Using pseudometrics in kernel density estimation

Sigve Hovda\textsuperscript{a,b,*}

\textsuperscript{a}Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway; \textsuperscript{b}Verdande Technology, Stiklestadveien 1, 7041 Trondheim, Norway

(Received 17 October 2013; accepted 8 July 2014)

Common kernel density estimators (KDE) are generalised, which involve that assumptions on the kernel of the distribution can be given. Instead of using metrics as input to the kernels, the new estimators use parameterisable pseudometrics. In general, the volumes of the balls in pseudometric spaces are dependent on both the radius and the location of the centre. To enable constant smoothing, the volumes of the balls need to be calculated and analytical expressions are preferred for computational reasons. Two suitable parametric families of pseudometrics are identified. One of them has common KDE as special cases. In a few experiments, the proposed estimators show increased statistical power when proper assumptions are made. As a consequence, this paper describes an approach, where partial knowledge about the distribution can be used effectively. Furthermore, it is suggested that the new estimators are adequate for statistical learning algorithms such as regression and classification.

Keywords: nonparametric density estimation; semiparametric density estimation; nonparametric regression; nonparametric functional data analysis; volume of ball in pseudometric space

1. Introduction

The multivariate kernel density estimator (KDE) with a fixed bandwidth is given by

\[ \hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - x_i), \]  

where \( K_H(x) : \mathbb{R}^m \to \mathbb{R} \) is a single modal, symmetric, nonnegative and zero-mean function that integrates to 1. The bandwidth matrix \( H \) describes the level of smoothing. It is convenient to write \( H = h^{m/2} \Phi \), where \( h \) is a positive scalar and \( |\det(\Phi)| = 1 \). With this parametrisation, the shape is determined by \( \Phi \) and the scale is determined by \( h \). This gives

\[ \hat{f}_H(x) = \frac{1}{nh^{m/2}} \sum_{i=1}^{n} K(H^{-1}(x - x_i)), \]  

where \( K \) is a kernel that is independent of \( H \). The first univariate version of Equation (2) was proposed by Fix and Hodges (1951), while the general class was investigated by Rosenblatt (1956) and Parzen (1962). The multivariate extension was outlined by Cacoullos (1966) and Epanechnikov (1969).

\textsuperscript{*}Email: sigveh@idi.ntnu.no

© 2014 The Author. Published by Taylor & Francis. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The moral rights of the author have been asserted.
Several authors have suggested a variable bandwidth version of Equation (2), where the two main approaches are the balloon estimator and the sample smoothing estimator. A comparison between KDE with fixed and variable estimators is given by Terrel and Scott (1992).

The balloon estimator lets \( h(x) \) to be equal to the distance to the \( k \)th nearest neighbour of \( x \). The estimator is given by

\[
\hat{f}_H(x) = \frac{1}{nh(x)^m} \sum_{i=1}^{n} K(H(x)^{-1}(x - x_i)).
\]

The first version of this estimator was published by Loftsgaarden and Quesenberry (1965) and has later been known as the \( k \)-nearest neighbour estimator. It is given by

\[
\hat{f}(x) = \frac{k}{nV_m r^m},
\]

where \( n \) is the sample size, \( V_m \) the volume of the \( m \)-dimensional Euclidean unit ball and \( r \) the Euclidean distance between \( x \) and its \( k \)th nearest neighbour. Equation (4) can be derived from Equation (3), if we choose \( H \) to be identity and a kernel that is a uniform density on the \( m \)-dimensional Euclidean unit ball. In Loftsgaarden and Quesenberry (1965), it was also proved that the estimator in Equation (4) is consistent for increasing \( n \).

The sample smoothing estimator was suggested by Breiman, Meisel, and Purcell (1977). Here, the bandwidth parameter is independent of the query point, but dependent on the sample points. The estimator is

\[
\hat{f}_H(x) = \frac{1}{nh(x_i)^m} \sum_{i=1}^{n} K(H(x_i)^{-1}(x - x_i)),
\]

where \( h(x_i) \) is the distance to the \( k \)-nearest neighbour of \( x_i \).

To improve readability, we recollect the following definitions.

**Definition 1.1** The kernel of a function \( f : X \to Y \) is an equivalence relation \( \equiv_f \) given by \( x \equiv_f y \Leftrightarrow f(x) = f(y) \), where \( x \) and \( y \) are elements in \( X \).

**Definition 1.2** A pseudometric on a set \( X \) is a function (distance function) \( d : X \times X \to \mathbb{R} \), which for all \( x, y, z \) in \( X \) satisfies the following conditions:

1. \( d(x, y) \geq 0 \) (non-negativity).
2. \( d(x, y) = d(y, x) \) (symmetry).
3. \( d(x, z) \leq d(x, y) + d(y, z) \) (triangle inequality).

If we add the requirement of identity, which is \( d(x, y) = 0 \) if and only if \( x = y \), then \( d \) is a metric as well.

**Definition 1.3** A pseudometric space \( M \) is an ordered pair \( (X, d) \). A closed \( \epsilon \)-ball on \( M \) that is centred at \( a \) is defined as

\[
B_d(a, \epsilon) := \{ x \in X \mid d(x, a) \leq \epsilon \},
\]

which is equivalent to the definition of a closed \( \epsilon \)-ball in a metric space. Note that \( \epsilon \) is called the radius of the ball.
This definition of pseudometric is exactly the same as the definition of a semimetric in the field of nonparametric functional data analysis (Ferraty and Vieu 2006). To avoid confusion in this paper, we will keep referring to pseudometrics, even though most of the literature in nonparametric functional data analysis uses the term semimetric.

In nonparametric functional data analysis, the explanatory variables are functional and most of the work has been done on regression and classification. An overview of the field is given in the book by Ferraty and Vieu (2006). In the case of density estimation, Delaigle and Hall (2010) proved that a probability density function generally does not exist for functional data, but developed notions of density and mode, when functional data are considered in the space determined by the eigenfunctions of the principal components. In the paper by Ferraty, Kudraszow, and Vieu (2012), an infinite dimensional analogue of a density estimator is discussed, when the small ball probability associated with the functional data can be approximated as a product of two independent functions; one depending on the centre of the ball and one depending on the radius.

In this paper, we consider only finite dimensional pseudometric spaces, where balls with finite radius have finite volumes. Such spaces exist as a specification of the theory of nonparametric functional data analysis (Ferraty et al. 2012). The theory in this paper is distinguished from that of nonparametric functional data analysis, because we consider pseudometric spaces, where volumes of balls with fixed radius are allowed to be dependent on locations of centres. In all previous work on using pseudometrics for functional data, the volumes of the balls are always considered as independent on the locations of the centres. This generalisation opens up for applications that are less linked to functional data and more related to restricted nonparametric and semiparametric estimators on real (non-functional) data.

To be more accurate, the parameters of the pseudometric precisely describe the kernel, or more roughly the level sets of the distribution. This is important for visualisation and interpretation of the estimated distribution. For instance, the complete multi-dimensional density that is estimated with an estimator using a pseudometric of family one (described later in this paper) can be described by a two-dimensional plot only.

The choice of the family of pseudometrics describes the constraints on the level sets of the distributions that can be modelled. In case of fitting parameters, it is therefore important that the level sets of the actual distribution are not in violation with these constraints. This is somewhat analogous to parametric models, where serious misbehaviour can be observed, if the distribution has a different shape than any of the members in the family of models.

An example of a similar approach can be found in a paper by Liescher (2005), where the level sets of the distributions were constrained to be of elliptic shape. The elliptic shape of the level sets was constrained by a transformation that was performed prior to applying the nonparametric density estimator. It was shown that the convergence rates were independent of the number of dimensions, except in the neighbourhood of the mean, and thereby overcoming the so-called curse of dimensionality. The estimators in this paper can be seen as a generalisation of the estimators of Liescher (2005), because a wider family of associated distributions is allowed. Convergence rates are not explicitly given in this paper, but simulations show that the performance close to the mode coincides with the common KDE and improves quickly away from the mode, which is similar to the results obtained by Liescher (2005).

In the context of reducing the curse of dimensionality of nonparametric density estimators, it is also worth mentioning the projection pursuit density estimator. 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. A special case is the flexible Bayesian estimator, where the projection matrix is identity. In essence, the flexible Bayesian method is assuming independence between the explanatory variables. The estimator will be used later in this paper and
S. Hovda

is plug-in

\[ \hat{f}_h(x) = \frac{1}{n^m h^m} \prod_{j=1}^{m} \sum_{i=1}^{n} K \left( \frac{x_j - x_{ij}}{h} \right), \]

where \( K \) is a univariate kernel, the bandwidth is \( h \), \( x_j \) is the \( j \)th coordinate of \( x \) and \( x_{ij} \) is the \( j \)th coordinate in sample \( i \). Consistency is proved under the assumption of independence (John and Langley 1995). This is a nonparametric variation of the old and well-known naive-Bayesian estimator, where the individual distributions are assumed Gaussian. A similarity between the estimators in this paper and the projection pursuit density estimators is that the resulting density can be visualised in lower dimensional spaces. In projection pursuit, it is enough to show the one-dimensional (ridge) functions along with the principal directions to understand the full distribution.

The proposed estimators have less in common with semiparametric models, where part of the model is fully parametric. For instance models which are a product of a parametric and a nonparametric component, such as in Naito (2004) and references therein. Another example is when some of the explanatory variables are modelled nonparametrically, while the rest are modelled parametrically, such as in Hoti and Holmström (2003).

In the next section, the underlying assumptions of the KDE with fixed and variable bandwidths are discussed. This is followed by formal definitions of the new estimators. A framework for finding suitable parametric families of pseudometrics is outlined in Section 3 and two families are identified. Some experiments are given in Section 4, where the new KDE are compared with common KDE on samples drawn from common distributions. In Section 5, open questions and ideas of how to use the new estimators on various problems are elaborated on, before the work is concluded.

2. Assumptions for kernel density estimation and definition of new estimators

It is convenient to start our discussion by investigating the \( k \)-nearest neighbour method, because it is conceptually easier than the other KDE. If we generalise Equation (4) for using any metric \( d \), then

\[ \hat{f}(x) = \frac{k}{n V_{B_d(x, d(x, x_k))}}, \]

where \( x_k \) is the \( k \)th nearest neighbour of \( x \) and \( V_{B_d(x, d(x, x_k))} \) is the volume of the ball in the metric space \( (\mathbb{R}^m, d) \) with radius \( d(x, x_k) \).

For the sake of discussion, a random variable \( Y \) is defined as a \( k \)-nearest neighbour estimate at a particular \( x \). Each \( k \)-nearest neighbour estimate requires drawing a new training sample and this is why \( Y \) is a random variable. The variance of \( Y \) is a consequence of which sample is chosen (because this determines the location of the \( k \)th-nearest neighbour), while a great portion of the bias can be explained by the shape of the ball.

This is potentially more clear if we realise that Equation (5) can be derived from two assumptions that were not stated in Loftsgaarden and Quesenberry (1965). The first assumption is that

\[ \int_{B_d(x, d(x, x_k))} \hat{f}(u) \, du = \frac{k}{n}. \]

The second assumption is that \( \hat{f}(x) \) is equal to the mean of \( \hat{f}(\cdot) \) inside \( B_d(x, d(x, x_k)) \), because the left-hand side of Equation (6) is equal to \( V_{B_d(x, d(x, x_k))} \hat{f}(x) \), which is how Equation (6) implies Equations (5) and (4). A measure for how much the second assumption is violated is identified as
\[ AV = \int_{\mathbb{R}^n} \left| \int_{\mathbb{R}^n} f(u) \frac{d u}{V_{\mathbb{R}^n}[x,d(x,x_k)]} - f(x) \right| d x. \]  

The parameter $AV$ is therefore highly dependent on which metric is chosen, because the metric determines the shape and the size of the region of integration. A great portion of the bias of $Y$ is a consequence of the second assumption. Also, we recognise that there is a connection between $f$ and the choice of metric. No metric is optimal for any $f$ and vice versa.

This argumentation can be transferred to all KDE with both fixed and variable bandwidths, because these methods are essentially weighting the contribution of each sample $x_i$ as a function of $V_{\mathbb{R}^n}[x,d(x,x_i)]$. For the KDE with fixed bandwidth, the balloon estimator and the sample smoothing estimator, respectively, we could write

\[ \hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K'_H(V_{\mathbb{R}^n}[x,d(x,x_i)]), \]

\[ \hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K'_{H(x)}(V_{\mathbb{R}^n}[x,d(x,x_i)]), \]

\[ \hat{f}_H(x) = \frac{1}{n} \sum_{i=1}^{n} K'_{H(x_i)}(V_{\mathbb{R}^n}[x,d(x,x_i)]), \]

where $K'_H$, $K'_{H(x)}$ and $K'_{H(x_i)}$ are nonnegative, strictly decreasing, one-sided univariate kernels that integrate to one, when integrated over $x$. If we let the random variable $Y$ be defined as any of the kernel density estimates at a particular $x$, then the bias of $Y$ is highly related to the shape of the balls in combination with the shape of the $f$ near $x$.

2.1. Definition of the $k$-nearest pseudo-neighbour estimator

A generalisation of the $k$-nearest neighbour estimator is proposed as

\[ \hat{f}(x) = \frac{k}{nV_{\mathbb{R}^n}[x,d_p(x,x_k)]}, \]

where $d_p$ is a pseudometric. Figure 1 illustrates the principles of $k$-nearest neighbour estimator as of Loftsgaarden and Quesenberry (1965) together with the $k$-nearest pseudo-neighbour estimator. This particular pseudometric is a special case of the parametric families of pseudometrics that are outlined later in this paper.

It is postulated that a proper choice of a pseudometric can reduce Equation (7) far more than a metric, because balls in pseudometric spaces span a richer variety of shapes. The key idea is to find a parametric family of pseudometrics and select parameters by either Bayesian inference or learning.

It is a crucial point that the parameters of a pseudometric family are typically associated with various sample statistics and associations between the explanatory variables. In the example shown in Figure 1, the centre of the dashed circles is a parameter of the pseudometric. This parameter can be estimated by the sample mean.

2.2. Definition of pseudometric KDE

A challenge with using pseudometrics is that the volumes of the balls with constant radius are generally dependent on the locations of the centres. This effect is neutralised by defining the
Figure 1. Description of the $k$-nearest pseudo-neighbour estimator. The scatter plot shows a training population with 100 points that are i.i.d and drawn from a two-dimensional normal distribution. A cross is marked at a query point and in $k$-nearest neighbour estimation, the $k$th-nearest neighbour is found by comparing the Euclidean distance to all other points. The smallest Euclidean ball that encapsulates the 5th-nearest neighbours is shown as a solid circle. Clearly, $f$ is not very constant inside this ball. In $k$-nearest pseudo-neighbour estimation, the $k$th-nearest neighbour is found by comparing the distance in the pseudometric space from the query point to all other points. The interior between the two dashed lines represents a ball in this pseudometric space, where the query point is in the centre. The radius of the ball (not of the dashed circles) is the distance to the 5th-pseudo-nearest neighbour. Obviously, $f$ is far more constant inside this ball. Consequently, improved statistical power is achieved when the equivalence classes with respect to the kernel of the distribution are assumed to be circles around the mean.

**equivalent Euclidean distance**, which is the function

$$
ed_{d_p}(x, y) = \left( \frac{V_{B_{d_p}(x, y)}}{V_m} \right)^{1/m},$$

where $V_m$ is the volume of the Euclidean unit $m$-ball. The equivalent Euclidean distance from $x$ to $y$ in a pseudometric space $(\mathbb{R}^m, d_p)$ is the radius of the Euclidean $m$-ball that has the same volume as $V_{B_{d_p}(x, y)}$. Note that $e_{d_p}$ is generally not a metric, nor a pseudometric, because the symmetry criteria are broken.
The idea is to weight the contributions of the \( x_i \)'s according to the equivalent Euclidean distances from the query point. This ensures that the level of smoothing is constant with respect to volume, which is the case for KDE with fixed bandwidths.

The pseudometric kernel density estimator (PKDE) is suggested as

\[
\hat{f}_{h,d_p}(x) = \int_0^{\infty} K_h(e_{d_p}(x, y)) \frac{dF}{de_{d_p}} \, de_{d_p} = \frac{1}{n} \sum_{i=1}^{n} K_h(e_{d_p}(x, x_i)), \tag{9}
\]

where \( K_h(u) \) is the kernel with bandwidth \( h \) and \( dF/de_{d_p} \) is equal to \( 1/n \) at the \( e_{d_p}(x, x_i)s \) and 0 elsewhere. An example of a kernel is the one-dimensional normal, which is given by

\[
K_h(u) = \frac{1}{h^n} K \left( \frac{u}{h} \right), \quad \text{where } K(u) = \frac{1}{(2\pi)^{m/2}} \exp \left( -\frac{1}{2} u^2 \right).
\]

This is realised because \( (d(V_m u^m)/du)K_h(u) \) integrates to 1, when integrated over \( u \). Notice that the kernel itself is not dependent on the choice of pseudometric. Only the number of dimensions is necessary input.

If we choose \( d_p(x, y) = \| \Phi^{-1}(x - y) \| \) where \( \det(\Phi) = 1 \) and a bandwidth \( h \), then Equation (1) becomes a special case of Equation (9), when \( H = h^m \Phi \) and normal kernels are used. The only difference is that the kernel in Equation (9) takes a scalar argument, while the kernel in Equation (1) takes a vector. A similar argumentation can be made for any kernel and this is why the PKDE is a generalisation of the standard KDE.

In the case of variable bandwidths, the pseudometric balloon estimator is suggested as

\[
\hat{f}_{h,d_p}(x) = \frac{1}{n h(x)^m} \sum_{i=1}^{n} K \left( \frac{e_{d_p}(x, x_i)}{h(x)} \right),
\]

where \( h(x) = e_{d_p}(x, k) \) and \( k \) is the \( k \)-nearest pseudo-neighbour of \( x \) (nearest in the sense \( d_p \)). Notice that if we choose \( d_p \) to be Euclidean and a kernel that has uniform density on the Euclidean ball, the \( k \)-nearest neighbour method falls out as a special case.

Moreover, the pseudometric sample smoothing estimator is suggested as

\[
\hat{f}_{h,d_p}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x_i)^m} K \left( \frac{e_{d_p}(x, x_i)}{h(x_i)} \right),
\]

where \( h(x_i) = e_{d_p}(x_i, k_i) \) and \( k_i \) is the \( k \)-nearest pseudo-neighbour of \( x_i \).

### 3. Choosing appropriate parametric families of pseudometrics

A clear challenge with using the new estimators is proposing suitable families of pseudometrics. It is therefore instructive to define the equivalence relation \( \sim_{d_p} \), where \( x \sim_{d_p} y \) if and only if \( d_p(x, y) = 0 \). While discussing the usefulness of a family of pseudometrics, it makes sense to ask which density functions are invariant to \( \sim_{d_p} \), and the next two definitions are motivated by this.

**Definition 3.1** A probability density function is said to be an associated distribution of a pseudometric (or metric) \( d_p \), denoted as \( f_{d_p} : X \rightarrow \mathbb{R}^+ \), when \( f_{\sim_{d_p}} \) is invariant to \( \sim_{d_p} \). Invariance to \( \sim_{d_p} \) means that \( x \sim_{d_p} y \), implies that \( f_{\sim_{d_p}}(x) = f_{\sim_{d_p}}(y) \) for all \( x, y \) in \( X \).

**Definition 3.2** An associated distribution \( f_{d_p} \) satisfies the strong condition if \( \sim_{d_p} \) is equal to \( =_{d_p} \), where \( =_{d_p} \) is the kernel of \( f_{\sim_{d_p}} \). This is the case when \( x \sim_{d_p} y \Leftrightarrow f_{\sim_{d_p}}(x) = f_{\sim_{d_p}}(y) \) for all \( x, y \) in \( X \).


Table 1. List of important symbols.

| Symbol | Explanation and properties |
|--------|----------------------------|
| \(m\)  | Number of dimensions       |
| \(m_1, m_2\) | Used in pseudometric \(dp\), where \(m_1\) and \(m_2\) are nonzero integers and \(m_1 + m_2 = m\) |
| \(A\)  | Linear transformation matrix of size \(m \times m\), where \(\det(A) \neq 0\) |
| \(c\)  | Centre point parameter (not centre of the balls). Size is either \(m\) or \(m_1\) depending on the context |
| \(p\)  | Vector of positive exponents in pseudometrics. Size is either \(m\) or \(m_1\) depending on the context |
| \(p\)  | Harmonic average of the elements in \(p\) |
| \(q\)  | Exponent that is used in pseudometric \(dp\) |
| \(I\)  | Identity matrix with size \(m \times m\) or \(m_1 \times m_1\) |
| \(0\)  | Null vector with size \(m, m_1\) or \(m_2\) |
| \(a\)  | Centre point of the balls. Length is \(m\) |
| \(\epsilon\) | Radius of the balls. \(\epsilon > 0\) |
| \(V_p\) | Volume of the generalised unit ball with parameter \(p\) |

3.1. Volumes of generalised balls

Two new families of pseudometrics are proposed, including the associated distributions and the derivations of the analytical expressions for the volumes of their balls. Before this, a definition of the closed linearly transformed generalised balls and a derivation of an analytical expression for the volumes are needed. Table 1 contains a list of symbols that are consistently used in rest of this paper.

**Definition 3.3** The closed linearly transformed generalised \(\epsilon\)-ball that is centred in \(a\) is denoted \(\mathbb{B}_{d_{sA,p}}[a, \epsilon]\), where

\[
d_{sA,p}(x, y) = \left( \sum_{i=1}^{m} |x_i' - y_i'|^{p_i} \right)^{1/p}.
\]

Here, \(p\) is a vector of the positive \(p_i\)s, while \(p\) is the harmonic average of the \(p_i\)s (i.e. \(p = (1/m \sum_{i=1}^{m} p_i^{-1})^{-1}\)). Moreover, \(x' = Ax\) and \(y' = Ay\), where \(A\) is a square transformation matrix with nonzero determinant.

In general \(d_{sA,p}\) is a semimetric, because the triangle inequality does not hold when at least one of the \(p_i\)s are less than 1. Depending on the \(p_i\)s, the generalised balls describe a wide range of multi-dimensional geometrical objects, where spheres, cubes, cylinders and stars are special cases. A few three-dimensional objects are shown in Figure 2.

In articles by Wang (2005) and Gao (2013), the volume of the generalised unit ball was derived.

**Theorem 3.4** The volume of the closed generalised unit ball \(V_p\) is

\[
V_p = V_{\mathbb{B}_{d_{sA,p}}[0,1]} = 2^m \prod_{i=1}^{m} \frac{\Gamma(1 + 1/p_i)}{\Gamma(1 + m/p)} ,
\]

where \(I\) is the identity matrix.

The proof by Wang used induction and was motivated by an idea from Folland (2001), while the proof by Gao used properties of the exponential distribution. An alternative proof of Equation (10) is included in the appendix.
Theorem 3.5  The volume of the closed linearly transformed generalised $\epsilon$-ball that is centred in $a$ is

$$V_{\mathbb{B}_{d,A_p}[a,\epsilon]} = \frac{1}{|\det(A)|} V_p \epsilon^m.$$
Figure 3. Balls in the pseudometric spaces \((\mathbb{R}^3, d_{p,1,0,p})\). One octant is removed in both figures, so it is possible to see more of the boundaries of the balls. (a) \(B_{d_{p,1,0,1}}(\{1,0,0\}, 0.1)\) and (b) \(B_{d_{p,1,0,2}}(\{1,0,0\}, 0.1)\).

This means that \(V_p\) and the absolute value of the determinant of \(A\) determine the overall scaling of the volume. This is a consequence of choosing the exponent in \(d_{s,A,p}\) to be \(1/p\).

3.2. Pseudometric family of type one

**Definition 3.6** The pseudometric family of type one \(d_{p,A,c,p}(x,y) : \mathbb{R}^m \rightarrow \mathbb{R}\) is given by

\[
d_{p,A,c,p}(x,y) = |d_{s,A,p}(x,c) - d_{s,A,p}(y,c)|,
\]

where \(c\) is a chosen point (for instance a mean or a median).

The parameters that can be trained for this pseudometric are \(A, c\) and \(p\). In the family of pseudometric spaces \((\mathbb{R}^3, d_{p,A,c,p})\), the balls can be viewed as generalisations of the spherical shell and two examples are shown in Figure 3.

3.2.1. Associated distributions

**Theorem 3.7** The family of associated distributions \(f_{\sim d_{p,A,c,p}}(x)\) is equal to a function \(g(u)\), which satisfies

\[
\frac{V_p m}{|\det(A)| p} \int_0^{\infty} u^{(m/p)1} g(u) \, du = 1,
\]

(11)

where \(u = d_{s,A,p}(x,c)^p\). Moreover, \(f_{\sim d_{p,A,c,p}}(x)\) satisfies the strong condition.

If we choose \(g(u)\) to be an exponential function \(a \exp(-u)\), then the coefficient \(a\) can be found from Equation (11). When inserting for \(g(u)\), the integral part of Equation (11) becomes \(a\Gamma(m/p)\),
Figure 4. The univariate generalised normal distribution is a parametric family of symmetric unimodal distributions that includes all normal and Laplace distributions, and as limiting cases it includes all continuous uniform distributions on bounded intervals of the real line. It includes a shape factor $\beta$ and a scale factor $\alpha$. Various choices of shape factors are shown in this figure, where the scale factor is 1 and the mean is 0.

so that

$$a = \frac{|\det(A)|}{V_p \Gamma(1 + m/p)} = \frac{|\det(A)|}{2^m \prod_{i=1}^{m} \Gamma(1 + 1/p_i)}.$$  

The parametric family of associated distributions is therefore a linear transformation of the $m$-dimensional generalised normal distribution (also known as the exponential power distribution)

$$f_{-\psi_{1,\alpha\beta}}(x) = \frac{|\det(A)|}{2^m \prod_{i=1}^{m} \Gamma(1 + 1/p_i)} \exp \left(-\sum_{i=1}^{m} |z_i|^{p_i} \right).$$

To interpret this result, a description of the univariate case is given by

$$f(x) = \frac{\beta}{2\alpha \Gamma(1/\beta)} \exp \left(-\left(\frac{|x - \mu|}{\alpha}\right)^{\beta} \right),$$

where more details are given in Figure 4. Also, if we choose $A$ to be diagonal with $1/\alpha_i$s on the diagonal then

$$f_{-\psi_{1,\alpha\beta}}(x) = \prod_{i=1}^{m} \frac{p_i}{2\alpha_i \Gamma(1/p_i)} \exp \left(-\left(\frac{|x_i - \mu_i|}{\alpha_i}\right)^{p_i} \right).$$

In this case all dimensions are independent and the $\alpha_i$s are the scale factors and $p_i$s are the shape factors.

3.2.2. Volumes of the balls

**Theorem 3.8** The volumes of the balls in the family of pseudometric spaces $(\mathbb{R}^m, d_{\psi_{1,\alpha\beta}})$ are given by

$$V_{B_{\psi_{1,\alpha\beta}}[a,\epsilon]} = \begin{cases} 
2V_p \sum_{i=0}^{m} \binom{m}{i} d_{s,\alpha\beta}(a,c)^{m-i} \epsilon^i & \text{if } \epsilon < d_{s,\alpha\beta}(a,c), \\
V_p \sum_{i=0}^{m} \binom{m}{i} d_{s,\alpha\beta}(a,c)^{m-i} \epsilon^i & \text{if } \epsilon \geq d_{s,\alpha\beta}(a,c), 
\end{cases}$$
where \( O \) is the set of positive odd numbers that are smaller than or equal to \( m \).

The continuity of the ball at \( \epsilon = d_{s,A,p}(a,c) \) is seen, when noting the fact that \( \sum_{i=0}^{m} \binom{m}{i} = 2^m \) and \( \sum_{i=0}^{O(\epsilon)} \binom{m}{i} = 2^{m-1} \). In the case when \( \epsilon < d_{s,A,p}(a,c) \), the volume is a polynomial with respect to either \( \epsilon \) or \( d_{s,A,p}(a,c) \). The same is true when \( \epsilon \geq d_{s,A,p}(a,c) \), but the polynomial is different. This demonstrates how the volumes of the balls are dependent on the locations of the centres, and this is why it was necessary to define the equivalent Euclidean distance in Equation (8).

### 3.3. Pseudometric family of type two

**Definition 3.9** The pseudometric family of type two \( d_{p2,A,c,p,q}(x,y) : \mathbb{R}^m \to \mathbb{R} \) is proposed as

\[
d_{p2,A,c,p,q}(x,y) = \left( \left( d_{p1,1,c'}(x_{1:m_1}, y_{1:m_1}) \right)^q + \sum_{i=m_1+1}^{m} |x_i - y_i|^q \right)^{1/q},
\]

where \( q \) is an integer larger than or equal to one, \( I \) is identity, \( x' = Ax \), \( y' = Ay \) and \( c' = A_{1:m_1,1:m_1}c \). The notations \( x_{1:b} \) means \( \{x_1, x_2, \ldots, x_b\} \) and \( A_{1:m_1,1:m_1} \) is the upper left sub-matrix of \( A \) with size \( m_1 \times m_1 \). Also we have \( m_1 + m_2 = m \).

The parameters that can be trained here is \( q, A, c \) and \( p \), but also which \( x_i, s \) are put into \( d_{p1,1,c'} \) (it does not have to be the \( m_1 \) first). In the limiting case when \( q \to \infty \)

\[
\lim_{q \to \infty} d_{p2,A,c,p,q}(x,y) = \max\{d_{p1,1,c'}(x_{1:m_1}, y_{1:m_1}), |x_{m_1+1} - y_{m_1+1}|, \ldots, |x_m - y_m|\}.
\]

The ball is as usual defined by

\[
\mathbb{B}_{d_{p2,A,c,p,q}}(a, \epsilon) := \{x \in X \mid d_{p2,A,c,p,q}(x,a) \leq \epsilon\}.
\]

If \( m_1 \) is less or equal to 1, \( \mathbb{B}_{d_{p2,A,c,p,q}}(a, \epsilon) \) is the familiar \( m \)-dimensional ball. The familiar torus and a deformation of a torus are other special cases that are shown in Figure 5.

#### 3.3.1. Associated distributions

In terms of discussing the associated distributions for \( d_{p2,A,c,p,q} \), it is an important observation that the equivalence classes of \( \mathbb{R}^m/\sim_{d_{p2,A,c,p,q}} \) are all topological manifolds of dimension \( m_1 - 1 \), while the equivalence classes of \( \mathbb{R}^m/\sim_{d_{p2,A,c,p,q}} \) are \( (m-1) \)-manifolds. This means that \( \mathbb{R}^m/\sim_{d_{p2,A,c,p,q}} \) can never be equal to \( \mathbb{R}^m/\sim_{d_{p2,A,c,p,q}} \), unless \( m_1 \) is equal to \( m \).

This dimension reduction may be desirable with regards to improving robustness, because the family of associated distributions spans wider. On the other side, it may reduce statistical power because AV defined in Equation (7) may not be minimised properly.

**Theorem 3.10** The parametric family of associated distributions \( f_{\sim_{d_{p2,A,c,p,q}}} \) is equal to a function \( g(u, x'_{1:m_1+1}) \), which satisfies

\[
\frac{V_{p/m_1}}{|\text{det}(A)|^p} \int_{\mathbb{R}^m} \int_{0}^{\infty} u^{(m_1/p-1)} g(u, x'_{1:m_1+1}) du \, dx_{m_1+1} \, dx_{m_1+2} \ldots \, dx_m = 1,
\]

where \( u = d_{s,1,p}(x'_{1:m_1}, c')^p \).
If we also assume that the \( x'_1, x'_2, \ldots, x'_m \) are all independent of \( x'_{m_1+1}, x'_{m_1+2}, \ldots, x'_m \) (and vice versa), then \( g(u, x'_{m_1+1:m}) = g_1(u)g_2(x'_{m_1+1:m}) \). Hence, \( g_1(u) \) is subject to the constraint

\[
\frac{V_{p_1} m_1}{|\det(A)|^p} \int_0^\infty u^{(m/p-1)} g_1(u) \, du = 1,
\]

while \( g_2(x'_{m_1+1:m}) \) is a probability density function that integrates to 1, when integrated over \( x'_{m_1+1:m} \). Clearly, the family of associated distributions \( f_{\sim d_{p,2}\mathcal{A}_p \cdot q}(\mathbf{x}) \) has the family of associated distributions \( f_{\sim d_{p,1}\mathcal{A}_p}(\mathbf{x}) \) and the family of all distributions as special cases.

### 3.3.2. Volumes of the balls

**Theorem 3.11** The volumes of the balls in the family of pseudometric spaces \((\mathbb{R}^m, d_{p,2}\mathcal{A}_p \cdot q)\) are given by

\[
V_{d_{p,2}\mathcal{A}_p \cdot q}[a, c] = V_p \left( \frac{2(1+1/q)\Gamma(m_2/q)}{|\det(A)| \Gamma(m_2/q)} \right) \left( \sum_{i=0}^{m_1} \left( m_1 \atop i \right) B_1 \left( \frac{m_2}{q}, \frac{i}{q} + 1 \right) d_{s,1,p}(a'_{l,m_1}, c')^{m_1-i} \epsilon^{m_2+i} \right) \\
+ 2 \sum_{i \in O} \left( m_1 \atop i \right) \left( B_1 \left( \frac{m_2}{q}, \frac{i}{q} + 1 \right) - B_1 \left( \frac{m_2}{q}, \frac{i}{q} + 1 \right) \right) d_{s,1,p}(a'_{l,m_1}, c')^{m_1-i} \epsilon^{m_2+i},
\]

where

\[
s = \begin{cases} 
0 & \text{if } \epsilon < d_{s,1,p}(a'_{l,m_1}, c'), \\
1 - \left( \frac{d_{s,1,p}(a'_{l,m_1}, c')}{\epsilon} \right)^q & \text{if } \epsilon \geq d_{s,1,p}(a'_{l,m_1}, c').
\end{cases}
\]

Here, \( B_1(a, b) \) is the incomplete beta function, \( a' = \mathbf{A}a \) and \( O \) is the set of positive odd numbers that are smaller than or equal to \( m_1 \).
In the case when $\epsilon < d_{s,1,p}(a'_{1,m_1}, c')$, the volume is a polynomial with respect to either $\epsilon$ or $d_{s,1,p}(a'_{1,m_1}, c')$. The same is true when $\epsilon \geq d_{s,1,p}(a'_{1,m_1}, c')$, but the polynomial is different.

**Theorem 3.12** The volumes of the balls in the family of pseudometric spaces $(\mathbb{R}^m, d_{p,2,\mathcal{A},p,q})$ in the limiting case when $q \to \infty$ are given by

$$
\lim_{q \to \infty} V_{B_{p,2,\mathcal{A},p,q}[a,x]} = \begin{cases} 
\frac{2^{m_1+1} V_p}{|\det(A)|} \sum_{i=0}^{m_1} \binom{m_1}{i} d_{s,1,p}(a'_{1,m_1}, c')^{m_1-i} \epsilon^{m_1+i} & \text{if } \epsilon < d_{s,1,p}(a'_{1,m_1}, c'), \\
\frac{2^m V_p}{|\det(A)|} \sum_{i=0}^{m_1} \binom{m_1}{i} d_{s,1,p}(a'_{1,m_1}, c')^{m_1-i} \epsilon^{m_1+i} & \text{if } \epsilon \geq d_{s,1,p}(a'_{1,m_1}, c'), 
\end{cases}
$$

where $O$ is the set of positive odd numbers that are smaller than or equal to $m_1$.

The continuity of the ball at $x = d_{s,1,p}(a'_{1,m_1}, c')$ is seen, when noting the fact that $\sum_{i=0}^{m_1} \binom{m_1}{i} = 2^m$ and $\sum_{i=0}^{m_1} \binom{m_1}{i} = 2^{m_1-1}$. This formula is simpler than the general formula given in Theorem 3.11 and it may be desirable in the context of reducing computational time.

### 3.4. Bandwidth selection

In Rosenblatt (1956), approximations for the bias, the variance, the mean squared error and the mean integrated squared error were derived in the univariate case. It was shown that by minimising the asymptotic mean integrated squared error (AMISE), an optimal bandwidth could be found when information of $\int_{-\infty}^{\infty} |d^2 f/dx^2|^2 \, dx$ exists. It is not possible to make such an elegant argumentation for pseudometric kernel density estimation. However, an expression, including one-dimensional integrals, for the AMISE of $f_{\sim d_{p,1,\mathcal{A},p}}$ is derived when $f$ is equal to $f_{\sim d_{p,1,\mathcal{A},p}}$.

**Theorem 3.13** The asymptotic mean squared error AMSE$(x)$ of $f_{d_{p,1,\mathcal{A},p}}(x)$ and the AMISE of $f_{d_{p,1,\mathcal{A},p}}$ are

$$
\text{AMSE}(x) = \sigma^2(f_{d_{p,1,\mathcal{A},p}}(x)) + (E(f_{d_{p,1,\mathcal{A},p}}(x)) - f(x))^2,
$$

$$
\text{AMISE} = \int_0^\infty \frac{dV_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{d\alpha} \text{AMSE}_1(\alpha) \, d\alpha,
$$

(12)

where $E(f_{d_{p,1,\mathcal{A},p}}(x))$ and $\sigma^2(f_{d_{p,1,\mathcal{A},p}}(x))$ are given by

$$
E(f_{d_{p,1,\mathcal{A},p}}(x)) = \frac{1}{h^m} \int_0^\infty K\left(\frac{V_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{V_m h^m}\right)^{1/m} \frac{dV_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{d\alpha} f_1(\alpha) \, d\alpha.
$$

(13)

and

$$
\sigma^2(f_{d_{p,1,\mathcal{A},p}}(x)) = \frac{1}{nh^{2m}} \left[ \int_0^\infty K^2\left(\frac{V_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{V_m h^m}\right)^{1/m} \frac{dV_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{d\alpha} f_1(\alpha) \, d\alpha - \left(\int_0^\infty K\left(\frac{V_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{V_m h^m}\right)^{1/m} \frac{dV_{B_{d_{p,1,\mathcal{A},p}[\alpha,\alpha]}}}{d\alpha} f_1(\alpha) \, d\alpha\right)^2 \right].
$$

(14)

Here, $\alpha_x = d_{s,1,p}(0,x)$, $\alpha_y = d_{s,1,p}(0,y)$, $f_1(\alpha_y) = f(y)$ and $\text{AMSE}_1(\alpha) = \text{AMSE}(x).$
4. Experiments

In this section, a few experiments are presented where various PKDE are compared with the standard KDE and the flexible Bayesian estimator. The experiments involve drawing repetitively an order pair \((x, X)\), \(n_{\text{rep}}\) times. Here, \(X\) is a sample of \(n\) points that are drawn from a known distribution \(f\) and \(x\) is a query point that is drawn randomly from an equivalence class \([\{r_i, 0, 0, \ldots, 0\}]\). The sequence \((r_i)_{i=1}^{n_{\text{eqc}}}\) starts at 0 and is uniformly spaced over the relevant domain of \(f\). For every \(r_i\), the mean and the quartiles are computed and compared with the known analytic solution. When plotted as a function of \(r_i\), the variances and the biases of the estimators are easily seen.

Seven experiments are conducted, where PKDE, KDE and flexible Bayesian are compared on a set of distributions, including the standard normal, the uniform, the one-sided normal and a one-sided normal that is split into two. Variable PKDE and variable KDE are compared only on the standard normal distribution. The parameters are summarised in Tables 2 and 3.

4.1. Results

In experiment one (Figure 6), the estimator is shown in various dimensions while the bandwidth is kept constant. The bandwidth is chosen to 0.5, which is the bandwidth that minimises AMISE (12) in six dimensions. The estimator seems to be unbiased in the tails, while it underestimates near

| Parameter | Argument |
|-----------|----------|
| Number of repetitions \(n_{\text{rep}}\) | 10,000 |
| Sample size \(n\) | 1000 |
| Number of equivalence classes \(n_{\text{eqc}}\) | 50 |
| Pseudometric family | Type one |
| Transformation matrix \(A\) | Identity |
| Centre point \(c\) | 0 |

Table 2. Parameters for all experiments.

| Distribution type | Dimensions\(^a\) | Method | Bandwidth | Exponents in pseudometric |
|-------------------|-----------------|--------|-----------|--------------------------|
| Standard normal   | 2, 4, 6         | PKDE   | 0.5       | All are 2                |
| Standard normal   | 6, 0.8          | PKDE   | 0.2       | All are 2                |
| Standard normal   | 6               | KDE    | 0.5       | NA                       |
| Flexible Bayesian |                 | Flexible Bayesian | 0.2 |                      |
| Standard normal   | 6               | Variable PKDE | 15th nearest neighbour | All are 2 |
| Flexible Bayesian |                 | Variable KDE |               |                          |
| Uniform on \([-0.5, 0.5]\) | 6               | PKDE   | 0.15      | All are 12               |
| Flexible Bayesian |                 | Flexible Bayesian | 0.15 |               |
| One-sided standard normal | 6 | PKDE | 0.25 | All are 2 |
| Flexible Bayesian |                 | Flexible Bayesian | 0.25 |                     |
| One-sided standard normal split into two | 6 | PKDE | 0.25 | All are 2 |
| Flexible Bayesian |                 | Flexible Bayesian | 0.25 |                     |

\(^a\)Number of dimensions.
the centre where the level of underestimation increases with the number of dimensions. The tails seem fairly unbiased, but as we move away from the centre, the mean normalised experimental error starts to fluctuate. From the quartile plots, it is seen that the skewness of the normalised experimental error increases with distance from the centre.

In experiment two (Figure 7), we see the effect of increasing and decreasing the bandwidth in six dimensions. When the bandwidth is increased, the underestimation near the centre increases while the variance decreases everywhere. When the bandwidth is decreased, the bias is reduced at the cost of a higher variance.

In experiment three (Figure 8), KDE with fixed bandwidth and flexible Bayesian is shown for comparison with PKDE in experiment one. Since the bandwidth in KDE is the same as for PKDE, the bias and the variance are the same near the centre, while KDE continues to underestimate much further away from the centre. Also, when the analytical graph is passed, KDE starts to overestimate. Obviously, an optimal bandwidth for KDE is smaller than for PKDE, since the bias in the tail is larger. This underestimation and overestimation pattern is also seen in the flexible Bayesian, but to a lower extent. In the case of the flexible Bayesian, we have chosen a bandwidth so that the variance near the centre is approximately the same as the KDE and the PKDE. This means that it is easier to compare the biases.

In experiment four (Figure 9), PKDE and KDE with variable bandwidths are shown for comparison. The number of neighbours is chosen so that the variances near the centres are approximately the same as in Figures 6 and 8, where the bandwidths are fixed. It is interesting to see that the variance of the normalised experimental error is fairly stable and not skewed compared to the fixed bandwidth graphs. The bias in the tails is clear in these plots. Notice also that the PKDE with variable bandwidth has a similar, or smaller, variance and bias than the flexible Bayesian inside the shaded area.

In experiment five (Figure 10), PKDE, KDE and flexible Bayesian are compared on a uniform population. The bandwidth of the flexible Bayesian is adjusted so the variance near the centre is similar for all estimators. Clearly, PKDE has the sharpest border, while KDE is outperformed by the other two.
Figure 7. Result of experiment two. Kernel density estimates for a standard normal distribution in six dimensions with various degrees of smoothing. The top row shows the probability density functions as functions of the distances from the centre, while the bottom row shows the corresponding normalised standard errors. The distance from the centre is chi distributed with order six. The shaded area is where the cumulative distribution function is between 0.025 and 0.975. (a) Bandwidth is 0.2 and (b) bandwidth is 0.8.

In experiment six (Figure 11), the estimators are evaluated on a one-sided normal distribution. When the bandwidth is half the bandwidth used in experiment one, PKDE performs identically to PKDE in experiment one. KDE, on the other hand, underestimates far more than in the two sided case in experiment three. This is particularly the case near the centre. The same effect is present for the flexible Bayesian.

In experiment seven (Figure 12), the one-sided normal distribution is split into two equally sized populations. The left population contains points where the radius is less than 2.3126 and the right population contains points where the radius is more than 2.3126. The bandwidth settings were the same as in experiment six, which is probably not optimal for a classification. However, it is still clear that PKDE would be favourable in a classification setting. This is because, in the case of PKDE and KDE, there is a trade off in how well the means are fitted on the two populations. The better fit we have on one population the worse the fit will be on the other population. Here, PKDE fits much better on both populations and is therefore favourable. The flexible Bayesian estimator obviously performs very poorly on the right population, because of the severe violation of the independence assumption.

4.2. Discussion of the experiments

KDE and variable KDE are known to be consistent, but in the above experiments the estimators are clearly biased. This is because the sample size is low compared to the number of dimensions.
Figure 8. Result of experiment three. KDE and flexible Bayesian estimates for a standard normal distribution in six dimensions, where the bandwidths are 0.5 and 0.2, respectively. The top row shows the probability density functions as functions of the distances from the centre, while the bottom row shows the corresponding normalised standard errors. The distance from the centre is chi distributed with order six. The shaded area is where the cumulative distribution function is between 0.025 and 0.975. (a) KDE and (b) flexible Bayesian.

In a practical setting, this is often the case and as a result, nonparametric estimators may suffer from poor statistical power.

In all experiments, the estimators are tested on associated distributions that satisfy the strong condition. In a practical setting, the knowledge about the parameters is less exact, where for instance only the priors are given. In a fair comparison, estimation of parameters should be part of the experiment, but this topic has not been treated in this manuscript.

It is worth noting that PKDE is fairly biased near the centre on the normal distribution. The reason for this is that the radii of the balls with constant volume are larger near the centre. Large contributions from points that are far away (in the pseudometric space) together with a high curvature of the distribution explains this effect. The underestimation is therefore also dependent on the number of dimensions and the choice of bandwidth. The bias near the centre is not seen in the uniform distribution, because the curvature is 0.

In the case of using a variable bandwidth, it seems that the skewness of the normalised experimental error is reduced everywhere at the cost of overestimation in the tails. The reason for the overestimation in the tails is that a query point that is far away from any point in the sample will always have contributions because of the increasing bandwidth. When choosing between variable and fixed PKDE (and also KDE), one should consider the need for accuracy in the tails. In case of classification, it may be desirable to detect very small densities, when the query points are far from any of the points in the population.
Figure 9. Result of experiment four. Variable PKDE and variable KDE for a standard normal distribution in six dimensions. Two standard deviations of the Gaussian kernel are defined to be the distance (in the pseudometric space) to the 15th nearest neighbour. The top row shows the probability density functions as functions of the distances from the centre, while the bottom row shows the corresponding normalised standard errors. The distance from the centre is chi distributed with order six and the shaded area is where the cumulative distribution function is between 0.025 and 0.975. (a) Variable PKDE and (b) variable KDE.

Figure 10. Result of experiment five. PKDE, KDE and flexible Bayesian on a six-dimensional uniform distribution on $[-0.5, 0.5]$. The bandwidth for the flexible Bayesian estimate is 0.04, while it is 0.15 for the other two. The shaded area is where the cumulative distribution function of the (pseudometric) distance to the centre is between 0.025 and 0.975. (a) PKDE, (b) KDE and (c) flexible Bayesian.

In the above experiments, it has also been revealed instances where KDE and flexible Bayesian perform poor. In the case of the uniform population, KDE struggles with the sharp border that is heavily smoothed to decrease variance. This is a great example where PKDE promises great potential over KDE. Notice also that PKDE performs even better than the flexible Bayesian.
Figure 11. Result of experiment six. PKDE, KDE and flexible Bayesian for a one-sided standard normal distribution in six dimensions, where the bandwidths are 0.1 for the flexible Bayesian estimates and 0.25 for the other methods. The top row shows the probability density functions as functions of the distances from the centre, while the bottom row shows the corresponding normalised standard errors. The distance from the centre is chi distributed with order six. The shaded area is where the cumulative distribution function is between 0.025 and 0.975. (a) PKDE, (b) KDE and (c) flexible Bayesian.

Figure 12. Result of experiment seven. Various kernel density estimates for two populations. The left population has a probability density function that is twice the one-sided standard normal distribution in six dimensions when the radius is less than 2.3126 and 0 elsewhere. The right population has a probability density function that is twice the one-sided standard normal distribution in six dimensions when the radius is more than 2.3126 and 0 elsewhere. The bandwidths are 0.1 for the flexible Bayesian estimates and 0.25 for the other methods. The shaded areas are where the cumulative distribution functions of the (pseudometric) distances to the centre are between 0.025 and 0.975. (a) Left population and (b) right population.

In the case of a one-sided population, it is clearly seen that KDE and flexible Bayesian underestimate severely near the centre and continue to underestimate far into the tails. The reason for this is that the estimators are smoothing over orthants where the distribution is 0 by definition. This does not happen in PKDE, because the volume calculation is only done on the relevant orthant.

The one-sided normal distribution that is split in two is added to symbolise a two class classification problem of practical interest. The flexible Bayesian struggles with the right population because of the obvious dependency between explanatory variables. KDE performs poorly on the left population as a result of the effect described in experiment six. It also performs poor on the
right population because the boundary is sharp. Clearly, PKDE would perform well if the density estimates were used as inputs to a classification rule.

To understand the full extent of experiment seven, it is important to note that similar results would be evident for a great variety of distributions. Each explanatory variable could be a generalised normal, which could also be either one-sided left, one-sided right or two sided. In addition, the whole distribution could be linearly transformed.

As a concluding remark of the experiments, it is clear that increased statistical power is observed for PKDE with pseudometric of type one, when proper assumptions about the distributions are made. Obviously, if poor assumptions were made, the estimator would suffer from poor robustness. PKDE with a pseudometric of type two has pseudometrics of type one and KDE as special cases. If care is taken when choosing parameters (or priors for parameters), it is therefore possible to increase statistical power at little expense of robustness.

5. Open problems and the use of the new estimators on other problems than density estimation

5.1. Asymptotic argumentation

The literature on kernel density and kernel regression is full of asymptotic argumentation for any variation of estimators. Description of consistency and rates of convergence started with the early works of Fix and Hodges (1951), Rosenblatt (1956) and Parzen (1962) in the one-dimensional case. Optimal rates of convergence for nonparametric estimators were described in Stone (1980).

It is an interesting question if it is possible to outline an argument for the rates of convergence and AMISE for the estimators described in this paper, provided that the data sample is drawn from an associated distribution. This seems difficult, since the results of the derivations of the bias and AMISE in Section 3.4 contain integrals that seem difficult to treat. However, it is possible that there are approximations that are not seen by the author of this manuscript.

5.2. Parameter selection

In the case of kernel density estimation, the parameters, usually the bandwidth matrix, have usually been selected using either plug-in or cross-validation methods. An introduction can be found in Wand and Jones (1994), while more recent articles are given by Duong and Hazelton (2003, 2005). In short, the plug-in selector is found by minimising an analytic expression of the AMISE. Since analytical expressions for AMISE do not exist yet, this path for fitting parameters seems difficult. It is far more likely that a cross-validation algorithm is a better route, because it is minimising the mean integrated squared error directly. This is a very important question, but the scope of it is so large that it is left as an open question in this paper.

If we succeed with fitting parameters in the future, the parameters that give the best fit describe the level sets of the estimated density. This is useful for exploration of data and interpreting results. As an example, it was enough to describe the density as a function of distances from the centre in the experiments in Section 4.1.

5.3. Kernel regression

The new nonparametric density estimators can be used on various problems. In terms of regression, most kernel smoothing algorithms can be generalised by using pseudometrics because they are essentially derived from the probability densities (Watson 1964). An overview of kernel smoothing
algorithms is given by Simonoff (1996) and Wand and Jones (1994). A generalisation of the Nadaraya–Watson kernel-weighted average is proposed as

$$\hat{Y}(x) = \frac{\sum_{i=1}^{n} K_h(e_{dp}(x, x_i)) Y(x_i)}{\sum_{i=1}^{n} K_h(e_{dp}(x, x_i))},$$

(15)

where the $Y(x_i)$s are the outputs related to the inputs $x_i$s in the training set. Notice that $\hat{Y}(x)$ is constant on subsets of the level sets of the estimated density. If the level of smoothing is decided, that is $h$, parameters can be found by minimising the squared residuals. In this case, the parameters that give the best fit describe the level sets of $\hat{Y}(x)$ (and generally not the level sets of the density, since the parameters have been fitted to minimise the residuals). The bandwidth parameter is really a question of how closely one wants to fit the data, thus obviously minimising the residuals will favour very small bandwidths. It is left as an open problem how to integrate the bandwidth selection with fitting all the other parameters. A potential path is to penalise roughness of the $\hat{Y}(x)$, similar to, for instance, the method laid out in Hurvich and Simonoff (1998).

The model above fits somewhat into the plethora of restricted nonparametric and semiparametric regression methods to reduce the effect of high dimensionality. The additive model has the form of a linear combination of one-dimensional smoothers. This method is strong when the explanatory variables are independent, while dependencies are only dealt with through the use of the back fitting algorithm. In projection pursuit, the additive model is generalised by allowing a projection of the data matrix of explanatory variables before application of the one-dimensional smoothers. The additive models and the projection pursuit models were first defined in Friedman and Stuetzle (1981), with the functional extension given in Ferraty, Goia, Salinelli, and Vieu (2013). The projection pursuit regressor has in common with Equation (15) that the visualisation and interpretation of the whole distribution are possible in lower dimensional spaces. However, the regressors will generally have very different forms, since in projection pursuit the regressor is a sum of one-dimensional curves, while Equation (15) has parts of the level sets described by the pseudometric.

In partially linear models the explanatory variables are divided into a linear part and a nonparametric part. The general principles are described in the book of Härdle, Liang, and Gao (2000) and the functional extension is described in Aneiros-Pérez and Vieu (2006). This method has less in common with Equation (15), since $\hat{Y}(x)$ is generally not linearly dependent on any of the explanatory variables.

### 5.4. Classification

The classification rule can use density estimates for each subpopulation (Taylor 1997). More precisely, the classification rule is

$$C(x) = \arg\max_i (\pi_i \hat{f}_{di}(x)),$$

where the $\pi_i$s are the prior probabilities and $\hat{f}_{di}$s are the density estimates for each subpopulation. Recognise that different pseudometrics can be used for each subpopulation.

### 5.5. Statistical learning

Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis as described in Mohri, Rostamizadeh, and Talwalkar (2012). The theory deals with the problem of finding a predictive function based on a set of data. The problems that
can be solved through this framework are many, including density estimation, classification and regression as discussed above. Another problem that is worth mentioning is data clustering. It is possible that the proposed estimators can be used in density-based clustering (Kriegel, Kröger, Sander, and Zimek 2011).

5.6. Some practical considerations related to training parameters for pseudometric kernel density estimation

Clearly, it is not a trivial challenge to find a recipe for training parameters on a wide range of distributions (or problems). Both searching the parameter space and avoiding over fitting are challenges. Limiting the hypothesis space is an important tool to improve computational time and avoid over fitting.

An interesting idea is to use a hypothesis test to find which explanatory variables are dependent on each other and rearrange them so they are grouped together. A semi-naive Bayesian estimator is proposed as

\[
\hat{f}_h(x) = \prod_{(o_j)_{j=2}^{\text{max}}} \left( \frac{1}{n} \sum_{i=1}^{n} K_h(e_{d_p}(x_{o_{j-1}+1:o_j}, x_{1:o_{j-1}+1:o_j})) \right),
\]

where \((o_j)\) is the sequence of the last element in each group and \(o_1\) equals 0.

Often, the bottleneck in computational time for kernel density estimation is the training of the parameters. When cross-validation is part of the training, the computational time is dominated by the calculations that are \(O(n^2)\). In pseudometric kernel density estimation, distance, volume and kernel function calculations are all \(O(n^2)\). However, all \(d_{A,p}(x, e)\) calculations and all coefficients in the volume polynomials can be precomputed before the cross-validation starts. This means that the new estimators can have a computational performance that is comparable to common KDE.

6. Conclusion

An idea for optimising kernel density estimation using pseudometrics is described, where assumptions on the kernel of the distribution can be given. In simulations, the proposed estimators show increased statistical power when proper assumptions are made. As a consequence, this paper describes an approach, where partial knowledge about the distribution can be used effectively. It is also argued that the new estimators may have a potential in statistical learning algorithms such as classification, regression and data clustering.

Acknowledgements

The author thanks Kay Steen, Frode Sørmo, Odd Erik Gundersen, Håkon Lorentzen, Waclaw Kusnierczyk and Helge Langseth and in particular the referees for many helpful comments.

Funding

This work has been partially supported by Verdande Technology and AMIDST. AMIDST is a project that is supported by the European Union’s Seventh Framework Programme for research, technological development and demonstration [grant agreement no. 619209].
References

Aneiros-Péres, G., and Vieu, P. (2006), ‘Semi-functional Partial Linear Regression’, Statistics and Probability Letters, 76, 1102–1110.

Breiman, L., Meisel, W., and Purcell, E. (1977), ‘Variable Kernel Estimates of Multivariate Densities’, Technometrics, 19, 135–144.

Cacoullos, T. (1966), ‘Estimation of a Multivariate Density’, Annals of the Institute of Statistical Mathematics, 18, 179–189.

Delaigle, A., and Hall, P. (2010), ‘Defining Probability Functions for a Distribution of Random Functions’, The Annals of Statistics, 38, 1171–1193.

Duong, T., and Hazelton, M.L. (2003), ‘Plug-in Bandwidth Matrices for Bivariate Kernel Density Estimation’, Journal of Nonparametric Statistics, 15, 17–30.

Duong, T., and Hazelton, M.L. (2005), ‘Cross-validation Bandwidth Matrices for Multivariate Kernel Density Estimation’, Scandinavian Journal of Statistics, 32, 485–506.

Epanechnikov, V.K. (1969), ‘Non-parametric Estimation of a Multivariate Probability Density’, Theory of Probability and Its Applications, 14, 153–158.

Fix, E., and Hodges, J.L. (1951), ‘Discriminatory Analysis – Nonparametric Discrimination: Consistency Properties’, Journal of Nonparametric Statistics, 24, 447–464.

Ferraty, F., Kudraszow, N., and Vieu, P. (2012), ‘Nonparametric Estimation of a Surrogate Density Function in Infinite-dimensional Spaces’, Journal of Nonparametric Statistics, 14, 231–240.

Friedman, J.H., Stuetzle, W., and Schroeder, A. (1984), ‘Projection Pursuit Density Estimation’, Journal of the American Statistical Association, 79, 599–608.

Fix, E., and Hodges, J.L. (1951), ‘Discriminatory Analysis – Nonparametric Discrimination: Consistency Properties’, Technical Report 4, United States Air Force School of Aviation Medicine, Randolph Field, TX.

Folland, G.B. (2001), ‘How to Integrate a Polynomial Over a Sphere’, The American Mathematical Monthly, 108, 446–448.

Friedman, J.H., and Stuetzle, W. (1981), ‘Projection Pursuit Regression’, Journal of the American Statistical Association, 76, 817–823.

Friedman, J.H., and Stuetzle, W. (1984), ‘Projection Pursuit Density Estimation’, Journal of the American Statistical Association, 79, 599–608.

Gao, F. (2013), ‘Volumes of Generalized Balls’, The American Mathematical Monthly, 120–130.

Härdle, W., Liang, H., and Gao, J. (2000), Partially Linear Models, New York: Springer.

Holmström, L. (2003), ‘A Semiparametric Density Estimation Approach to Pattern Recognition’, Pattern Recognition, 37, 409–419.

Hurvich, C.M., and Simonoff, J.S. (1998), ‘Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion’, Journal of the Royal Statistical Society: Series B, 60, 271–293.

John, G.H., and Langley, P. (1995), ‘Estimating Continuous Distributions in Bayesian Classifiers’, in Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence, August, San Mateo, CA: Morgan Kaufmann, pp. 338–345.

Kriegel, H.P., Kröger, P., Sander, J., and Zimek, A. (2011), ‘Density-Based Clustering’, Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery, 1, 231–240.

Liescher, E. (2005), ‘A Semiparametric Density Estimator Based on Elliptical Distributions’, Journal of Multivariate Analysis, 92, 205–225.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012), Foundations of Machine Learning, Cambridge: The MIT Press.

Parzen, E. (1962), ‘On Estimation of a Probability Density Function and Mode’, The Annals of Mathematical Statistics, 33, 1065–1076.

Rosenblatt, M. (1956), ‘Remarks on Some Nonparametric Estimates of a Density Function’, The Annals on Mathematical Statistics, 27, 832–837.

Simonoff, J.S. (1996), Smoothing Methods in Statistics, New York: Springer Series in Statistics.
Appendix 1. Proofs

A.1 Proof of Theorem 3.4

We use that the \( m \)-dimensional generalised normal distribution \( f(x) \) can be expressed as a function \( g(\epsilon) \), where \( \epsilon = \sum_i |x_i|^{p_i} \), so

\[
f(x) = \prod_{i=1}^{m} \frac{p_i}{2\Gamma(1/p_i)} \exp(-|x_i|^{p_i}) = g(\epsilon) = \frac{1}{2^m \prod_{i=1}^{m} \Gamma(1 + 1/p_i)} \exp(-\epsilon).
\]

Since \( f(x) \) integrates to 1, then the following also holds for \( g(\epsilon) \)

\[
\int_{0}^{\infty} \frac{dV_{\mathbb{B}_{p_1,p_2,...,p_m}[\epsilon]}}{d\epsilon} g(\epsilon) \, d\epsilon = 1,
\]

where the ball with radius \( \epsilon \) is defined as

\[
\mathbb{B}_{p_1,p_2,...,p_m}[\epsilon] = \left\{ x \in \mathbb{R}^m \mid \sum_i |x_i|^{p_i} \leq \epsilon \right\}.
\]

Let \( B \) be a diagonal matrix, where the \( i \)-th diagonal element \( b_{ii} = \epsilon^{-1/p_i} \), then we recognise that

\[
\mathbb{B}_{p_1,p_2,...,p_m}[\epsilon] = \left\{ x \in \mathbb{R}^m \mid \sum_i |b_{ii}x_i|^{p_i} \leq 1 \right\},
\]

which means that \( \mathbb{B}_{p_1,p_2,...,p_m}[\epsilon] \) is the generalised unit ball that has been linearly transformed by \( B^{-1} \). The volume of \( \mathbb{B}_{p_1,p_2,...,p_m}[\epsilon] \) is therefore

\[
V_{\mathbb{B}_{p_1,p_2,...,p_m}[\epsilon]} = \frac{1}{|\det(B)|} V_{\mathbb{B}_{p_1,p_2,...,p_m}[1]} = V_{\mathbb{B}_{p_1,p_2,...,p_m}[1]}^{m/p},
\]

where \( p \) is the harmonic average of the \( p_i \)'s. This means that Equation (A1) reduces to

\[
\int_{0}^{\infty} \frac{dV_{\mathbb{B}_{p_1,p_2,...,p_m}[\epsilon]}}{d\epsilon} \epsilon^{m/p-1} \exp(-\epsilon) \, d\epsilon = 1.
\]

Since the integral is \( \Gamma(m/p) \), then \( V_{\mathbb{B}_{p_1,p_2,...,p_m}[1]} \) must be

\[
V_{\mathbb{B}_{p_1,p_2,...,p_m}[1]} = \frac{2^m \prod_{i=1}^{m} \Gamma(1 + 1/p_i)}{\Gamma(1 + m/p)}.
\]

A.2 Proof of Theorem 3.5

Let \( B \) be a diagonal matrix, where the \( i \)-th diagonal element is \( \epsilon^{-p_i/p_i} \), then we recognise that

\[
\mathbb{B}_{d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}}[0, \epsilon] = \{ x \in \mathbb{R}^m \mid d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}(Bx, 0) \leq 1 \},
\]

which means that \( \mathbb{B}_{d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}}[0, \epsilon] \) is the generalised unit ball that has been linearly transformed by \( B^{-1} \mathcal{A}^{-1} \). The volume of \( \mathbb{B}_{d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}}[0, \epsilon] \) is therefore

\[
V_{\mathbb{B}_{d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}}[0, \epsilon]} = V_{\mathbb{B}_{d_{\Delta_{\mathcal{A}}} \mathcal{A} \mathcal{P}}[0, \epsilon]} = \frac{1}{|\det(\mathcal{A})| |\det(B)|} \frac{V_{p}}{V_{p}},
\]

\[
= \frac{1}{|\det(\mathcal{A})|} V_{p}^{m}.
\]
A.3 Proof of Theorem 3.7

We recognise that
\[ \mathbb{R}^m/\sim_{d_{1,A,p}} = \{ x \in \mathbb{R}^m \mid d_{1,A,p}(x,c) \leq \epsilon \} \forall \epsilon \in \mathbb{R}^+. \]

Each equivalence class in \( \mathbb{R}^m/\sim_{d_{1,A,p}} \) is therefore uniquely determined by \( u \), thus \( f_{\sim_{d_{1,A,p}}}(u) = g(u) \). Moreover, \( f_{\sim_{d_{1,A,p}}} \) must integrate to 1, which means that the following is valid for \( g(u) \)

\[ \int_0^\infty \frac{dV_{\mathbb{R}^m}}{du} g(u) \, du = 1, \]
\[ \frac{V_{n(m/p-1)}}{\det(A)} \int_0^\infty u^{m/p-1} g(u) \, du = 1. \]

The strong condition is satisfied, because
\[ \mathbb{R}^m/\sim_{d_{1,A,p}} = \{ x \in \mathbb{R}^m \mid d_{1,A,p}(x,c) \leq \epsilon \} \forall \epsilon \in \mathbb{R}^+. \]

A.4 Proof of Theorem 3.8

When \( \epsilon < d_{1,A,p}(a,c) \)

\[ V_{\mathbb{R}^m/\sim_{d_{1,A,p}}} \leq V_{\mathbb{R}^m/\sim_{d_{1,A,p}}(a,c+c)} - V_{\mathbb{R}^m/\sim_{d_{1,A,p}}(a,c-c)} \]
\[ = \frac{V_p}{\det(A)} \left( (d_{1,A,p}(a,c) + \epsilon)^m - (d_{1,A,p}(a,c) - \epsilon)^m \right) \]
\[ = \frac{2V_p}{\det(A)} \sum_{i=0}^m \binom{m}{i} (d_{1,A,p}(a,c))^{m-i} \epsilon^i. \]

When \( \epsilon \geq d_{1,A,p}(a,c) \)

\[ V_{\mathbb{R}^m/\sim_{d_{1,A,p}}} \leq V_{\mathbb{R}^m/\sim_{d_{1,A,p}}(a,c+c)} \]
\[ = \frac{V_p}{\det(A)} (d_{1,A,p}(a,c) + \epsilon)^m \]
\[ = \frac{V_p}{\det(A)} \sum_{i=0}^m \binom{m}{i} (d_{1,A,p}(a,c))^{m-i} \epsilon^i. \]

A.5 Proof of Theorem 3.10

We recognise that
\[ \mathbb{R}^m/\sim_{d_{2,A,p,q}} = \{ x \in \mathbb{R}^m \mid d_{1,p}(x_1,1,1) = \epsilon \land x_{m_1+1,m} = q \} \forall \epsilon \in \mathbb{R}^+ \land \forall q \in \mathbb{R}^m, \]

which means that each equivalence class in \( \mathbb{R}^m/\sim_{d_{2,A,p,q}} \) is uniquely determined by \( u \) and \( x_{m_1+1,m} \), thus \( f_{\sim_{d_{2,A,p,q}}}(u) = g(u, x_{m_1+1,m}) \).

Since \( f_{\sim_{d_{2,A,p,q}}} \) integrates to 1, then the following is valid for \( g(u, x_{m_1+1,m}) \)

\[ \int_{\mathbb{R}^m} g(u, x_{m_1+1,m}) \, dx_1 \, dx_2 \ldots \, dx_m = 1, \]
\[ \int_{\mathbb{R}^{m_1}} \int_{\mathbb{R}^{m_2}} \int_{\mathbb{R}^{m_3}} \ldots \int_{\mathbb{R}^{m_{m-1}}} \frac{dV_{\mathbb{R}^m}}{du} g(u, (x_{m_1+1,m})) \, dx_1 \, dx_2 \ldots \, dx_m = |\det(A)|, \]
\[ \int_{\mathbb{R}^{m_1}} \int_{\mathbb{R}^{m_2}} \int_{\mathbb{R}^{m_3}} \ldots \int_{\mathbb{R}^{m_{m-1}}} \frac{dV_{\mathbb{R}^m}}{du} g(u, (x_{m_1+1,m})) \, dx_1 \, dx_2 \ldots \, dx_m = |\det(A)|, \]
\[ \frac{V_{n(m/p-1)}}{\det(A)} \int_{\mathbb{R}^{m_1}} \int_{\mathbb{R}^{m_2}} \int_{\mathbb{R}^{m_3}} \ldots \int_{\mathbb{R}^{m_{m-1}}} \frac{dV_{\mathbb{R}^m}}{du} g(u, (x_{m_1+1,m})) \, dx_1 \, dx_2 \ldots \, dx_m = 1. \]
A.6 Proof of Theorem 3.11

By a simple change of variables, we recognize that

\[ V_{B_{p,2A,e,q}}[e] = \frac{1}{|\text{det}(A)|} V_{B_{p,2A,e,q}}[\alpha,e] \]

and also that \( V_{B_{p,2A,e,q}}[\alpha,e] \) can be found without loss of generality, by choosing \( \alpha_{m1+1:m} = 0 \). Therefore

\[ V_{B_{p,2A,e,q}}[\alpha,e] = \int \int_{\alpha, q} \ dx_1 \ dx_2 \ldots \ dx_m \]

\[ = \int_{\alpha, q} \sum_{i=1}^{m} |x_i|^q \ dx_1 \ dx_2 \ldots \ dx_m \]

Now let \( \alpha^q = \sum_{i=1}^{m} |x_i|^q \) and \( \beta^q = \epsilon^q - \alpha^q \), then the inner integral is

\[ \int_{\alpha, q} \sum_{i=1}^{m} |x_i|^q \ dx_1 \ dx_2 \ldots \ dx_m = V_{B_{p,2A,e,q}}[\alpha_{1:m}] \]

If we substitute for \( \beta \), we see that the inner integral is a function \( U \) of \( \alpha \)

\[ U(\alpha) = \begin{cases} 
V_p \sum_{i=1}^{m} \binom{m}{i} d_{1:p}(\alpha_{1:m1}, c')m1-i (\epsilon^q - \alpha^q) i/q & \text{if } \alpha \leq \epsilon^{1/q}, \\
2V_p \sum_{i=0}^{m} \binom{m}{i} d_{1:p}(\alpha_{1:m1}, c')m1-i (\epsilon^q - \alpha^q) i/q & \text{if } \alpha > \epsilon^{1/q},
\end{cases} \]

where \( s \) is defined in Theorem 3.11. The full integral is

\[ V_{B_{p,2A,e,q}}[\alpha,e] = \int_{\alpha,e} U(\alpha) \ dx_{m1+1} \ dx_{m1+2} \ldots \ dx_m, \]

and if we let \( V_{m2}^\alpha(\alpha) \) be the volume of the \( L^q \) ball with radius \( \alpha \) in \( m_2 \)-dimensions, then we can integrate over \( \alpha \), so that

\[ V_{B_{p,2A,e,q}}[\alpha,e] = \int_{0}^{\epsilon} U(\alpha) \frac{dV_{m2}^\alpha(\alpha)}{d\alpha} \ d\alpha, \quad \text{where} \quad \frac{dV_{m2}^\alpha(\alpha)}{d\alpha} = \frac{q(2(1+1/q))^{m2}}{\Gamma(m2/q)} \alpha^{m2-1}.\]

Substituting and reorganising gives

\[ V_{B_{p,2A,e,q}}[\alpha,e] = V_p \frac{q(2(1+1/q))^{m2}}{\Gamma(m2/q)} \left( \sum_{i=0}^{m} \binom{m}{i} d_{1:p}(\alpha_{1:m1}, c')m1-i \int_{\epsilon^{1/q}}^{\epsilon} (\epsilon^q - \alpha^q) i/q \alpha^{m2-1} \ d\alpha \right) \]

\[ + 2 \sum_{i=0}^{m} \binom{m}{i} d_{1:p}(\alpha_{1:m1}, c')m1-i \int_{\epsilon^{1/q}}^{\epsilon} (\epsilon^q - \alpha^q) i/q \alpha^{m2-1} \ d\alpha \]

If we change the variables in the integrals with \( \alpha\int / \epsilon^q \), we see that the integral parts can be expressed with incomplete beta functions. The two integrals are

\[ \frac{1}{q} B_i \left( \frac{m2}{q} + i + 1 \right) \epsilon^{m2+i} \quad \text{and} \quad \frac{1}{q} B_i \left( \frac{m2}{q} + i + 1 \right) - \frac{1}{q} B_i \left( \frac{m2}{q} + i + 1 \right) \epsilon^{m2+i}, \]

respectively, so

\[ V_{B_{p,2A,e,q}}[\alpha,e] = V_p \frac{q(2(1+1/q))^{m2}}{\Gamma(m2/q)} \sum_{i=0}^{m} \binom{m}{i} B_i \left( \frac{m2}{q} + i + 1 \right) d_{1:p}(\alpha_{1:m1}, c')m1-i \epsilon^{m2+i} \]

\[ + 2 \sum_{i=0}^{m} \binom{m}{i} \left( B_i \left( \frac{m2}{q} + i + 1 \right) - B_i \left( \frac{m2}{q} + i + 1 \right) \right) d_{1:p}(\alpha_{1:m1}, c')m1-i \epsilon^{m2+i} \]
A.7 Proof of Theorem 3.12

In the case when \( \epsilon < d_{1,p}(\alpha_{1,m_1}', c') \), \( s \) is zero and in the case when \( \epsilon \geq d_{1,p}(\alpha_{1,m_1}', c') \), \( s \to 1 \) as \( q \to \infty \). From Theorem 3.11, we see that

\[
\lim_{q \to \infty} V_{B_{\beta_{2,A,p,q}}[a,c]} = \begin{cases} 
V_p^{2m_2+1} \frac{\Gamma(1+1/q)^{m_1}}{|\det(A)| \Gamma(m_2/q)} \sum_{i=0}^m \left( \frac{m_1}{i} \right) B(1, \frac{m_2}{q}, i+1) , \\
V_p^{2m_2} \frac{\Gamma(1+1/q)^{m_1}}{|\det(A)| \Gamma(m_2/q)} \sum_{i=0}^m \left( \frac{m_1}{i} \right) B(1, \frac{m_2}{q}, i+1) , \\
V_p^{2m_2} \frac{1}{q^2} \frac{\Gamma(1+1/q)^{m_1}}{|\det(A)| \Gamma(m_2/q)} \sum_{i=0}^m \left( \frac{m_1}{i} \right) B(1, \frac{m_2}{q}, i+1) , \\
V_p^{2m_2} \frac{1}{q^2} \frac{\Gamma(1+1/q)^{m_1}}{|\det(A)| \Gamma(m_2/q)} \sum_{i=0}^m \left( \frac{m_1}{i} \right) B(1, \frac{m_2}{q}, i+1) , \\
\end{cases}
\]

if \( \epsilon < d_{1,p}(\alpha_{1,m_1}', c') \).

The proof is completed by noting that, \( B(1, \frac{m_2}{q}, i+1) = \Gamma(m_2/q) \Gamma(i/q + 1) / \Gamma((m_2 + i)/q + 1) \) and that all the gamma functions either cancel out or go to 1 as \( q \to \infty \).

A.8 Proof of Theorem 3.13

The expected value of \( f_{\bar{d}_{1,A,p}}(x) \) is

\[
E(f_{\bar{d}_{1,A,p}}(x)) = E \left( \frac{1}{h^n} \int_0^\infty K \left( \frac{e_{\bar{d}_{1,A,p}}(x,y)}{h} \right) \tilde{f}(y) dy \right)
= \frac{1}{h^n} \int_0^\infty K \left( \frac{e_{\bar{d}_{1,A,p}}(x,y)}{h} \right) f(y) dy,
\]

where \( \tilde{f}(y) = 1/n \) at the \( x_i \)'s and 0 elsewhere. This is seen, because \( E(\tilde{f}(y)) = f(y) \). Moreover, the variance of \( f_{\bar{d}_{1,A,p}}(x) \) is

\[
\sigma^2(f_{\bar{d}_{1,A,p}}(x)) = \frac{1}{n} \left[ \frac{1}{h^{2m}} \int_0^\infty K^2 \left( \frac{e_{\bar{d}_{1,A,p}}(X,Y)}{h} \right) f(y) dy - \left( \frac{1}{h^n} \int_0^\infty K \left( \frac{e_{\bar{d}_{1,A,p}}(x,y)}{h} \right) f(y) dy \right)^2 \right],
\]

because \( E(f_{\bar{d}_{1,A,p}}(x)f_{\bar{d}_{1,A,p}}(x)) \) is equal to \( 1/n \) times the first part plus \( (n-1)/n \) times the second part in the square brackets. Equations (13) and (14) are derived by changing variables to \( \alpha \). The bottom equation in Equation (12) is obtained by integrating AMSE\((x)\) over \( x \) and changing variables to \( \alpha \).