topology on $\mathbb{R}^m$.

Theorem 2.2. *If* $(\mathbb{R}^m, d_p)$ *is a pseudometric space and* $(\mathbb{R}^m, \{\mathbb{B}_{d_p}(\boldsymbol{x},\epsilon) \,|\, \epsilon > 0 \ \&\ \boldsymbol{x} \in \mathbb{R}^m\})$ *is a topological space, then* $d_{t,V_t,u}$ *is a transmetric of order* u, *if*

$$d_{t,V_t,u}(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{V_{\mathbb{B}_{d_p}(\boldsymbol{x},d_p(\boldsymbol{x},\boldsymbol{y}))}}{V_t}\right)^{\frac{1}{u}},$$

where V_t *is chosen to be the volume of the unit ball of the transmetric space* $(\mathbb{R}^m, d_{t,V_t,u})$.

Proofs of theorem 2.1 and 2.2 are found in Hovda (2016) To fix ideas, the framework described in this section is required to discuss the transmetrics as of Hovda (2014). Moreover, the next section demonstrates that many state-of art problems can be described using this framework.

3. Relationship to State-of-the-art Problems

The purpose of this section is to describe the relationship to other methods and to show the flexibility of modeling problems with transmetric density estimators. The discussion is far from extensive, but the selection is chosen to show some of the generality of definition 2.3.

3.1 Distributions with Elliptic Level Sets

In a paper by Liescher (2005), the level sets of the distributions are constrained to be of elliptic shape. The elliptic shape of the level sets is constrained by a transformation that is performed prior to applying the nonparametric density estimator. The estimators in Hovda (2014) and Hovda (2016) can be viewed as a generalization of the estimators of Liescher (2005). This is because a wider family of distributions is allowed. In Liescher (2005), it is shown that the convergence rates are independent of the number of dimensions, except in the neighborhood of the mode. This result complies with the results in Hovda (2014) and Hovda (2016)

3.2 Nonparametric Functional Data Analysis

As said in the introduction, pseudometrics are commonly used in nonparametric functional data analysis. A function can be viewed as a point in an infinite dimensional space. It is worth noting that the number of dimensions in the definition 2.3 can approach infinity, but the number of dimensions can only be countable infinite. This is less general than the uncountable infinite number of dimensions that is needed when the domain of the functions are for instance $\mathbb{R}$. In this respect, definition 2.3 is a specification of what is common in nonparametric functional data analysis. There are two reasons for this specification. First, the specification has little practical implication as functions are usually discretized. Second, it is probably possible, but outside the scope of this paper, to generalize definition 2.3 to also include variables of uncountable infinite dimensions.

In chapter 3.4.1 in the book of Ferraty and Vieu (2006), some finite dimensional pseudometrics are given. As an example, we mention a pseudometric that is based on the first c principal components of the data

$$d_t(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{c} \left(\sum_{j=1}^{m} (x_j - y_j) w_j v_{ij} \right)^2}, \tag{6}$$

where the discretized representations of the functions $\boldsymbol{x}$ and $\boldsymbol{y}$ have m entries each and the w_js are the weights that defines the approximate integration. Moreover, v_{ij} is the jth coordinate of the ith eigenvector of the covariance matrix of a relevant dataset.

This is a transmetric of order c, where V_t is clearly infinite and $V_t = lim_{L\to\infty} L^{m-c} V_c$, where V_c is the volume of the c-sphere. The c dimensional kernel as defined in equation (2) is an associated kernel. If we choose $\mathbf{H}^{\frac{1}{2}} = h\mathbf{I}$, the estimator in our notation is simply

$$\hat{f}_{\mathbf{H},d_t}(\boldsymbol{x}) = n^{-1} h^{-1} \sum_{i=1}^{n} K(h^{-1} d_t(\boldsymbol{x}, \boldsymbol{X}_i)), \tag{7}$$

which is analog to what is found in Ferraty and Vieu (2006). It is worth noting that the associated distributions of this transmetric is the family of all functions that can be described as a linear combination of the first c eigenvectors.

Other pseudometrics that are discussed in Ferraty and Vieu (2006) are either based on partial least squares or the L^2-norm of the derivatives of some order. Without derivations, it should be clear that the finite versions of these pseudometrics are also transmetrics.

3.3 Linear Regression

We define a transmetric space $(\mathbb{R}^m, d_t)$, where

$$d_t(\boldsymbol{x}, \boldsymbol{y}) = \left| (x_1 - y_1) - \sum_{j=2}^{m} \beta_{j-1} (x_j - y_j) \right|. \tag{8}$$

This is a transmetric of order one, where $V_t = lim_{L\to\infty} 2L^{m-1}$. The one-dimensional uniform kernel taken from equation (3) is associated with d_t. If we choose $\mathbf{H}^{\frac{1}{2}} = \mathbf{I}$, then

$$\hat{f}_{\mathbf{I},d_t}(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} K(d_t(\boldsymbol{x}, \boldsymbol{X}_i)).$$

The regression model is found by investigating the expected value of x_1 given $x_2, x_3, ...x_m$. Therefore,

$$
\begin{aligned}
E(x_1|x_2, x_3, ...x_m) = \int_{\mathbb{R}} x_1 f(x_1|x_2, x_3, ...x_m) dx_1 &= \frac{\int_{\mathbb{R}} x_1 f(\boldsymbol{x}) dx_1}{\int_{\mathbb{R}} f(\boldsymbol{x}) dx_1} \\
&\approx \frac{\int_{\mathbb{R}} x_1 \hat{f}_{\mathbf{I},\boldsymbol{d_t}}(\boldsymbol{x}) dx_1}{\int_{\mathbb{R}} \hat{f}_{\mathbf{I},\boldsymbol{d_t}}(\boldsymbol{x}) dx_1} \\
&= n^{-1} \sum_{i=1}^{n} \int_{\mathbb{R}} x_1 K(d_t(\boldsymbol{x}, \boldsymbol{X_i})) dx_1 \\
&= n^{-1} \sum_{i=1}^{n} \left(X_{i,1} - \sum_{j=2}^{m} \beta_{j-1}(X_{i,j} - x_j) \right) \\
&= \sum_{j=2}^{m} \beta_{j-1} x_j + n^{-1} \sum_{i=1}^{n} \left(X_{i,1} - \sum_{j=2}^{m} \beta_{j-1}(X_{i,j}) \right),
\end{aligned}
$$

which describes the linear regression model. The intercept is zero when the data points are divided by the individual sample means. Obviously, this result has no effect on the parameter selection method, which is for instance minimizing the squared residuals. Based on this, it should be clear that other parametric regression models, such as polynomial regression, can be described in the sense of transmetric density estimation.

3.4 Partially Linear Models

Partially linear models is given by

$$
X_{i,1} = \sum_{j=2}^{m_1} (\beta_{j-1} X_{i,j}) + g(X_{i,m_1+1}, X_{i,m_1+2}, .., X_{i,m}) + u_i,
$$

where $\sum_{j=2}^{m_1} (\beta_{j-1} X_{i,j})$ is the linear part and $g(X_{i,m_1+1}, X_{i,m_1+2}, ..., X_{i,m})$ is the nonparametric part. The data points are assumed to be i.i.d. and $E(u_i|X_{i,2}, X_{i,3}, , ..., X_{i,m}) = 0$. An estimator for $g(x_{m_1+1}, x_{m_1+2}, ..., x_m)$ is found by investigating

$$
\begin{aligned}
&E(x_1 - \sum_{j=2}^{m_1} \beta_{j-1} x_j | x_{m_1+1}, x_{m_1+2}, ..., x_m) \\
&= \int_{\mathbb{R}^{m_1}} (x_1 - \sum_{j=2}^{m_1} \beta_{j-1} x_j) f(x_1, x_2, ..., x_{m_1} | x_{m_1+1}, x_{m_1+2}, ..., x_m) dx_1 dx_2 ... dx_{m_1} \\
&= \frac{\int_{\mathbb{R}^{m_1}} (x_1 - \sum_{j=2}^{m_1} \beta_{j-1} x_j) f(\boldsymbol{x}) dx_1 dx_2 ... dx_{m_1}}{\int_{\mathbb{R}^{m_1}} f(\boldsymbol{x}) dx_1 dx_2 ... dx_{m_1}}.
\end{aligned}
$$

We make an estimate of g, denoted $\hat{g}$, by inserting an estimator for $f(\boldsymbol{x})$. We choose a tuple of transmetrics $\boldsymbol{d_t}$, where $d_{t,1}$ is a m_1-dimensional version of equation (8), associated with a m_1-dimensional kernel as of equation (3). The other transmetrics are one-dimensional, associated with one-dimensional kernels. In this case

$$
\hat{f}_{\mathbf{H},\boldsymbol{d_t}}(\boldsymbol{x}) = n^{-1} |\mathbf{H}|^{-\frac{1}{2}} \sum_{i=1}^{n} \prod_{j=1}^{q} K_j(d_{t,j}(\boldsymbol{x_j}, \boldsymbol{X_{i,j}})), \tag{9}
$$

where $q = m - m_1 + 1$. If we choose $\mathbf{H}$ so that $\boldsymbol{x_1} = \{x_1, x_2, ..., x_{m_1}\}$, then

$$
\begin{aligned}
&\hat{g}(x_{m_1+1}, x_{m_1+2}, ..., x_m)\\
&= \sum_{i=1}^{n} \Bigg[\int_{\mathbb{R}^{m_1}} (x_1 - \sum_{j=2}^{m_1} \beta_{j-1} x_j) K_1(d_{t,1}(\boldsymbol{x_1}, \boldsymbol{X_{1,i}})) d\boldsymbol{x_1} \times \\
&\frac{\prod_{j=2}^{q} K_j(d_{t,j}(\boldsymbol{x_j}, \boldsymbol{X_{i,j}}))}{\sum_{i=1}^{n} \int_{\mathbb{R}^{m_1}} K_1(d_{t,1}(\boldsymbol{x_1}, \boldsymbol{X_{1,i}})) d\boldsymbol{x_1} \prod_{j=2}^{q} K_j(d_{t,j}(\boldsymbol{x_j}, \boldsymbol{X_{i,j}}))} \Bigg] \\
&= \sum_{i=1}^{n} w_i(x_{m_1+1}, x_{m_1+2}, ..., x_m)(X_{1,i} - \sum_{j=2}^{m_1} \beta_{j-1} X_{j,i}) \quad \text{where} \\
&w_i(x_{m_1+1}, x_{m_1+2}, ..., x_m) = \frac{\prod_{j=2}^{q} K_j(d_{t,j}(\boldsymbol{x_j}, \boldsymbol{X_{i,j}}))}{\sum_{i=1}^{n} \prod_{j=2}^{q} K_j(d_{t,j}(\boldsymbol{x_j}, \boldsymbol{X_{i,j}}))}.
\end{aligned}
$$

Here $w_i(x_{m_1+1}, x_{m_1+2}, ..., x_m)$ are the weight functions in the partially linear model (Häardle et al., 2000).

Projection Pursuit

In the context of reducing the curse of dimensionality of nonparametric density estimators, it is also worth mentioning the *projection pursuit density estimators*. This method was first introduced in Friedman, Stuezle and Schroeder (1984), and a parametric extension was given by Welling, Zemel and Hinton (2003). In projection pursuit one projects the explanatory variables into principal directions and fits one-dimensional smooth density functions to these projections. The resulting density is the product of these densities.

The analogous projection pursuit density estimator can be described as a product of transmetric density estimators as

$$
\hat{f}_{pp}(\boldsymbol{x}) = \prod_{j=1}^{Q} \hat{f}_{\mathbf{H_j}, \boldsymbol{d}_{t,j}}(\boldsymbol{x_j}),
$$

where $\boldsymbol{x_j}$ is a vector in the projected space, which only contain a subset of the coordinates. In the special case, when all the $\boldsymbol{d_{t,j}}$s contains one transmetric each and all transmetrics are one-dimensional, the usual projection pursuit model appears.

A similarity between the transmetric density estimators and the projection pursuit density estimators is that the resulting density can be visualized in a lower dimensional space. In projection pursuit, it is enough to show the one-dimensional ridge functions along with the principal directions to understand the full distribution.

4. Parameter Estimation by Cross-validation

The global error criteria to be minimized in the unbiased cross-validation method is the mean integrated squared error that is defined as

$$
\begin{aligned}
\text{MISE}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}) &= E\left(\int_{\mathbb{R}^m} (\hat{f}_{\mathbf{H},\boldsymbol{d}_t}(\boldsymbol{x}) - f(\boldsymbol{x}))^2 d\boldsymbol{x} \right) \\
&= \int_{\mathbb{R}^m} \left(\text{Bias}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}(\boldsymbol{x})) \right)^2 d\boldsymbol{x} + \int_{\mathbb{R}^m} \text{Var}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}(\boldsymbol{x})) d\boldsymbol{x}.
\end{aligned} \tag{10}
$$

This expression is also known as the L^2 risk function. The unbiased cross-validation (UCV) method aims to minimize MISE and employs the objective function

$$
\begin{aligned}
\text{UCV}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}) &= \int_{\mathbb{R}^m} (\hat{f}_{\mathbf{H},\boldsymbol{d}_t}(\boldsymbol{x}))^2 d\boldsymbol{x} - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{\mathbf{H},\boldsymbol{d}_t(-i)}(\boldsymbol{X_i}) \quad \text{where} \\
\hat{f}_{\mathbf{H},\boldsymbol{d}_t(-i)}(\boldsymbol{x}) &= (n-1)^{-1} |\mathbf{H}|^{-\frac{1}{2}} \sum_{\substack{j=1,\\ j \neq i}}^{n} \prod_{k=1}^{q} K_k(d_{t,k}(\boldsymbol{x_k}, \boldsymbol{X_{k,j}}))
\end{aligned} \tag{11}
$$

is a *leave-one-out* estimator of f. The function $\text{UCV}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t})$ is unbiased in the sense that the expected value of $\text{UCV}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t})$ is equal to $\text{MISE}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}) - R(f)$. Here, $R(f)$ is the L^2-norm of f. The second term of equation (11) can be expanded to

$$-2n^{-1}(n-1)^{-1}|\mathbf{H}|^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{\substack{j=1,\\ j\neq i}}^{n}\prod_{k=1}^{q}K_k(d_{t,k}(\boldsymbol{X_{k,i}},\boldsymbol{X_{k,j}})) \tag{12}$$

which is straightforward to treat computationally. The first term of equation (11) can be expanded to

$$n^{-2}|\mathbf{H}|^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}\prod_{k=1}^{q}\int_{\mathbb{R}^m}K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,i}}))K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,j}}))d\boldsymbol{x_k}. \tag{13}$$

In the case when i is equal to j, equation (13) becomes

$$\begin{aligned}&n^{-1}|\mathbf{H}|^{-\frac{1}{2}}\prod_{k=1}^{q}R(K_k)+n^{-1}(n-1)^{-1}|\mathbf{H}|^{-\frac{1}{2}}\sum_{i=1}^{n}\sum_{\substack{j=1,\\ j\neq i}}^{n}\prod_{k=1}^{q}c_{ijk},\\ &\text{where}\quad c_{ijk}=\int_{\mathbb{R}^m}K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,i}}))K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,j}}))d\boldsymbol{x_k}\end{aligned} \tag{14}$$

and $R(K_k)$ is defined as

$$R(K_k)=\int_{\mathbb{R}^{m_k}}K_k^2(d_{t,k}(\boldsymbol{x_k},\boldsymbol{y_k}))d\boldsymbol{y_k}.$$

If all transmetrics in $\boldsymbol{d_t}$ has an order, theorem 2.1 can be used to obtain analytical expressions. Notice that the second part of equation (14) is an expression of the variance part of MISE. The challenge with estimating UCV$((\hat{f}_{\mathbf{H},\boldsymbol{d_t}})$ is to estimate the c_{ijk}s in equation (14). Two special cases are treated below.

4.1 Spesial Case When Transmetrics are of the Type Described in Hovda (2014)

Theorem 4.1. *If we define* $\beta_{x,k}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}d_{s,k,\boldsymbol{p_k}}(\boldsymbol{x_k},\boldsymbol{c_k})$, $\beta_{x,k,i}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}d_{s,k,\boldsymbol{p_k}}(\boldsymbol{X_{k,i}},\boldsymbol{c_k})$ *and* $\beta_{x,k,j}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}d_{s,k,\boldsymbol{p_k}}(\boldsymbol{X_{k,j}},\boldsymbol{c}$ *then*

$$\begin{aligned}&\int_{\mathbb{R}^m}K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,i}}))K_k(d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,j}}))d\boldsymbol{x_k}=\\ &m_kV_{t,k}\int_0^\infty K_k(U_{m_k}(\beta_{x,k},\beta_{x,k,i}))K_k(U_{m_k}(\beta_{x,k},\beta_{x,k,j}))\beta_{x,k}^{(m_k-1)}d\beta_{x,k}.\end{aligned} \tag{15}$$

Proof. We start by defining $\alpha_{x,k}=d_{s,k,\boldsymbol{p_k}}(\boldsymbol{x_k},\boldsymbol{c_k})$, $\alpha_{x,k,i}=d_{s,k,\boldsymbol{p_k}}(\boldsymbol{X_{k,i}},\boldsymbol{c_k})$ and $\alpha_{x,k,j}=d_{s,k,\boldsymbol{p_k}}(\boldsymbol{X_{k,j}},\boldsymbol{c_k})$. Furthermore, we define the function $d_{t1,k}(\alpha_{x,k},\alpha_{x,k,i})=d_{t,k}(\boldsymbol{x_k},\boldsymbol{X_{k,i}})$ and since $\alpha_{x,k}$ is a transmetric of order m_k, we can apply theorem 2.1. The integral has therefore reduced the number of dimensions to one

$$m_kV_{\boldsymbol{p_k}}\int_0^\infty K_k(d_{t1,k}(\alpha_{x,k},\alpha_{x,k,i}))K_k(d_{t1,k}(\alpha_{x,k},\alpha_{x,k,j}))\alpha_{x,k}^{(m_k-1)}d\alpha_{x,k}.$$

The proof is concluded by changing variables to $\beta_{x,k}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}\alpha_{x,k}$, $\beta_{x,k,i}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}\alpha_{x,k,i}$ and also $\beta_{x,k,j}=(V_{\boldsymbol{p_k}}/V_{t,k})^{1/m_k}\alpha_{x,k,j}$. This means that the arguments in the kernels can be expressed by the U_{m_k} function. □

The integral is now one-dimensional, but unfortunately the integral is not trivial to solve for arbitrary m_k. For most choices of kernel, we must rely on numerical solutions. In those cases, it is only necessary to integrate in regions that are close to $\beta_{x,k,i}$ and $\beta_{x,k,j}$.

If we define these two functions

$$\begin{aligned}C_{1,s}(u,v)&=\begin{cases}3^{-\frac{s}{2}}\left[\min(u+r_s(u),v+r_s(v))^s-\max(u-r_s(u),v-r_s(v))^s\right]\\ \qquad\text{for}\quad u-r_s(u)-r_s(v)\leq v\leq u+r_s(u)+r_s(v)\\ 0\qquad\text{else}\end{cases}\\ C_{2,s}(u,v)&=\begin{cases}1\qquad\text{for}\quad u-r_s(u)\leq v\leq u+r_s(u)\\ 0\qquad\text{else,}\end{cases}\end{aligned} \tag{16}$$

choose a uniform kernel and let $V_{t,k} = 2^{m_k}$, then equation (15) is simply

$$3^{-\frac{m_k}{2}} 2^{-m_k} C_{1,m_k}(\beta_{x,k,i}, \beta_{x,k,j}), \tag{17}$$

which is computationally tractable. This is because each $\beta_{x,k,i}$ and each $r_{m_k}(\beta_{x,k,i})$ can be pre-computed before calculating the double sums in equation (14). Note that $C_{1,m_k}(\beta_{x,k,i}, \beta_{x,k,j})$ is bounded by one, which happens when $\beta_{x,k,i} = \beta_{x,k,j}$.

To summarize, the estimator of UCV for the uniform kernel is therefore

$$\begin{aligned} \mathrm{UCV}(\hat{f}_{\mathbf{H},d_t}) = {} & n^{-1}|\mathbf{H}|^{-\frac{1}{2}} \prod_{k=1}^{q} R(K_k) - \\ & 2n^{-1}(n-1)^{-1}|\mathbf{H}|^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{\substack{j=1,\\ j\neq i}}^{n} \prod_{k=1}^{q} K_k(d_{t,k}(\boldsymbol{X_{k,i}}, \boldsymbol{X_{k,j}})) + \\ & n^{-1}(n-1)^{-1}|\mathbf{H}|^{-\frac{1}{2}} \sum_{i=1}^{n} \sum_{\substack{j=1,\\ j\neq i}}^{n} \prod_{k=1}^{q} 3^{-\frac{m_k}{2}} 2^{-m_k} C_{1,m_k}(\beta_{x,k,i}, \beta_{x,k,j}). \end{aligned} \tag{18}$$

This can be reduced to

$$\begin{aligned} & \mathrm{UCV}(\hat{f}_{\mathbf{H},d_t}) = n^{-1}|\mathbf{H}|^{-\frac{1}{2}} 3^{-\frac{m}{2}} 2^{-m} \times \\ & \left[1 + (n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1,\\ j\neq i}}^{n} \left[\prod_{k=1}^{q} C_{1,m_k}(\beta_{x,k,i}, \beta_{x,k,j}) - 2 \prod_{k=1}^{q} C_{2,m_k}(\beta_{x,k,i}, \beta_{x,k,j}) \right] \right]. \end{aligned} \tag{19}$$

It is worth noting that $C_{1,s}$ is generally smaller than one and that it takes values other than zero more often than $C_{2,s}$. It is also worth noting that the two points $\beta_{x,k,i}$ and $\beta_{x,k,j}$, contribute maximally to decrease UCV when $\beta_{x,k,j} = \beta_{x,k,i} \pm r_{m_k}(\beta_{x,k,i})$. By contrast, when $\beta_{x,k,j}$ is slightly more than $\beta_{x,k,i} + r_{m_k}(\beta_{x,k,i})$ or slightly less than $\beta_{x,k,i} - r_{m_k}(\beta_{x,k,i})$, these two points contribute maximally to increase UCV. This effect is also seen for UCV on common kernel density estimation, when the kernel is uniform.

4.2 Special Case when Transmetrics are Metrics

It is straightforward to see that UCV for the common kernel density estimator, with a uniform kernel as of Duong and Hazelton (2005) is given by

$$\begin{aligned} & \mathrm{UCV}(\hat{f}_{\mathbf{H}}) = n^{-1}|\mathbf{H}|^{-\frac{1}{2}} 3^{-\frac{m}{2}} 2^{-m} \times \\ & \left[1 + (n-1)^{-1} \sum_{i=1}^{n} \sum_{\substack{j=1,\\ j\neq i}}^{n} \left(\prod_{k=1}^{m} C_3(\boldsymbol{X_{k,i}}, \boldsymbol{X_{k,j}}) - 2 \prod_{k=1}^{m} C_4(\boldsymbol{X_{k,i}}, \boldsymbol{X_{k,j}}) \right) \right]. \\ & C_3(u,v) = \begin{cases} 1 - \frac{|u-v|}{2\sqrt{3}} & \text{for} \quad |u-v| \le 2\sqrt{3} \\ 0 & \text{else} \end{cases} \\ & C_4(u,v) = \begin{cases} 1 & \text{for} \quad u - \sqrt{3} \le v \le u + \sqrt{3}, \\ 0 & \text{else,} \end{cases} \end{aligned} \tag{20}$$

which is analogous to the generalized method.

5. Comparison of Transmetric Density Estimation and Kernel Density Estimation Using Monte Carlo Simulations

This section focuses on comparing transmetric density estimation with common kernel density estimation. In particular, AMISE is compared to optimal MISE and optimal UCV is compared to optimal MISE subtracted by $R(f)$. The kernels are uniform and the data sets are drawn from multi-dimensional normal distributions, where the number of dimensions and population sizes are varied.

Table 2. Parameters that are common for all experiments

Parameter	Argument
Distribution type	Standard normal $\mathcal{N}[0,1]$
Number of dimensions (m)	{2,4,8}
Kernel shape	Uniform
Parameters of $\hat{f}_{\mathbf{H}}$	$\mathbf{H}^{1/2} = \lvert\mathbf{H}\rvert^{1/(2m)}\mathbf{I}$
Parameters of $\hat{f}_{\mathbf{H},\boldsymbol{d}_t}$	$\boldsymbol{c} = \mathbf{0}$, the elements of $\boldsymbol{p}$ are 2 and $\mathbf{H}^{1/2} = \lvert\mathbf{H}\rvert^{1/(2m)}\mathbf{I}$

Table 3. Parameters in experiment one and two

Parameter	Argument
Population sizes (n)	{10, 31, 100, ... 100000}
Number of MSE repetitions n_{MSE}	100000
$\mathrm{H}_{\mathrm{AMISE}}(\hat{f}_{\mathbf{H}})$ and $\mathrm{AMISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H}})$	Equation (20) in Hovda (2016), where $M = m(m+2)2^{-m-2}\pi^{-m/2}$
$\mathrm{H}_{\mathrm{AMISE}}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t})$ and $\mathrm{AMISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t})$	Equation (19) in Hovda (2016), where approximation limit $L = 0.05$

5.1 Monte Carlo Simulations of MISE and AMISE in the Case When $q = 1$

From equation (19) in Hovda (2016), we have analytical expressions for the optimal AMISE of $f_{\mathbf{H},d_t}$, $\mathrm{AMISE}_{\mathrm{opt}}(f_{\mathbf{H},d_t})$ with the corresponding optimal bandwidth $\mathrm{H}_{\mathrm{AMISE}}(f_{\mathbf{H},d_t})$. Equation (20) in the same article gives analytical expressions for $\mathrm{AMISE}_{\mathrm{opt}}(f_{\mathbf{H}})$ and $\mathrm{H}_{\mathrm{AMISE}}(f_{\mathbf{H}})$. All these expressions are given as functions of n.

In order to verify these approximations, we have chosen to simulate MISE when $|\mathbf{H}|^{\frac{1}{2}} = \mathrm{H}_{\mathrm{AMISE}}$. This value is denoted $\mathrm{MISE}(\mathrm{H}_{\mathrm{AMISE}})$ and it is compared with $\mathrm{AMISE}_{\mathrm{opt}}$. Moreover, we have also estimated minimal MISE, that is $\mathrm{MISE}_{\mathrm{opt}}$, by varying $|\mathbf{H}|^{\frac{1}{2}}$ around $\mathrm{H}_{\mathrm{AMISE}}$. The $|\mathbf{H}|^{\frac{1}{2}}$ that correspond to $\mathrm{MISE}_{\mathrm{opt}}$ is denoted $\mathrm{H}_{\mathrm{MISE}}$.

The procedure for MISE calculations is straightforward. A region R_f of volume V_{R_f} is identified, where f is assumed to be close to zero outside this region. For fixed f, $\mathbf{H}$ and $\boldsymbol{d_t}$, the following are repeated n_{MSE} times. In iteration i, n data points are drawn randomly from f and placed in a dataset $\boldsymbol{X}^i$. A single sample $\boldsymbol{T}^i$ is drawn from a uniform distribution on R_f. The estimator $\hat{f}^i$ is calculated based on the dataset $\boldsymbol{X}^i$ and the estimator of MISE is calculated by

$$\widehat{\mathrm{MISE}}(\hat{f}_{\mathbf{H},\boldsymbol{d}_t}) = V_{R_f} n_{MSE}^{-1} \sum_{i=1}^{n_{rep}} (\hat{f}^i(\boldsymbol{T}^i) - f(\boldsymbol{T}^i))^2. \tag{21}$$

In experiment one, $\hat{f}_{\mathbf{H}}$ is computed and in experiment two, $\hat{f}_{\mathbf{H},\boldsymbol{d}_t}$ is computed. To summarize, MISE is estimated by taking the average of a number of mean square error MSE estimates with random points taken uniformly within R_f. The relevant parameters are listed in the table 2 and table 3.

5.1.1 Results:

The result of experiment one is shown in figure 1. Here, the common kernel density estimator is evaluated. It is clear that $\mathrm{H}_{\mathrm{AMISE}}$ seems to overestimate $\mathrm{H}_{\mathrm{MISE}}$ for small n, but this effect is smaller for larger n. Since the graphs are on a log scale, the reducing gap between them means that the normalized experimental error $|\mathrm{H}_{\mathrm{AMISE}} - \mathrm{H}_{\mathrm{MISE}}|/\mathrm{H}_{\mathrm{MISE}}$ is also reducing.

Moreover, it seems that $\mathrm{AMISE}_{\mathrm{opt}}$ is an overestimation of $\mathrm{MISE}_{\mathrm{opt}}$, but the normalized experimental error is reduced with increased n. The reason for this is related to that $\mathrm{H}_{\mathrm{AMISE}}$ overestimates $\mathrm{H}_{\mathrm{MISE}}$ on this distribution.

The fact that $\mathrm{MISE}(\mathrm{H}_{\mathrm{AMISE}})$ seems to lie between $\mathrm{MISE}_{\mathrm{opt}}$ and $\mathrm{AMISE}_{\mathrm{opt}}$ is a verification of the approximations given in equation (20) in Hovda (2016) It is also worth noting that this experiment is a verification of the convergence orders for the common kernel density estimators. On a log scale the convergence orders are proportional to the slopes of the graphs and it is clear that the slopes of AMISE are similar to those for MISE. It is also clear that the slopes are levelling out when the number of dimensions is increasing. This is expected.

The result of experiment two is shown in figure 2. Here the transmetric density estimator is evaluated. Similar to the result

Table 4. Parameters in experiment three and four

Parameter	Argument
Population sizes (n)	{10, 31, 100, ... 10000}
Number of minimizations n_{MIN}	1000
Number of MSE repetitions n_{MSE}	100000
UCV($\hat{f}_{\mathbf{H}}$)	Equation (20)
UCV($\hat{f}_{\mathbf{H},d_t}$)	Equation (19)

of experiment one it is clear that $\mathrm{H}_{\mathrm{AMISE}}$ and $\mathrm{AMISE}_{\mathrm{H}}$ are overestimates of $\mathrm{H}_{\mathrm{MISE}}$ and $\mathrm{MISE}_{\mathrm{H}}$, respectively.

Moreover, the normalized experimental errors of both $\mathrm{H}_{\mathrm{AMISE}}$ and $\mathrm{AMISE}_{\mathrm{H}}$ improve for increasing n. This is expected, because the restriction in equation (18) in Hovda (2016), suggest that the approximations of AMISE are only valid for n larger than 10^6. Unfortunately, limited computational power has restricted us from evaluating larger population sizes. However, the convergence of $\mathrm{MISE}_{\mathrm{H}}$ towards $\mathrm{AMISE}_{\mathrm{H}}$ gives trust to the approximations of equation (19) in Hovda (2016) Another verification is that MISE($\mathrm{H}_{\mathrm{AMISE}}$) seems to follow $\mathrm{AMISE}_{\mathrm{opt}}$.

It is clearly worth commenting that for the smallest population sizes, $\mathrm{MISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H},d_t})$ is similar to $\mathrm{MISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H}})$ and $\mathrm{H}_{\mathrm{MISE}}(\hat{f}_{\mathbf{H},d_t})$ is similar to $\mathrm{H}_{\mathrm{MISE}}(\hat{f}_{\mathbf{H}})$. This is because the two methods coincide, when n is small. However, it is obvious that $\mathrm{MISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H},d_t})$ is substantially smaller than $\mathrm{MISE}_{\mathrm{opt}}(\hat{f}_{\mathbf{H}})$ for larger n. This is seen in the left subfigure of figure 5, where the ratio of $\mathrm{MISE}_{opt}(\hat{f}_{\mathbf{H},d_t})$ and $\mathrm{MISE}_{opt}(\hat{f}_{\mathbf{H}})$ is plotted. In eight dimensions and when $n = 100000$, MISE of the transmetric density estimator is 20 times smaller.

5.2 Experiments on cross-validation

In this section, the optimal UCV($\hat{f}_{\mathbf{H},d_t}$) and UCV($\hat{f}_{\mathbf{H}}$) are calculated for various choices of bandwidths. Again the multi-dimensional normal distributions are considered.

The optimal bandwidth $\mathrm{H}_{\mathrm{UCV}}$ is the bandwidth that minimizes UCV, that is UCV_{opt}. The mean integrated squared error that correspond to $\mathrm{H}_{\mathrm{UCV}}$ is denoted $\mathrm{MISE}_{\mathrm{UCV}}$. For a given population size (n), each search for $\mathrm{H}_{\mathrm{UCV}}$ is repeated n_{MIN} times. A distribution of estimates of $\mathrm{H}_{\mathrm{UCV}}$ is therefore available. $\mathrm{MISE}_{\mathrm{UCV}}$ is found by calculating MISE based on the average of n_{MSE} MSE calculations. The bandwidths are chosen randomly from the $\mathrm{H}_{\mathrm{UCV}}$ distribution.

In the experiments, the distributions are normal and known and therefore UCV + $R(f)$ is compared to the estimates of MISE that is found in experiment one and two. The parameters of experiment three and four are listed in table 4.

5.2.1 Results

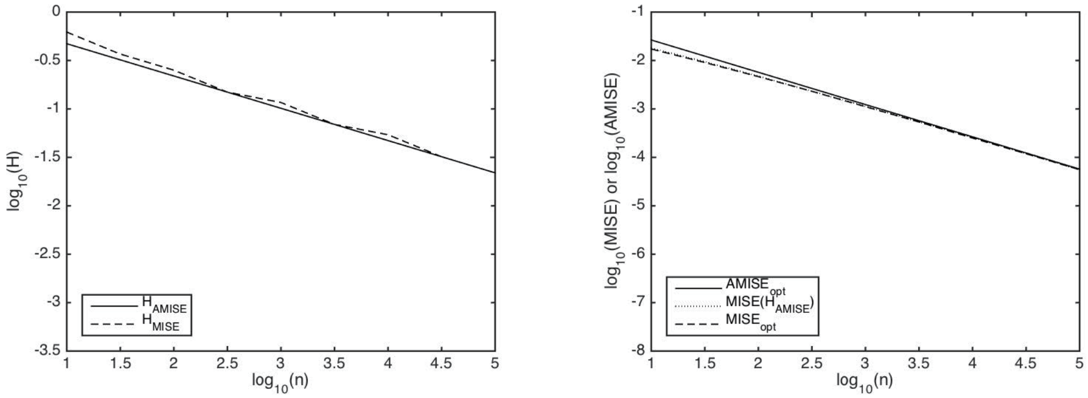
When using transmetric density estimation on multi-dimensional normal distributions, the convergence order of the mean integrated squared error, away from the mode, is approaching $4/5$ for large population sizes. This result is independent of the number of dimensions. This result complies with the asymptotic arguments in Hovda (2016) Moreover, it is shown that parameters can be trained using unbiased cross-validation. However, the convergence order is slower for the transmetric density estimatior when the number of dimensions is small.

The result of experiment three is shown in figure 3. It is clear that using UCV on the common density estimator is a method that converges for increasing n in all dimensions. The fact that the gaps between the quartiles of H_{UCV} and H_{MISE} decrease on the log scale, indicate that also the normalized experimental error decrease for increasing n. Moreover, it is seen that the expected value of UCV($\hat{f}_{\mathbf{H}}$) + $R(f)$ converges towards MISE($\hat{f}_{\mathbf{H}}$) as n increases. Notice that the normalized experimental error MISE(H_{UCV}) is decreasing as a function of n.

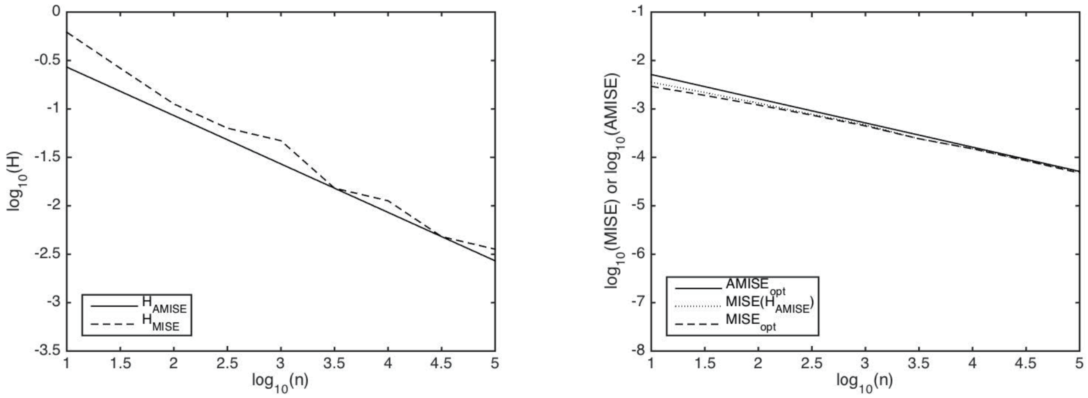
The result of experiment four is shown in figure 4. Here, UCV of the transmetric density estimator is shown. Clearly, the method seems to pick too small bandwidths, and this effect is most evident in two dimensions. This effect is increasing with n. This may not be too surprising since this method coincide with the common kernel density estimator for small n.

It is not completely clear whether the expected value of UCV($\hat{f}_{\mathbf{H}}$) + $R(f)$ converges towards MISE($\hat{f}_{\mathbf{H}}$) or not. On one side, since all graphs are decreasing everywhere on these log plots, it is clear that the experimental error is decreasing as a function of n. Moreover, in two and four dimensions the normalized experimental error seems constant, which is an important property for indicating convergence. However this is not the case in eight dimensions, which could be a sign that this will level out for very large n.

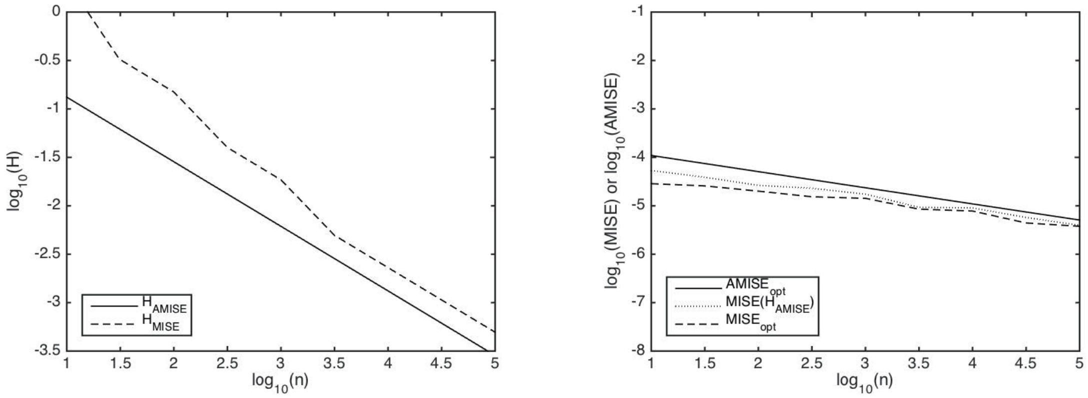
However, it is encouraging that this effect is not seen in MISE(H_{UCV}). It seems that the normalized experimental error is constant everywhere.



Two dimensions. Two dimensions.

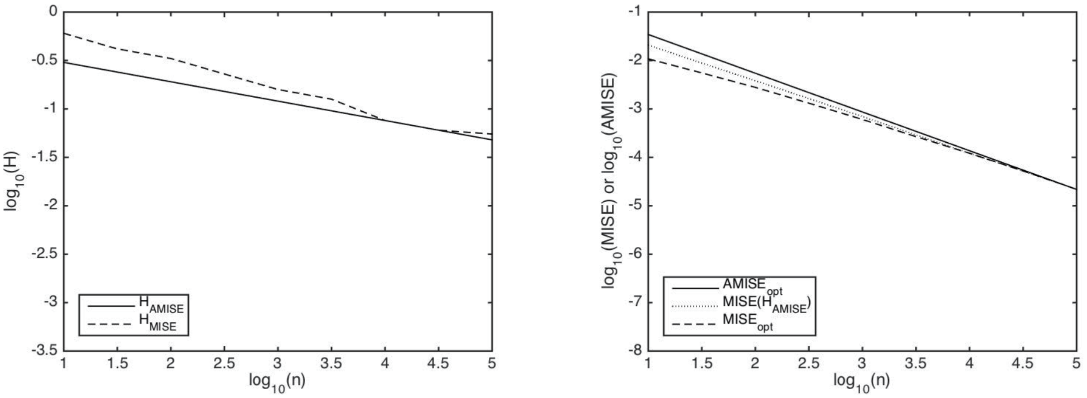


Four dimensions. Four dimensions.

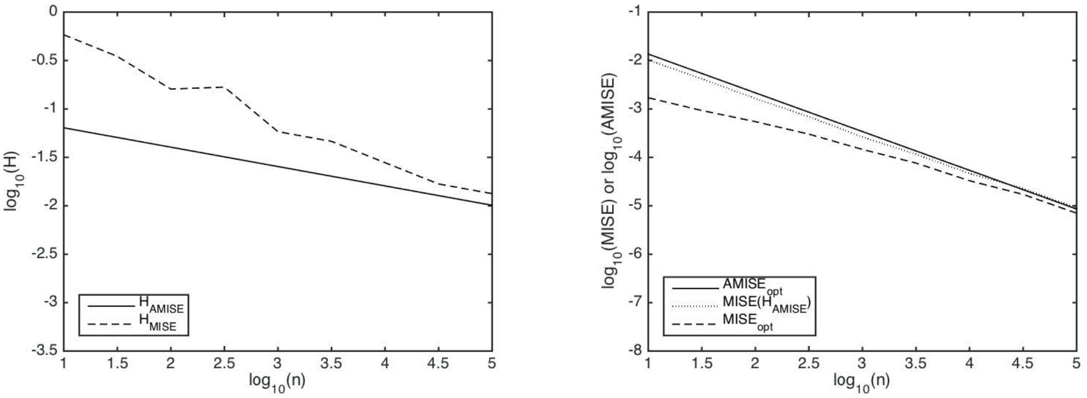


Eight dimensions. Eight dimensions.

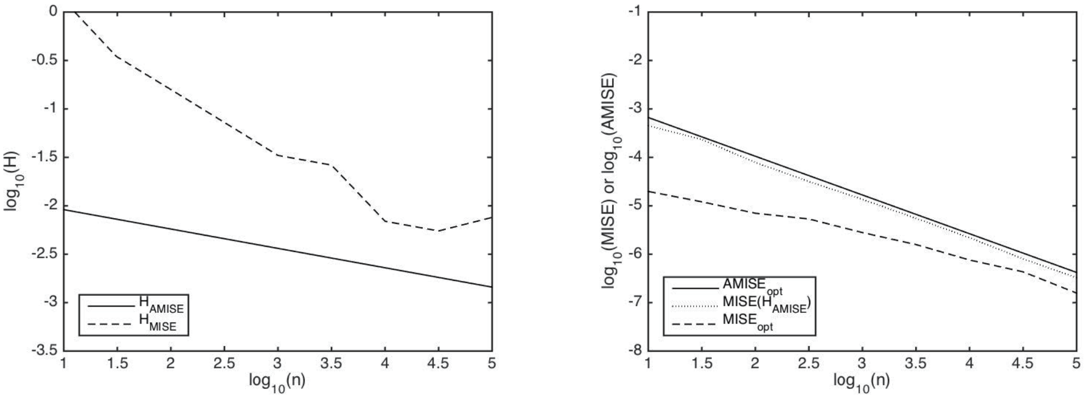
Figure 1. Result of experiment one, where the common kernel density estimator is evaluated. The left column shows log plots of optimal AMISE$_{opt}$, MISE($\mathrm{H}_{\mathrm{AMISE}}$) and MISE$_{opt}$ as functions of population size n. The distributions vary in the number of dimensions, but are all standard normal. The right column shows the corresponding optimal bandwidths in the sense of MISE and AMISE. The plots clearly show that optimal AMISE and optimal MISE decrease with approximately $n^{-4/(4+m)}$, while the optimal bandwidths are proportional with $n^{-1/(4+m)}$.



Two dimensions. Two dimensions.



Four dimensions. Four dimensions.



Eight dimensions. Eight dimensions.

Figure 2. Result of experiment two, where the transmetric density estimator is evaluated. The left column shows log plots of optimal AMISE$_{opt}$, MISE(H$_{\text{AMISE}}$) and MISE$_{opt}$ as functions of population size n. The distributions vary in the number of dimensions, but are all standard normal. The right column shows the corresponding optimal bandwidths in the sense of MISE and AMISE. The plots clearly show that optimal AMISE and optimal MISE decrease with approximately $n^{-4/5}$, while the optimal bandwidths are proportional with $n^{-1/5}$.

It is worth noting that the poorer performance of UCV on the transmetric density estimator is somewhat surprising and the question of error in the computer program is always valid. To mitigate this, the author has implemented the code in both Java and Matlab with identical results.

The right plot in figure 5, summarizes the cross-validation experiments. The ratio of the two methods is shown, where parameter estimation is part of the methods. In general, some of the gain that is achieved by using transmetrics is lost by poorer performance of the cross-validation method. In fact, in two dimensions, the common kernel density estimator is working better when parameter estimation is part of the equation. However, in more dimensions, it seems that the transmetric density estimator is superior.

6. Conclusion

A great variety of state-of-the-art problems can be defined using transmetric density estimation. This has clarified how methods relate to each other and opened up new ways on how they can be combined. Moreover, unbiased cross-validation is possible even when the distribution is not an associated distribution.

Using Monte Carlo simulations, it is shown that parameters such as the scaling of the bandwidth matrix can be estimated using unbiased cross-validation. Although, the method seems to underestimate the bandwidth in two dimensions, the method seems appropriate when the number of dimensions is higher. The experimental error of the unbiased cross-validation method seems constant with increasing population size. Moreover, Monte Carlo simulations have verified the asymptotic properties that were outlined in Hovda (2016)

Acknowledgements

The author gratefully acknowledge the university for supporting this research. Moreover, the reviewers have given valuable feedback.

Supplementary material

- Code repository at GitHub®: `https://github.com/sigveh/ipt.git`.

References

Cacoullos, T. (1966). Estimation of a Multivariate Density. *Annals of the Institute of Statistical Mathematics, 18*, 179C189. http://dx.doi.org/10.1007/BF02869528

Deheuvels, P. (1977). Estimation non parametrique de la densite par histogrammes generalises (II). *Publ. lInst. Statist. lUniv., 22*, 1C23.

Duong, T., & Hazelton, M.L. (2003). Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation. *Journal of Nonparametric Statistics, 15*, 17C30. http://dx.doi.org/10.1080/10485250306039

Duong, T., & Hazelton, M.L. (2005). Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation. *Scandinavian Journal of Statistics, 32*, 485C506. http://dx.doi.org/10.1111/j.1467-9469.2005.00445.x

Epanechnikov, V.K. (1969). Non-parametric Estimation of a Multivariate Probability Density. *Theory of Probability and its Applications, 14*, 153C158. http://dx.doi.org/10.1137/1114019

Ferraty, F.Y., & Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice, Springer.

Fix, E., & Hodges, J.L. (1951). Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties. Technical report 4, United States Air Force School of Aviation Medicine, Randolph Field, Texas.

Friedman, J.H., Stuezle, W., & Schroeder, A. (1984). Projection Pursuit Density Estimation. *Journal of the American Statistical Association, 79*, 599C608.

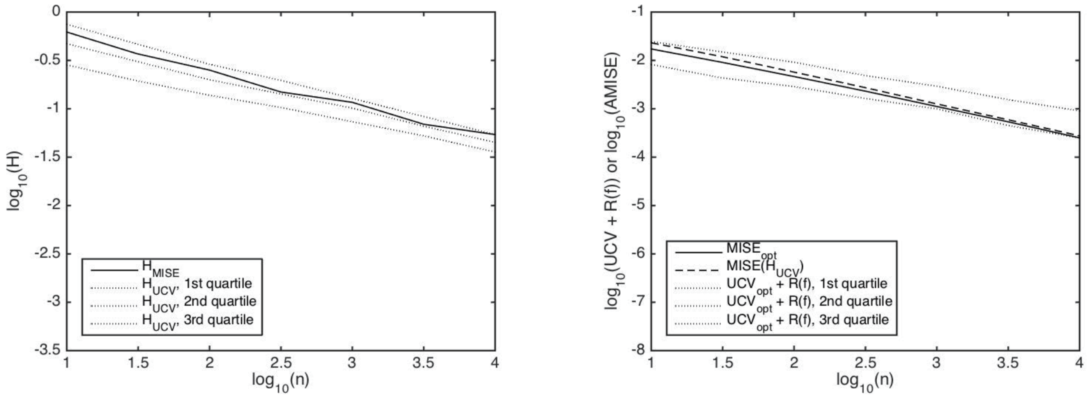
Härdle, W., Liang, H., & Gao, J. (2000). Partially Linear Models, Springer Contributions to Statistics.

Hovda, S. (2014). Using pseudometrics in kernel density estimation. *Journal of Nonparametric Statistics, 26*, 669C696. http://dx.doi.org/10.1080/10485252.2014.944524

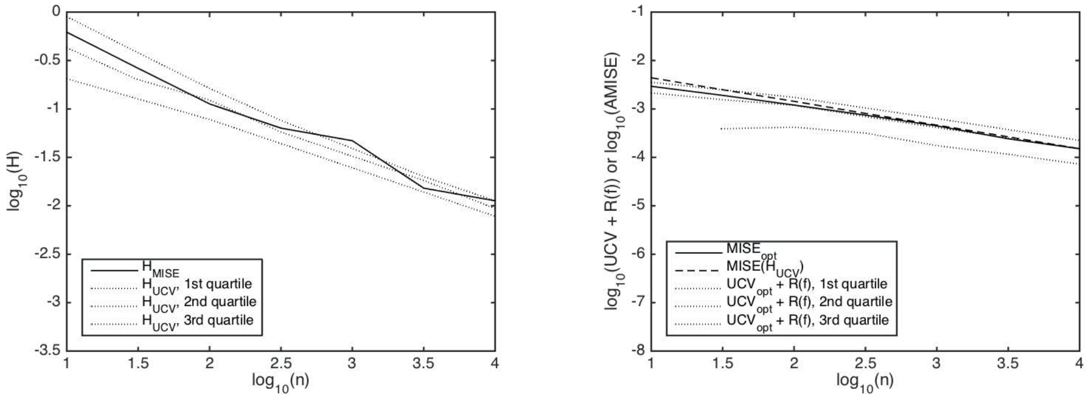
Hovda, S. (2016). Transmetric density estimation. *International Journal of Statistics and Probability, 5,* xxx. http://dx.doi.org/10.5539/ijsp.v5n2p35

Liescher, E. (2005). A Semiparametric Density Estimator Based on Elliptical Distributions. *Journal of Multivariate Analysis, 92,* 205C225.

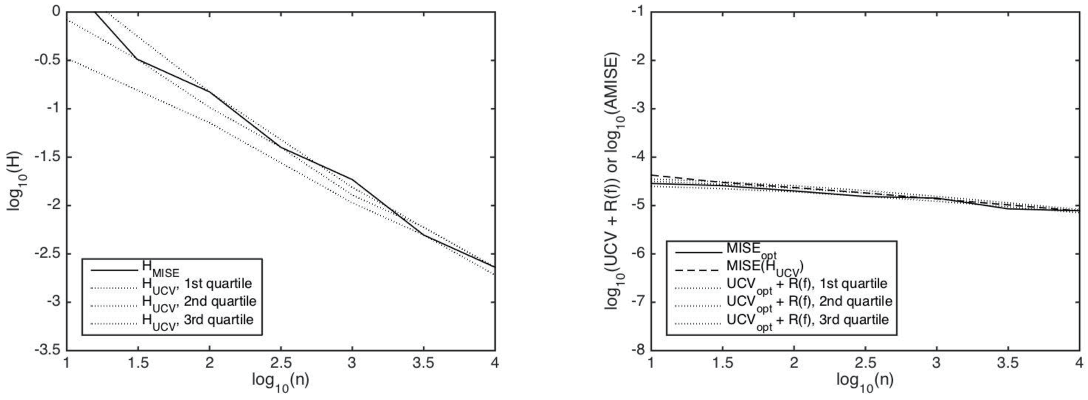
Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics, 33,* 1065C1076. http://dx.doi.org/10.1214/aoms/1177704472



Two dimensions. Two dimensions.

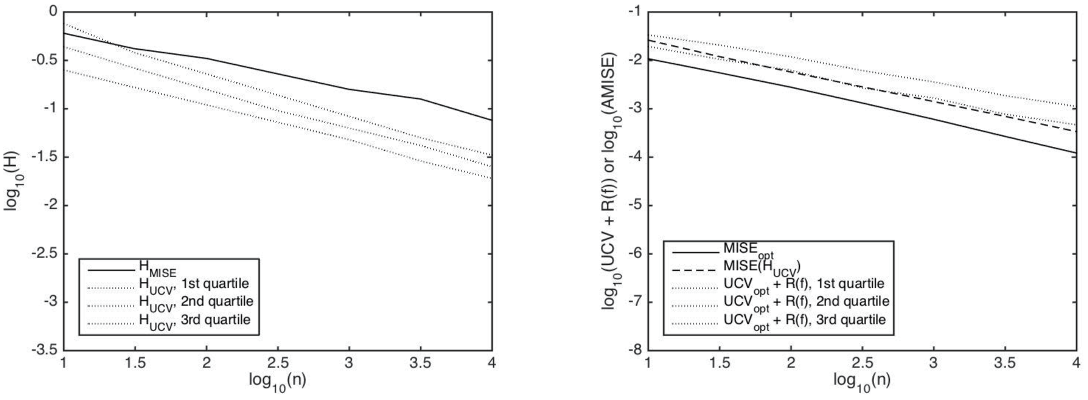


Four dimensions. Four dimensions.

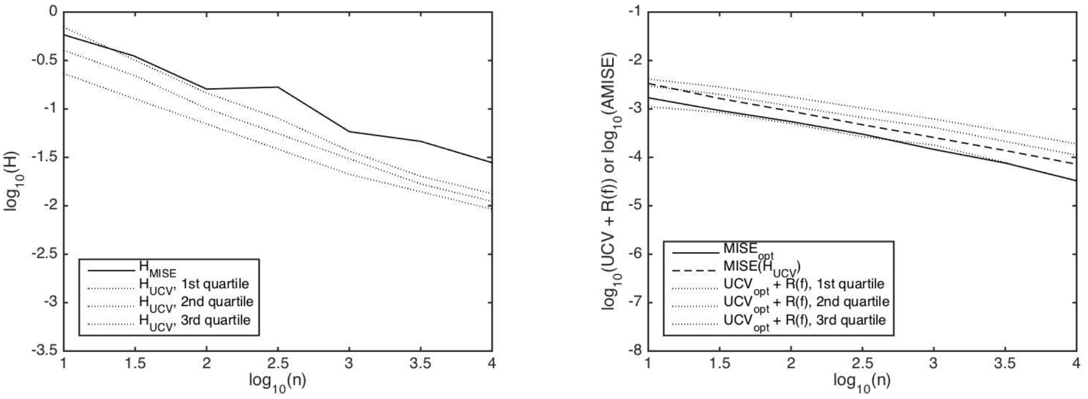


Eight dimensions. Eight dimensions.

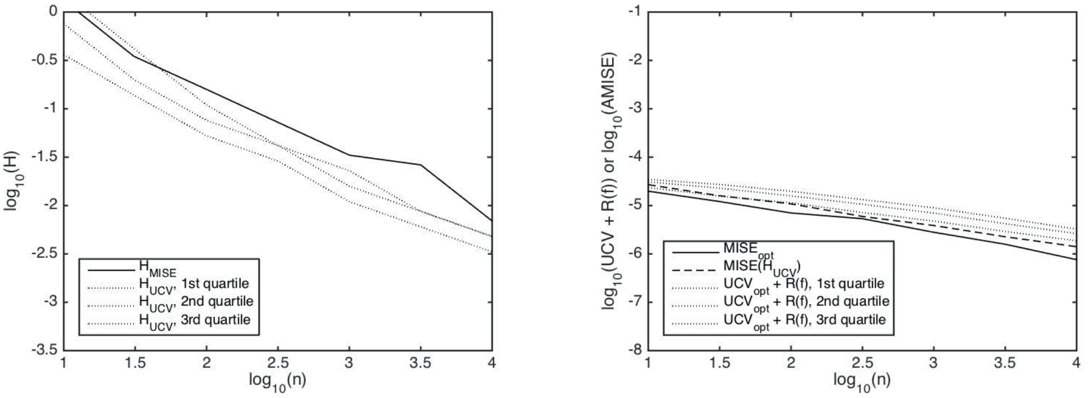
Figure 3. Result of experiment three, where the common kernel density estimator is evaluated. The left column shows log plots of quartiles of H_{UCV} together with H_{MISE} and H_{AMISE}. The next column shows the quartiles of UCV $+R(f)$ together with MISE(H_{UCV}) and MISE_{opt}.



Two dimensions. Two dimensions.

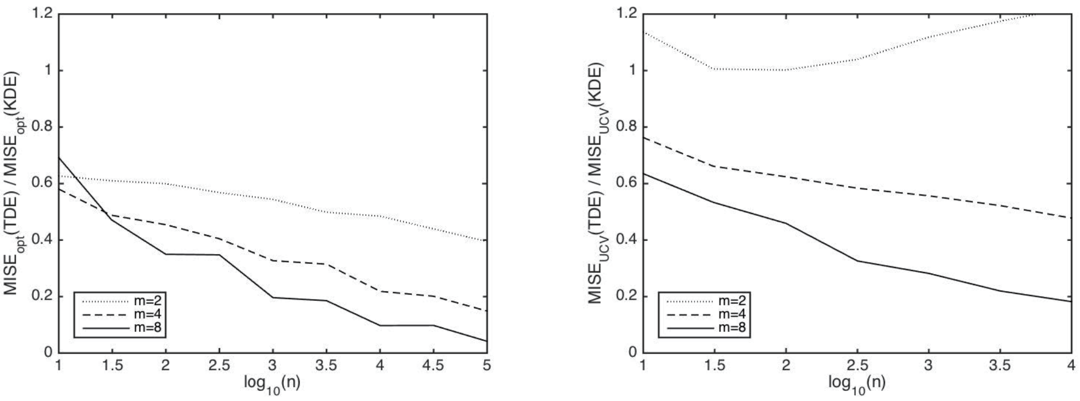


Four dimensions. Four dimensions.



Eight dimensions. Eight dimensions.

Figure 4. Result of experiment four, where the transmetric density estimator is evaluated. The left column shows log plots of quartiles of H_{UCV} together with H_{MISE} and H_{AMISE}. The next column shows the quartiles of UCV $+R(f)$ together with MISE(H_{UCV}) and $MISE_{opt}$.



Without UCV.

With UCV.

Figure 5. Comparison of transmetric density estimation and common kernel density estimators. The left figure shows a plot of MISE$_{opt}(\hat{f}_{\mathbf{H},d_t})$/ MISE$_{opt}(\hat{f}_{\mathbf{H}})$, while the right figure shows a plot of MISE$_{UCV}(\hat{f}_{\mathbf{H},d_t})$/ MISE$_{UCV}(\hat{f}_{\mathbf{H}})$ in various dimensions.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals on Mathematical Statistics, 27,* 832C837. http://dx.doi.org/10.1214/aoms/1177728190

Welling, M., Zemel, R.S., & Hinton, G.E. (2003). Efficient Parametric Projection Pursuit Density Estimation in UAI, Acapulco, Mexico, August, Morgan Kaufmann, pp. 575C582.

Copyrights