# Projection Pursuit via Decomposition of Bias Terms of Kernel Density Estimators

Mark E. Marson[1] and David W. Scott[2]

May 29, 1997

**ABSTRACT:**

Dimension reduction of data, $\Re^d \rightarrow \Re^p$ $(p << d)$, to be used for clustering has specific requirements that are not generally met by generic dimension reduction algorithms such as principal components. Projection pursuit, on the other hand, has a growing variety of criteria that target holes, skewness, etc., using information measures, density functionals, sample moments, etc. With the exception of $q = 3$ by Nason (19xx), the target dimension is only one or two.

In this paper, we focus on features of a multivariate kernel density estimator to find a sequence of basis vectors defining our subspace. We begin by deriving two new variations on the formula for obtaining the asymptotic integrated squared bias (AISB) of a general multivariate kernel density estimator. These formulae are then used to compute orthogonal projections of data onto lower dimensional subspaces in which as much as possible of the AISB is retained. The methods we use for dimension reduction favor those subspaces on which the clustering information of the data is prominent. The methods are not of the

[1]Mark E. Marson is Cryptologic Mathematician, National Security Agency, 9800 Savage Road, Fort George G. Mead, MD 20755.

[2]David W. Scott is Professor, Department of Statistics, Rice University, Houston, TX 77251. This research was supported in part by the National Science Foundation under grant DMS-96-15912 and the National Security Agency under contract MOD ??????.

"least normal" variety, since the normal density is not the "easiest" estimable density, but rather a certain low order polynomial density.

KEY WORDS: Projection Pursuit, Dimension Reduction, Kernel Density Estimation, Asymptotic Integrated Squared Bias.

# 1 Introduction

For a multivariate histogram with regular cubical mesh of volume $h^d$, the asymptotic mean integrated squared error (AMISE) is the sum of the asymptotic integrated variance (AIV) and the asymptotic integrated squared bias (AISB):

$$\mathrm{AMISE}(h) = \mathrm{AIV}(h) + \mathrm{AISB}(h) = \frac{1}{nh^d} + \frac{1}{12}h^2 \int_{\Re^d} \nabla f(\mathbf{x})^T \nabla f(\mathbf{x}) d\mathbf{x}, \tag{1}$$

where $\nabla f(\mathbf{x})$ denotes the gradient of the density. (See Scott, 1992). In 1985 Terrell proposed a method called informative components analysis (ICA), which chooses new coordinates by decomposing the AISB term of the AMISE. The informative components are simply the eigenvectors of the expectation of the Hessian of the pdf $f$. This is easily seen to hold since $\nabla f(\mathbf{x})^T \nabla f(\mathbf{x}) = \mathrm{tr}\left[\nabla f(\mathbf{x})^T \nabla f(\mathbf{x})\right] = \mathrm{tr}\left[\nabla f(\mathbf{x}) \nabla f(\mathbf{x})^T\right]$. Integrating by parts gives

$$\int_{\Re^d} \nabla f(\mathbf{x})^T \nabla f(\mathbf{x}) d\mathbf{x} = -\mathrm{tr}\left[\int_{\Re^d} \nabla^2 f(\mathbf{x})\, f(\mathbf{x}) d\mathbf{x}\right] = -\mathrm{tr}\left[\mathrm{E}\left(\nabla^2 f(\mathbf{X})\right)\right],$$

where $\nabla^2 f(\mathbf{X})$ denotes the Hessian of the density at $\mathbf{x}$. Since $\mathrm{tr}\,(A)$ is the sum of its eigenvalues for $A$ symmetric, the eigenvectors associated with the largest eigenvalues contribute the most to the bias term. Those eigenvectors define the ICA subspace.

The ICA coordinates were renamed principle Hessian directions (PHD) when they were proposed independently by Li in 1992 [ ]. One interpretation of this method is that it is a particular measure of the roughness of the data. Specifically, they choose coordinates by finding those directions $v$ in $\Re^d$ such that

$$\int_{\Re^d} |v^T \nabla f(\mathbf{x})|^2 d\mathbf{x} = \|f_v\|_2^2 \tag{2}$$

is large.

In this paper we derive analogous formulae for general multivariate kernel density estimators. In particular, we also decompose the AISB term of the AMISE to choose new coordinates for the data. One can interpret our method as finding those directions $v$ in $\Re^d$ for which

$$\int_{\Re^d} \left(v^T \nabla^2 f(\mathbf{x}) v\right)^2 d\mathbf{x} = \|v^T \nabla^2 f v\|_2^2 \tag{3}$$

is large.

We feel (3) is a more intuitive, natural, and powerful measure of roughness than (2), although both methods often focus on similar features in the data. Also, the method presented in this paper to estimate the AISB of a kernel density estimator is identical to the one suggested by Scott in [ ] for computing informative components. Li offers a different method of computation, which may also be applicable to our situation here. (expand?)

## 2 AISB Calculation

In this section we derive an expression for computing the AISB of a general multivariate kernel density estimator. This formula is then used as the basis for our dimension reduction strategies.

Let $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \Re^d$, be samples from a pdf $f$. We consider general multivariate kernel density estimators of $f$ with the form

$$\hat{f}(\mathbf{x}) = \frac{1}{n|H|} \sum_{i=1}^{n} K\left(H^{-1}(\mathbf{x} - \mathbf{x}_i)\right), \tag{4}$$

where $H$ is a $d \times d$ nonsingular matrix and $K : \Re^d \to \Re^1$ a multivariate kernel satisfying the moment conditions

$$\begin{aligned} \int_{\Re^d} K(\mathbf{w}) d\mathbf{w} &= 1, \\ \int_{\Re^d} \mathbf{w} K(\mathbf{w}) d\mathbf{w} &= \mathbf{0}, \\ \int_{\Re^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w}) d\mathbf{w} &= I_d. \end{aligned} \tag{5}$$

One can show (Scott, 1992) that the asymptotic integrated variance (AIV) of (4) is

$$\mathrm{AIV}(h) = \frac{R(K)}{nh^d}, \tag{6}$$

where

$$R(K) = \int_{\Re^d} K(\mathbf{x})^2 d\mathbf{x}. \tag{7}$$

We are currently concerned, however, with computing the AISB.

By making the change of variables $\mathbf{w} = H^{-1}(\mathbf{x}-\mathbf{y})$, and using first two moment conditions and the Taylor expansion of $f$ to second order, the expectation of $\hat{f}$ can be approximated by

$$\begin{aligned}
\mathrm{E}\hat{f}(\mathbf{x}) &= \frac{1}{|H|}\int_{\Re^d} K\left(H^{-1}(\mathbf{x}-\mathbf{y})\right) f(\mathbf{y})d\mathbf{y} \\
&= \int_{\Re^d} K(\mathbf{w})f(\mathbf{x}-H\mathbf{w})d\mathbf{w} \\
&= \int_{\Re^d} K(\mathbf{w})\left[f(\mathbf{x}) - \mathbf{w}^T H\nabla f(\mathbf{x}) + \frac{1}{2}\mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H\mathbf{w}\right] d\mathbf{w} \\
&= f(\mathbf{x}) + \frac{1}{2}\int_{\Re^d} K(\mathbf{w})\mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H\mathbf{w}d\mathbf{w} \qquad (8)
\end{aligned}$$

Using the trace formula tr$\{AB\}$=tr$\{BA\}$ and the last moment condition in (5), the integral in the last line of (8) simplifies to

$$\begin{aligned}
\frac{1}{2}\int_{\Re^d} K(\mathbf{w})\mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H\mathbf{w}d\mathbf{w} &= \frac{1}{2}\int_{\Re^d} \mathrm{tr}\{K(\mathbf{w})\mathbf{w}^T H^T \nabla^2 f(\mathbf{x}) H\mathbf{w}\}d\mathbf{w} \\
&= \frac{1}{2}\int_{\Re^d} \mathrm{tr}\{\mathbf{w}\mathbf{w}^T K(\mathbf{w}) H^T \nabla^2 f(\mathbf{x}) H d\mathbf{w}\} \\
&= \frac{1}{2}\mathrm{tr}\left\{\int_{\Re^d} \mathbf{w}\mathbf{w}^T K(\mathbf{w})d\mathbf{w} \cdot H^T \nabla^2 f(\mathbf{x}) H\right\} \\
&= \frac{1}{2}\mathrm{tr}\left\{H^T \nabla^2 f(\mathbf{x}) H\right\}. \qquad (9)
\end{aligned}$$

Now, define $A$ to be the $d \times d$ matrix and $h > 0$ the scalar such that

$$H = hA \quad \text{where} \quad |A| = 1. \qquad (10)$$

Let $v_i$ denote the $i^{\underline{th}}$ column of $A$. By combining (8) and (9), the pointwise bias of $\hat{f}(\mathbf{x})$ can then be written as

$$\mathrm{Bias}\{\hat{f}(\mathbf{x})\} = \mathrm{E}\hat{f}(\mathbf{x}) - f(\mathbf{x}) = \frac{1}{2}h^2 \sum_{i=1}^{n} v_i^T \nabla^2 f(\mathbf{x}) v_i. \qquad (11)$$

By squaring and integrating (11) we get the AISB formula

$$\mathrm{AISB}(h) = \frac{1}{4}h^4 \int_{\Re^d} \left(\sum_{i=1}^{d} v_i^T \nabla^2 f(\mathbf{x}) v_i\right)^2 d\mathbf{x}. \qquad (12)$$

Combining (6) and (12) we therefore make the following expression for the AMISE.

**Proposition 1.** Let $\hat{f}(\mathbf{x})$ be a general multivariate kernel density estimator as in (4). Then the AMISE for $\hat{f}(\mathbf{x})$ is given by

$$\mathrm{AMISE}(h) = \mathrm{AIV}(h) + \mathrm{AISB}(h) = \frac{R(K)}{nh^d} + \frac{1}{4}h^4 \int_{\Re^d} \left(\sum_{i=1}^{d} v_i^T \nabla^2 f(\mathbf{x}) v_i\right)^2 d\mathbf{x}. \qquad (13)$$

We wish to find a convenient expression to calculate the AISB. A typical term of (12) is

$$\begin{aligned}\frac{1}{4}h^4 \int_{\Re^d} \left(v_i^T \nabla^2 f(\mathbf{x}) v_i\right) \left(v_j^T \nabla^2 f(\mathbf{x}) v_j\right)\, d\mathbf{x}& \\ = \frac{1}{4}h^4 v_i^T \left(\int_{\Re^d} \nabla^2 f(\mathbf{x}) v_i v_j^T \nabla^2 f(\mathbf{x})\, d\mathbf{x}\right) v_j.& \qquad (14)\end{aligned}$$

The integrand of (14) is a $d \times d$ matrix with $(k, l)$ entry

$$\left(\sum_{p=1}^{d} f_{kp}(\mathbf{x}) v_{i,p}\right)\left(\sum_{q=1}^{d} v_{j,q} f_{ql}(\mathbf{x})\right) = \sum_{p,q=1}^{d} v_{i,p} v_{j,q} f_{kp}(\mathbf{x}) f_{ql}(\mathbf{x}), \qquad (15)$$

where $v_{i,p}$ denotes the $p^{\underline{th}}$ component of $v_i$, and the second partial $f_{kp}(\mathbf{x}) = \partial^2 f(\mathbf{x})/\partial x_k \partial x_p$. Assuming sufficient regularity of the density function $f$, we can integrate (15) and perform integration by parts twice to obtain

$$\begin{aligned}\sum_{p,q=1}^{d} v_{i,p} v_{j,q} \int_{\Re^d} f_{kp}(\mathbf{x}) f_{ql}(\mathbf{x})\, d\mathbf{x} &= \sum_{p,q=1}^{d} v_{i,p} v_{j,q} \int_{\Re^d} f_{klpq}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} \\ &= \sum_{p,q=1}^{d} v_{i,p} v_{j,q} \left(\mathrm{E}(\nabla^2 f_{kl})\right)_{(p,q)} \\ &= v_i^T \mathrm{E}(\nabla^2 f_{kl}) v_j, \qquad (16)\end{aligned}$$

where $\nabla^2 f_{kl}$ denotes the Hessian of $f_{kl}$ and $\mathrm{E}(\nabla^2 f_{kl})$ its expectation. Therefore, the $(k, l)$ entry of the matrix of integrals in (14) is

$$\left(\int_{\Re^d} \nabla^2 f(\mathbf{x}) v_i v_j^T \nabla^2 f(\mathbf{x})\, d\mathbf{x}\right)_{(k,l)} = v_i^T \mathrm{E}(\nabla^2 f_{kl}) v_j. \qquad (17)$$

The entire matrix of integrals can be expressed as follows. Let $J$ be the $d^2 \times d^2$ block matrix where block $(k, l)$ is $d \times d$ matrix $\mathrm{E}(\nabla^2 f_{kl})$. Recall that if $A$ is a $p \times q$ matrix and $B$ is an $r \times s$ matrix, then the *Kronecker product* of $A$ and $B$, written $A \otimes B$, is the $pr \times qs$ block matrix where block $(i, j)$ is the $r \times s$ matrix $a_{ij} B$. Let $V_i$ denote the $d^2 \times d$ matrix $V_i = I_d \otimes v_i$. Then we can write the matrix of integrals in (14) as

$$\int_{\Re^d} \nabla^2 f(\mathbf{x}) v_i v_j^T \nabla^2 f(\mathbf{x})\, d\mathbf{x} = V_i^T J V_j. \qquad (18)$$

Since $V_i v_i = (I_d \otimes v_i) v_i = v_i \otimes v_i$, we can now write (14) as

$$\frac{1}{4}h^4 v_i^T \left(\int_{\Re^d} \nabla^2 f(\mathbf{x}) v_i v_j^T \nabla^2 f(\mathbf{x})\, d\mathbf{x}\right) v_j = \frac{1}{4}h^4 v_i^T V_i^T J V_j v_j = \frac{1}{4}h^4 (v_i \otimes v_i)^T J (v_j \otimes v_j). \qquad (19)$$

Finally, (19) is only one of the $d^2$ terms in the sum for the AISB. Summing over all the terms gives

$$\text{AISB} = \frac{1}{4}h^4 \sum_{i,j=1}^{d} (v_i \otimes v_i)^T J (v_j \otimes v_j) = \frac{1}{4}h^4 u^T J u \tag{20}$$

where $u = \sum_{i=1}^{d}(v_i \otimes v_i)$.

We have therefore proved the following theorem.

**Theorem 1** (*AISB formula*). Let $\tilde{f}$ be a general multivariate kernel density estimator defined as in (4). Then for a sufficiently smooth density function $f$ the AISB can be calculated by

$$\text{AISB} = \frac{1}{4}h^4 u^T J u,$$

where the matrix $J$ and vector $u$ are defined as above.

We write out the formulae in $\Re^2$ to help illustrate the above theorem. Let $\text{E}(f_{ijkl})$ denote the expectation

$$\text{E}(f_{ijkl}) = \int_{\Re^d} \frac{\partial^4 f(\mathbf{x})}{\partial x_i \partial x_j \partial x_k \partial x_l} f(\mathbf{x}) \mathbf{dx}. \tag{21}$$

Then $J$ is the $4 \times 4$ matrix

$$J = \begin{pmatrix} \text{E}(f_{1111}) & \text{E}(f_{1112}) & \text{E}(f_{1112}) & \text{E}(f_{1122}) \\ \text{E}(f_{1112}) & \text{E}(f_{1122}) & \text{E}(f_{1122}) & \text{E}(f_{1222}) \\ \text{E}(f_{1112}) & \text{E}(f_{1122}) & \text{E}(f_{1122}) & \text{E}(f_{1222}) \\ \text{E}(f_{1122}) & \text{E}(f_{1222}) & \text{E}(f_{1222}) & \text{E}(f_{2222}) \end{pmatrix} \tag{22}$$

and $(v_i \otimes v_i)$ is the $4 \times 1$ vector

$$v_i \otimes v_i = (v_{i,1}^2, v_{i,1}v_{i,2}, v_{i,1}v_{i,2}, v_{i,2}^2)^T. \tag{23}$$

There is clearly a lot of repetition in the above matrix. A simple counting argument shows that, in general, the $d^2 \times d^2$ symmetric matrix $J$ has only $(d^4 + 6d^3 + 11d^2 + 6d)/24$ unique entries. Also, the dimension of the nullspace of $J$ is at least $d(d-1)/2$, since there are that many pairs of identical columns. In general, $J$ is neither positive nor negative semi-definite.

Since the number of unique terms in $J$ grows like $O(d^4)$, its estimation may become expensive when the dimension gets large. Fortunately, we can reduce to $d \times d$ the size of the matrix we need to compute if we assume that $A$ is orthogonal.

**Theorem 2** (*Orthogonal AISB formula*)**.** Let $\tilde{f}$ be a general multivariate kernel density estimator defined as in (4), where $H = hA$ and $A$ is orthogonal. Let $J'$ be the $d \times d$ matrix with $(i, j)$ entry $\mathrm{E}(f_{iijj})$, and $w = (1, 1, \ldots, 1)$. Then the AISB for a sufficiently smooth density function $f$ is

$$\mathrm{AISB} = \frac{1}{4} h^4 w^T J' w = \frac{1}{4} h^4 \sum_{i,j} J'(i, j). \tag{24}$$

Proof. Let $\{v_1, \ldots, v_d\}$ be the columns of $A$, and $\{e_1, \ldots, e_d\}$ the standard basis. Then

$$u = \sum_{i=1}^{d} (v_i \otimes v_i) = (e_1^T, \ldots, e_d^T)^T \tag{25}$$

This is easily seen from the identity $AA^T = I_d$. The theorem then follows from using (25) in the AISB formula in (20). Another proof of (24) is to use the trace formula on (9) to get $\frac{1}{2}\mathrm{tr}(H^T \nabla^2 f(x) H) = \frac{1}{2}\mathrm{tr}(HH^T \nabla^2 f(x)) = \frac{1}{2}\mathrm{tr}(\nabla^2 f(x))$, and then repeat the integration by parts arguments used to derive (20).

**Proposition 2.** The matrix $J'$ defined above is positive definite, with all positive entries.

Proof. The entries of $J'$ are all positive, since

$$\begin{aligned} J'(i,j) &= \int_{\Re^d} \frac{\partial^4 f(\mathbf{x})}{\partial x_i \partial x_i \partial x_j \partial x_j} f(\mathbf{x}) d\mathbf{x} \\ &= \int_{\Re^d} \left( \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right)^2 d\mathbf{x} \\ &> 0. \end{aligned} \tag{26}$$

To see that $J'$ is positive definite, let $\triangle^2 f(\mathbf{x}) = (f_{11}(\mathbf{x}), \ldots, f_{dd}(\mathbf{x}))^T$ and note that

$$J' = \int_{\Re^d} (\triangle^2 f(\mathbf{x}))(\triangle^2 f(\mathbf{x}))^T d\mathbf{x}. \tag{27}$$

Then for any nonzero vector $v$ in $\Re^d$,

$$\begin{aligned} v^T J' v &= v^T \left( \int_{\Re^d} (\triangle^2 f(\mathbf{x}))(\triangle^2 f(\mathbf{x}))^T d\mathbf{x} \right) v \\ &= \int_{\Re^d} v^T (\triangle^2 f(\mathbf{x}))(\triangle^2 f(\mathbf{x}))^T v \, d\mathbf{x} \\ &= \int_{\Re^d} \left( (\triangle^2 f(\mathbf{x}))^T v \right)^2 d\mathbf{x} \\ &> 0. \end{aligned} \tag{28}$$

Hence $J'$ is positive definite.

In the next section we discuss methods for estimating $J$ and $J'$. For now we are concerned with using these matrices to find interesting projections of the data onto lower dimensional subspaces. The basic strategy is to remove those coordinates which contribute least to the AISB. The remaining coordinates are the ones that contain the important features the data. In particular, our methods tends to favor those coordinates in which the clustering information of the data is prominent.

Consider the diagonal terms of the AISB in (13). These are the terms of the form $(v_i \otimes v_i)^T J(v_i \otimes v_i)$. Using the same arguments in the AISB derivation, it can be seen that

$$\begin{aligned}(v_i \otimes v_i)^T J(v_i \otimes v_i) &= \int_{\Re^d} \left(v_i^T \nabla^2 f(\mathbf{x}) v_i\right) \left(v_i^T \nabla^2 f(\mathbf{x}) v_i\right) \, d\mathbf{x} \\ &= \|v_i^T \nabla^2 f v_i\|_2^2. \end{aligned} \tag{29}$$

Therefore, the diagonal terms of the AISB are a measure the roughness of the density function along the directions $v_i$. Our premise is that those directions in which (29) is large will contain the most interesting features of the data. Conversely, the data are relatively flat and uninteresting in those directions for which (29) is small.

To reduce the dimension of the data with $J$, we remove the uninteresting dimensions one at a time. One option is to use a numerical minimizer to find the unit vector $v_d$ which minimizes $(v \otimes v)^T J(v \otimes v)$. We then proceed recursively, finding the unit vector $v_i$ which is orthogonal to vectors $\{v_{i+1}, \ldots, v_d\}$ and minimizes $(v \otimes v)^T J(v \otimes v)$. The data are then reduced in dimension by projecting it onto the subspace spanned by $\{v_1, \ldots, v_{d_1}\}$. Another option is to successively project the data onto hypersurfaces normal to the unit vector which minimizes $(v \otimes v)^T J(v \otimes v)$, and then re-estimate $J$ for the projected data.

The strategies above choose the new coordinates to be orthogonal, but this is not required by Theorem 1. Also, these methods do not take the cross terms of the AISB directly into account. In particular, the portion of the AISB used to choose the coordinates is

$$\sum_{i=1}^{d} v_i J v_i = \mathrm{AISB} - \sum_{i \neq j} v_i J v_j. \tag{30}$$

It might be an interesting exercise to find other criteria which completely decompose the AISB, or which use nonorthogonal coordinates.

The above methods work well when the dimension of the data is not too large, say $d \leq 10$. When the dimension gets much larger, however, these methods become less useful. As previously mentioned, estimating $J$ becomes expensive. Also, numerical minimization routines become much less effective and reliable as the dimensions increases, either not converging at all, or converging to local minima. For larger dimensions, we use Theorem 2 and $J'$ for dimension reduction.

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$ denote the eigenvalues of $J'$. Then from Theorem 2 and the trace identity $\operatorname{tr}(J') = \sum_{i=1}^{d} \lambda_i$,

$$\text{AISB} = \sum_{i,j=1}^{d} J'(i,j) = \sum_{i=1}^{d} \lambda_i + \sum_{i \neq j} J'(i,j). \tag{31}$$

In practice, the contribution of the eigenvalues to the total AISB is much larger than that of the off-diagonal terms. Therefore, the eigenvectors of $J'$ are a natural choice for the new coordinates for the data. Those coordinates corresponding to the smaller eigenvalues are then discarded.

Since $J'$ only has $d(d+1)/2$ unique terms, its estimation in large dimensions is much less expensive than that of $J$. It is therefore useful for removing at once many of the "noise" dimensions from high dimensional data. On the other hand, in lower dimensions the methods which use $J$ often find subtle features of the data that are missed by the eigenvectors of $J'$. This suggests a strategy of using $J'$ to first reduce the dimension of the data, and then estimating and using $J$ on the projected data if further reduction is desired.

The methods above can be considered "greedy," in that they recursively choose the best (or eliminate the worst) coordinates one at a time. This is obviously the case when we use the minimizer on the full matrix $J$, but can also be considered true of the eigenvectors of $J'$. In particular, the unit eigenvector $v_{i+1}$ corresponding to the eigenvalue $\lambda_{i+1}$ is the vector orthogonal to $v_1, \ldots, v_i$ which maximizes $v^T J' v$.

It is therefore natural to ask whether the projections found this way are the "best" possible ones. The simple example in the next section shows this is not always the case. Since our algorithm orders the coordinates by the roughness along each direction separately, the projections of the data onto the subspaces spanned by these coordinates can sometimes be surprising. Still, these projections usually offer interesting information about the data.

This raises the related question of measuring the overall roughness of the projected data, rather than along one direction at a time. Let $S$ denote a $d_1$ dimensional subspace of $\Re^d$, and $P = [p_1, \cdots, p_{d_1}]$ and $Q = [q_1, \cdots, q_{d_2}]$ orthonormal bases for $S$ and $S^\perp$ respectively. The marginal of the density function $f$ on $S$ is

$$f_S(\mathbf{p}) = \int_S f(\mathbf{x}) d\sigma(\mathbf{x}) = \int_{\Re^{d_2}} f\Big([P, Q] \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}\Big) d\mathbf{q}. \tag{32}$$

The projection of the data points onto $S$ is given by $\mathbf{y}_i = P^T \mathbf{x}_i$, and the corresponding kernel density estimator is then

$$\hat{f}_S(\mathbf{p}) = \frac{1}{nh^d} \sum_{i=1}^{n} K\Big(\frac{\mathbf{p} - \mathbf{y}_i}{h}\Big). \tag{33}$$

Repeating the arguments of this section gives the pointwise bias of (33),

$$\text{Bias}\{\hat{f}_S(\mathbf{p})\} = \frac{1}{2} h^2 \sum_{i=1}^{n} \frac{\partial^2 f_S(\mathbf{p})}{\partial p_i \partial p_i}, \tag{34}$$

and the AISB,

$$\text{AISB}_S(h) = \frac{1}{4} h^4 w^T \Big( \int_{\Re^{d_1}} f_S(\mathbf{p}) J_S(\mathbf{p}) d\mathbf{p} \Big) w, \tag{35}$$

where $w = (1, 1, \cdots, 1)^T$ and $J_S(\mathbf{p})$ is the $d_1 \times d_1$ matrix with $(i, j)$ entry given by $\partial^4 f_S(\mathbf{p}) / \partial p_i^2 \partial p_j^2$.

Our conjecture is that (35) is a better measure of bias for dimension reduction than (20) or (24), and determining the subspace $S$ which maximizes it would lead to interesting and useful projections of the data. Unfortunately, estimating (35) and optimizing over all possible subspaces appears to be a very difficult and expensive problem. No experiments have been run yet to test our hypothesis.

# 3 Implementation

In this section we present a method for estimating the fourth derivative expectations in (21). It is based on the method of cross validation. While we make no claims that our estimates are optimal, they are easy to implement and usually quite effective.

The data are first preprocessed through principal components, which may be used to eliminate redundant variables, leaving the data normalized by "sphering." Let $\hat{f}$ be the normal product kernel estimator for the density function $f$,

$$\hat{f}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} h^d n} \sum_{i=1}^{n} e^{-||\mathbf{x}-\mathbf{x}_i||^2/2h^2}. \tag{36}$$

The corresponding leave-one-out estimator where the data point $\mathbf{x}_m$ is omitted is

$$\hat{f}_{-m}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} h^d (n-1)} \sum_{\substack{i=1 \\ i \neq m}}^{n} e^{-||\mathbf{x}-\mathbf{x}_i||^2/2h^2}. \tag{37}$$

We then estimate (21) by computing $\partial^4 \hat{f}_{-m}(\mathbf{x})/\partial x_i \partial x_j \partial x_k \partial x_l$, evaluating at the omitted point, and averaging the results. This gives the estimate

$$\begin{aligned} \widehat{\mathrm{E}} \frac{\partial^4 f}{\partial x_i \partial x_j \partial x_k \partial x_l} &= \frac{1}{n} \sum_{m=1}^{n} \frac{\partial^4 \hat{f}_{-m}(\mathbf{x}_m)}{\partial x_i \partial x_j \partial x_k \partial x_l} \\ &= \frac{2}{(2\pi)^{d/2} h^d n(n-1)} \sum_{m=1}^{n} \sum_{p=m+1}^{n} \frac{\partial^4 e^{-||\mathbf{x}_m-\mathbf{x}_p||^2/2h^2}}{\partial x_i \partial x_j \partial x_k \partial x_l}. \end{aligned} \tag{38}$$

If $\mathbf{X}$ denotes the random vector corresponding to the distribution from which the data samples are drawn, then each term $\partial^4 \hat{f}_{-m}(\mathbf{x}_m)/\partial x_i \partial x_j \partial x_k \partial x_l$ is a consistent estimator for

$$\mathrm{E}_{\mathbf{X}} \Big( \int_{\Re^d} \frac{\partial^4 \hat{f}(\mathbf{x})}{\partial x_i \partial x_j \partial x_k \partial x_l} f(\mathbf{x}) \, d\mathbf{x} \Big); \tag{39}$$

see the Appendix. Averaging over all the leave-one-out estimators helps smooth over anomalies due to outliers in the data. We therefore expect (38) to be a reasonable estimate for (21).

While the estimate (38) of course depends on the bandwidth $h$, the coordinates chosen using the methods described in the previous section tend to vary both slowly and smoothly as $h$ varies. Good estimates are obtained for a wide range of bandwidths. Since the AISB is a measure of the overall roughness of the density function, we would like to minimize the amount of artificial roughness introduced by its estimation. For this reason we recommend using bandwidths which oversmooth, rather than undersmooth, the estimates. A useful initial choice for the bandwidth is Scott's rule (for sphered data),

$$h = n^{-1/(d+4)}. \tag{40}$$

# 4 Examples

In this section, we present a small selection of examples that compare principal components, ICA, and PPICKER. We note that PPICKER (pronounced p-picker) contains PP=projection pursuit, IC=informative components, and KER=kernel. Indeed we are picking an appropriate subspace of $p$ variables. Our general approach is to examine the original variables and possibly apply simple transformations as desired. Then principal components are examined and highly correlated variables dropped. The remaining variables are re-expressed in principal component space (i.e., are "sphered"). Thus all second-order information has been taken account of and the sample covariance matrix of any new orthonormal projection will be the identity matrix.

The first example (eightball) consist of 8 quadravariate normal clusters with $\sigma = 1$ and $n = 100$ for each cluster, which are centered at $(\pm 4, \pm 4, \pm 4, \pm 4)$. Eight of the 16 possible centers are eliminated by the additional constraint that the third and fourth coordinates of the center point are identical (in sign).

The data were rotated using ICA and PPICKER; see Figures 1 and 2, respectively. Note how the first dimensions of ICA prefer a few large clusters in the projected space (so as to maximize $\int f'(\cdot)^2$), where as PPICKER attempts to find as many bumps as possible (so as to maximize $\int f''(\cdot)^2$). This observation is confirmed by examining density estimates (averaged shifted histogram, Scott (1985)) of the data expressed on the 4 (ordered) ICA and PPICKER dimensions; see Figures 3 and 4, respectively. ICA's behavior is reminiscent of the original Friedman and Tukey (1974) projection pursuit algorithm which Huber (198?) later showed was essentially trying to maximize $\int f(\cdot)^2$.

Recall that ICA finds the bias maximizing subspace of dimension $p$ "exactly" since the total bias has the exact eigenvalue decomposition. PPICKER, on the other hand in simpler implementation, uses the eigenvalue decomposition on the smaller matrix $J'$ (much smaller than $J$), which provides only an approximation to the total bias, omitting the off-diagonal terms in $J$. Hence, the first two PPICKER dimensions do not find all eight clusters; however, PPICKER is doing an interesting job of finding structure in the first several dimensions. Principal components does not show anything???

Our final example uses the $n = 500$ data points in the PRIM9 data discussed by Friedman and Tukey and found in the Splus system. We eliminated the ninth variable (discrete with even integer values) and retained the remaining eight variables, which are all continuous when examined one-by-one. In the upper left frame of Figure 5, the first two principal components of the data are displayed, and reveal two elongated clusters. Both ICA and PPICKER find at least six clusters in the first two dimensions. Note that the first dimension of ICA strongly superimposes the six bumps into two, while the first dimension of PPICKER arranges the six clusters on the diagonal to reveal six small bumps. The addition of a second variable in each case is especially beneficial. Finally, searching the remaining six principal coordinate dimensions revealed that the second and eighth (!!) variables showed the six clusters. We have performed many smaller examples and larger, but the lessons are similar to what we have briefly recorded here. Details such as the choice of smoothing parameter were handled by using Scott's simple rule $n^{-1/(d+4)}$ with occasional perturbation to investigate sensitivity to the roughness functional estimates.

# 5 Appendix

We outline the proof that our leave-one-out estimator for the roughness constants is consistent. The multivariate case is only slightly more difficult that the univariate case, so we only give details there. Suppose we attempt to estimate $R(f'')$ by using the identity $\int f''(x)^2 dx = \mathrm{E}\left[f^{iv}(X)\right]$ and the leave-one-out estimator

$$\widehat{R(f'')} = \frac{1}{n}\sum_{k=1}^{n}\frac{1}{(n-1)h^5}\sum_{i\neq k} K^{iv}\left(\frac{x_k - x_i}{h}\right).$$

Then the expected value of our estimator is given by

$$\mathrm{E}\widehat{R(f'')} = \frac{1}{h^5}\int_y\int_x K^{iv}\left(\frac{x-y}{h}\right) f(x)f(y)\,dx\,dy = \frac{1}{h^5}\int_y\int_w K^{iv}(w)f(y+hw)f(y)\,h\,dw\,dy.$$

Taking a Taylor's series of $f(y+hw)$ around $y$ up to order 6, noting that $\int w^k K^{iv}(w)dw$ vanishes for $k < 4$ and $k$ odd (assuming $K$ is symmetric and sufficiently smooth), then the inner integral over $w$ only has the terms

$$\frac{1}{24}h^4 f^{iv}(y)\int_w w^4 K^{iv}(w)\,dw + \frac{1}{720}h^6 f^{vi}(y)\int_w w^6 K^{iv}(w)\,dw.$$

Integration by parts yields $\int w^4 K(w)\,dw = 24$ and $\int w^6 K(w)\,dw = 360\sigma_K^2$ so that

$$\mathrm{E}\widehat{R(f'')} = \frac{1}{h^4}\int_y \left[h^4 f^{iv}(y) + \frac{1}{2}\sigma_K^2 h^6 f^{vi}(y)\right] f(y)\,dy = \int f''(y)^2\,dy - \frac{1}{2}\sigma_K^2 h^2 \int f'''(y)^2\,dy$$

after integrating $\int f^{iv}(y)f(y)\,dy$ and $\int f^{vi}(y)f(y)\,dy$ by parts. Thus the asymptotic bias equals $-\frac{1}{2}\sigma_K^2 h^2 \int f'''(y)^2\,dy$.

The variance of our estimator is relatively straightforward, once a proper count of terms is obtained. Specifically, $n^2(n-1)^2 h^{10} \times \mathrm{Var}\widehat{R(f'')}$ equals

$$\left[2n(n-1)\,\mathrm{Var}K^{iv}\left(\frac{x_k - x_i}{h}\right) + 4n(n-1)(n-2)\,\mathrm{Cov}\left(K^{iv}\left(\frac{x_k - x_i}{h}\right), K^{iv}\left(\frac{x_k - x_j}{h}\right)\right)\right]$$

(where different subscripts $i, j, k$ are taken to be distinct) and noting that

$$\mathrm{Cov}\left(K^{iv}\left(\frac{x_k - x_i}{h}\right), K^{iv}\left(\frac{x_\ell - x_j}{h}\right)\right) = 0$$

by independence of the four (distinct) random variables. Similar calculations performed using a Taylor's series shows that

$$\mathrm{Var}\,K^{iv}\left(\frac{x_k - x_i}{h}\right) \approx hR(K^{iv})\int f(y)^2 dy$$

and

$$\mathrm{Cov}\left(K^{iv}\left(\frac{x_k - x_i}{h}\right), K^{iv}\left(\frac{x_k - x_j}{h}\right)\right) \approx h^2 R(K^{iv})\int f(y)^3 dy.$$

Combining gives

$$\mathrm{MSE}\left[\widehat{R(f'')}\right] \approx \frac{4R(K^{iv})\int f(y)^3 dy}{nh^8} + \frac{1}{4}\sigma_K^4 h^4 R(f''')^2,$$

for which the optimal choice of bandwidth is $O(n^{-1/12})$. For all unknown quantities being standard normal, $h^* = 1.28n^{-1/12}$. In general, MSE $= O(n^{-1/3})$, proving consistency of our estimator. Alternatively, plug-in estimates of the roughness quantities can be considered (Hall and Marron, 1988), but we have elected to go with the leave-one-out estimators based subjectively on our collective experience. The corresponding multivariate roughness results are quite similar. For example, the bias of the multivariate product kernel estimator of $\int_{\mathbf{y}} [\partial^4 f(\mathbf{y})/\partial x_i^4]^2\,d\mathbf{y}$ has the bias term $\frac{1}{2}\sigma_K^2 h^2 \int_{\mathbf{y}} f_{iii}(\mathbf{y})^2 d\mathbf{y}$ and additional terms of the same order containing integrals such as $\int_{\mathbf{y}} f_{iij}(\mathbf{y})^2 d\mathbf{y}$ for $j \neq i$.

## References


Friedman, J.L. and Tukey, J.W. (1974). "PP"

Hall, P.?. and Marron, S.?. (1988). "Estimating Functional.." Prob Stat Letters.

Huber, P.?. (1985?). "PP" AOS.

Li, K.-C. (1992). "Sliced Inverse Regression...???"

Nason, G. (19xx) Projection Pursuit...

Scott, D.W. (1985). "Averaged Shifted Histogram," AOS.

Scott, D.W. (1992). *Multivariate Density Estimation:...*

Terrell, G.R. (1985). "ICA..."

XGOBI group....SIAM paper.