L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Multivariate density estimation by probing depth

Ricardo Fraiman

Universidad de la Republica, Montevideo, Uruguay

Regina Y. Liu

Rutgers University, Piscataway, USA

Jean Meloche

The University of British Columbia, Vancouver, Canada

Abstract: In high dimension, the estimation of a density is difficult because the observed data gets increasingly sparse with the dimension. This is known as the curse of dimensionality. For that reason, in high dimension, universally consistent estimators such as the kernel density estimator are not practical. In this paper, we consider a class of multivariate densities, within which a density function f can be expressed as $f = g \circ D$ for some given notion of data depth D and some real function g. We propose a density estimator which is shown to be consistent within the class, and it converges at the same rate as the *univariate* kernel density estimator.

Key words: Multivariate density estimation, Data depth, rate of convergence.

AMS subject classification: 62G07, 62H12.

1 Introduction

Let $X_1, \ldots, X_n$ be an i.i.d. sample from an unknown density $f : \mathcal{R}^p \to [0, \infty)$. When p is large, a kernel density estimator of the form

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^p} K\left(\frac{x - X_i}{h}\right)$$

where h is the smoothing parameter, is impractical because of the extremely large number of observations needed to "fill" the p–dimensional space in order to ensure sufficient observations in each "bin" of the kernel. This phenomenon of sparsity of data in high dimensional space is referred to as the "curse of dimensionality" in Bellman (1961).

Many approaches have been developed in the literature in order to address this problem, and in particular projection pursuit techniques (Friedman, Stuelzle and Schroeder, 1984). Projection pursuit avoids this problem by working in low-dimensional linear projections. However, as pointed by Huber (1985), projection pursuit is poorly suited to deal with highly nonlinear structures. For these reasons, the analysis of high-dimensional data sets is often made under some additional restrictions. One common practice is to assume that the density f belongs to some parametric family, so the estimation of the density amounts to the estimation of finitely many parameters. For example, if the underlying density is normal, then one only needs to estimate the mean and the variance. In the same spirit but without the firm grip of parametric assumptions, the so-called "tailor-design density estimates" (cf. Devroye, 1987) are designed to perform well for a particular class of densities. This class can but does not have to be parametric. In general "tailor-design density estimates" are not *universally* consistent, since they are tailored to suit a specific target class of densities. A typical example is the Grenander estimator (Grenander, 1956) which concerns only monotone densities.

In this paper we rely on the general nonparametric smoothing principle to provide a multivariate density estimator, with the idea of enlarging the *neighbourhood* for smoothing so as to include sufficiently many data points even when the dimension is high. Roughly speaking, our approach may be viewed as a generalized version of the following simple nonparametric density estimator

$$\hat{f}_h(x; A_h) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \in A_h(x))}{m_p(A_h(x))}$$

where $\mathbf{1}$ denotes the indicator function, m_p denotes the p–dimensional Lebesgue measure and $A_h(x) = \{t \ni \|x - t\| < h\}$. The above estimate takes advantage of the smoothness of the unknown density f, and assumes that f is nearly constant in the neighbourhood $A_h(x)$. As indicated above, the difficulty with this approach in p dimensions is that the volume of the neighbourhood $A_h(x)$ decreases rapidly with p. As a result, the variance of the estimator increases rapidly with p and one is forced to increase the bandwidth h to obtain a balance between the variance and the squared bias of the estimator. On the other hand, $A_h(x)$ is not the only set over which

f can be presumed constant. The density f is also nearly constant over the larger set $B_h(x) = \{t \ni \|f(x) - f(t)\| < h\}$. Used as a neighbourhood, $B_h(x)$ has a large volume even in high dimension, so one is not forced to use a large bandwidth to obtain a balance between the variance and the squared bias. In contrast with the neighbourhood $A_h(x)$ which does not depend on f, the neighbourhood $B_h(x)$ does and needs to be estimated. Of course, the estimation of the neighbourhood $B_h(x)$ is difficult and may very well offset the improvement resulting from enlarged neighbourhoods.

Our estimator is based on a *neighbourhood* which is between the above two extreme cases corresponding to the sets $A_h(x)$ and $B_h(x)$. Assume that $f = g \circ D$ for some $g : \mathcal{R} \to [0, \infty)$ and some transformation $D : \mathcal{R}^p \to [0, \infty)$ that may depend on f. Under this restriction, f is constant whenever D is constant so that f is nearly constant over the set $C_h(x) = \{t \ni \|D(x) - D(t)\| < h\}$. The estimator we propose essentially amounts to using the set $C_h(x)$ as the neighbourhood.

In recent years, the class of ellipsoidal densities $f = g \circ D$ with $D(x) = (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$ for some function g has received considerable attention because it enables an analysis of multivariate data which does not rely on the validity of the classical multivariate normal theory. The estimator we propose goes hand in hand with such developments, providing an estimate of the density that outperforms other density estimators within that class. The class of ellipsoidal densities is one example among other possible generalisations of the classical multivariate normal family. Various possibilities will be discussed in what follows.

Let m_d denote the d–dimensional Lebesgue measure. Throughout the paper, we will assume that the measure $m_p \circ D^{-1}$ is absolutely continuous with respect to m_1 and $\mathcal{L}_D$ will denote the Radon Nykodym derivative of $m_p \circ D^{-1}$ with respect to m_1. Under this assumption,

$$\Pr\{D(X_1) \in A\} = \int \mathbf{1}\{D(t) \in A\}\, g(D(t))\, dt = \int_A g(r)\, \mathcal{L}_D(r)\, dr$$

showing that $D(X_1)$ has a density, which we will denote f_D, and that $f_D = g\mathcal{L}_D$. The relationship $f_D = g\mathcal{L}_D$ can also be expressed as

$$f(x) = g(D(x)) = \frac{f_D(D(x))}{\mathcal{L}_D(D(x))}.$$

Thus, if D is a known transformation, we can estimate $f(x)$ by

$$\tilde{f}(x; D, h) = \frac{\hat{f}_D(D(x))}{\mathcal{L}_D(D(x))} = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(D(x) - D(X_i))}{\mathcal{L}_D(D(x))}.$$

Example 1 *Consider the case of a spherical density f. This assumption is equivalent to the assumption $f = g \circ D$ for $D(x) = \|x\|^2$ and some $g : \mathcal{R} \to \mathcal{R}$. Here, D is a fixed transformation and does not need to be estimated. It is easy to see that*

$$\mathcal{L}_D(r) = \frac{\pi^{p/2}}{\Gamma(p/2)} r^{p/2-1}$$

is the Radon Nykodym derivative of $m_p \circ D^{-1}$ with respect to m_1. Thus, in the case of a spherical density f, we propose the estimator

$$\tilde{f}(x; D, h) = \frac{\Gamma(p/2)}{\pi^{p/2}} (D(x))^{1-p/2} \frac{1}{n} \sum_{i=1}^{n} K_h(D(x) - D(X_i)).$$

The estimator $\tilde{f}(x; D, h)$ basically amounts to an estimator of the univariate density f_D so that we would expect a one-dimensional nonparametric rate of convergence to $f(x)$ for $\tilde{f}(x; D, h)$.

In general, D could depend on f and would then need to be estimated. If $\hat{D}$ is an estimator of D, we can estimate $f(x)$ by

$$\tilde{f}(x; \hat{D}, h) = \frac{\hat{f}_{\hat{D}}(\hat{D}(x))}{\mathcal{L}_{\hat{D}}(\hat{D}(x))} = \frac{1}{n} \sum_{i=1}^{n} \frac{K_h(\hat{D}(x) - \hat{D}(X_i))}{\mathcal{L}_{\hat{D}}(\hat{D}(x))}.$$

Example 2 *Consider the case of an ellipsoidal density f. This assumption is equivalent to the assumption $f = g \circ D$ for $D(x) = (x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)$ and some $g : \mathcal{R}^p \to \mathcal{R}$. Here, D depends on unknown parameters μ and Σ that need to be estimated. It is easy to see that*

$$\mathcal{L}_D(r) = \sqrt{|\Sigma|} \frac{\pi^{p/2}}{\Gamma(p/2)} r^{p/2-1}.$$

is the Radon Nykodym derivative of $m_p \circ D^{-1}$ with respect to m_1. Thus, in the case of an ellipsoidal density f we propose the estimator

$$\tilde{f}(x; \hat{D}, h) = \frac{\Gamma(p/2)}{\sqrt{|\hat{\Sigma}|} \pi^{p/2}} (\hat{D}(x))^{1-p/2} \frac{1}{n} \sum_{i=1}^{n} K_h(\hat{D}(x) - \hat{D}(X_i))$$

where $\hat{D}(x) = (x - \hat{\mu})^{\mathrm{T}} \hat{\Sigma}^{-1} (x - \hat{\mu})$ and $\hat{\mu}$ and $\hat{\Sigma}$ are some estimates of μ and Σ. The transformation D involves parameters μ and Σ for which there exist estimates converging at speed $1/\sqrt{n}$. Since the rate of convergence for $\hat{D}$ is less than $1/\sqrt{n}$, we expect the asymptotic behaviour of $\tilde{f}(x; \hat{D}, h)$ to be unaffected by the estimation of D and a one-dimensional nonparametric rate of convergence to $f(x)$ for $\tilde{f}(x; \hat{D}, h)$.

Stute and Werner (1991) focus on this last case and propose an estimate of the density based on the above formula, using independent estimates of μ and Σ (based on the data of some preliminary sample).

The above two examples are unusual in that an explicit expression for $\mathcal{L}_D$ is available. When deriving exact expressions for $\mathcal{L}_D$, the following relationships are useful. Let D be some depth. If $D_{\mu,\Sigma}(x) = D((x-\mu)^T\Sigma^{-1}(x-\mu))$, we have $\mathcal{L}_{D_{\mu,\Sigma}}(r) = \sqrt{|\Sigma|}\mathcal{L}_D(r)$. If ρ is a monotone transformation and if $D_\rho(x) = \rho(D(x))$, we have $\mathcal{L}_{D_\rho}(r) = (\mathcal{L}_D/|\rho'|)(\rho^{-1}(r))$. Nevertheless, an explicit expression for $\mathcal{L}_D$ is usually not available and an approximation for the denominator must be used. Note that provided $\mathcal{L}_D$ is smooth,

$$\begin{aligned}\overline{\mathcal{L}}_D(D(x)) &= \int K_{h_f}(D(x) - D(t))\, dt \\ &= \int K(u)\, \mathcal{L}_D(D(x) - h_f u)\, du \to \mathcal{L}_D(D(x))\end{aligned}$$

as h_f converges to 0. Thus, our purpose in this paper is to investigate the properties of

$$\hat{f}(x; D, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(D(x) - D(X_i))}{\int K_{h_f}(D(x) - D(t))\, dt}$$

and

$$\hat{f}(x; \hat{D}_n, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(\hat{D}_n(x) - \hat{D}_n(X_i))}{\int K_{h_f}(\hat{D}_n(x) - \hat{D}_n(t))\, dt}$$

where $K_h(x) = K(x/h)/h$ for some kernel $K : \mathcal{R} \to [0, \infty)$ and some bandwidths h and h_f. The particular case with $h_f = 0$ corresponds to the situation where there exist an explicit expression for $\mathcal{L}_D$ and, since h_f is merely used to provide a simple approximation for $\mathcal{L}_D(D(x))$, our intention is to let h_f converge to 0 faster that h does. The estimators $\hat{f}(x; D, h)$ and $\hat{f}(x; \hat{D}_n, h)$ are always non–negative but do not integrate to 1 and, in practice, they need to be normalized.

Notions of multivariate depth are interesting candidates for D because they can usually be estimated at the usual parametric rate $1/\sqrt{n}$. Since f_D is estimated at a one–dimensional nonparametric rate of convergence, the estimation of D should not affect the overall rate of convergence. This implies that for the class of densities such that $f = g \circ D$, we get a one–dimensional nonparametric rate of convergence in a p–dimensional density estimation problem. This assertion will be proved in the next Section. Many notions of multivariate depth have been defined and studied in the literature (see Small, 1990) , including

- Mahalanobis depth (Mahalanobis, 1936)

$$D(x) = 1/\left(1 + (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right).$$

- Tukey's depth (Tukey, 1975)

$$D(x) = \inf\left\{\int_H dP(x)\ H \text{ is a closed half–space containing } x\right\}.$$

- Simplicial depth (Liu, 1990) $D(x) = \Pr\{x \in S[X_1, \ldots, X_{p+1}]\}$ where $S[X_1, \ldots, X_{p+1}]$ is the simplex with vertices $X_1, \ldots, X_{p+1}$.
- APL depth (Fraiman and Meloche, 1996) $D(x) = K_\gamma * f(x)$ for some kernel K and some fixed smoothing parameter γ.

All of the above depths can be estimated at the $1/\sqrt{n}$ parametric rate so that they can all be estimated at no cost in terms of the asymptotic behaviour of $\hat{f}(x; \hat{D}_n, h)$. The depth, however, does have an impact on the asymptotic bias and variance of $\hat{f}(x; \hat{D}_n, h)$. More importantly, the depth D determines the class of densities f of the form $f = g \circ D$. As described in Example 2, for $D(x) = (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$, $f = g \circ D$ if and only if f is ellipsoidal. In the case of Simplicial depth, although we know the level curves of D must be convex, it is not clear how large the class of densities f of the form $f = g \circ D$ is. The level curves for APL depth don't even need to be convex but the equation $f = g \circ D$ does not appear to be satisfied except for ellipsoidal densities.

As noted before, density estimators that perform particularly well under some class of density are called "tailor design density estimate" by Devroye. One can regard the estimator we propose as one that will take advantage of the relationship $f = g \circ D$ for a given notion of depth. The proposed estimates are not universally consistent (they converge only if $f = g \circ D$) but they provide better performance than the universally consistent estimate on the class of densities f of the form $f = g \circ D$.

2 Main Results

In this section, we present results concerning the strong convergence and the asymptotic normality of

$$\hat{f}(x; D, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(D(x) - D(X_i))}{\int K_{h_f}(D(x) - D(t))\, dt}$$

and

$$\hat{f}(x; \hat{D}_n, h) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(\hat{D}_n(x) - \hat{D}_n(X_i))}{\int K_{h_f}(\hat{D}_n(x) - \hat{D}_n(t))\, dt}$$

where $\hat{D}_n$ is an estimator of D. We summarize below the assumptions that are needed. Proofs can be found in the Appendix.

H0: The bandwidth sequence $h = h_n$ is such that $nh^5 \to \beta^2 \in (0, \infty)$. The bandwidth sequence $h_f = h_{nf}$ converges to zero faster than h does: $h_f/h \to 0$.

H1: The kernel K is symmetric, has a bounded support, integrates to 1 and has three bounded and continuous derivatives.

H2: $X_1, \ldots, X_n$ are i.i.d. with some density $f : \mathcal{R}^p \to [0, \infty)$ such that $f = g \circ D$ for some function $g : \mathcal{R} \to [0, \infty)$. Both g and f are bounded and have two bounded and continuous derivatives.

H3: There exist a $1 - 1$ and continuously differentiable transformation $T : \mathcal{R} \times [0,1]^{p-1} \to \mathcal{R}^p$ such that $D(T(r, \theta)) = r$ for all $r \in \mathcal{R}$ and all $\theta \in [0,1]^{p-1}$. The transformations T and its Jacobian J_T have two bounded and continuous partial derivatives with respect to r.

H4: The inverse image by D of a bounded set is bounded and x is in the interior of the support of D.

Assumptions **H0** and **H1** are more or less the standard assumption for the bandwidth and the kernel in kernel density estimation. The bounded support of the kernel is not usually needed but simplifies the proofs. The necessity of Assumptions **H2** can be explained as follows. Since $\hat{f}(x; D, h)$ is a function of $D(x)$ so that we can only hope to get consistency if $f = g \circ D$ for some $g : \mathcal{R} \to [0, \infty)$. Note that, by virtue of **H3**,

$$\begin{aligned}
&\Pr\{D(X_1) \in A\} = \int \mathbf{1}\{D(t) \in A\}\, g(D(t))\, dt \\
&= \int\int \mathbf{1}\{D(T(r,\theta)) \in A\}\, g(D(T(r,\theta)))\, |J_T(r,\theta)|\, d\theta dr \\
&= \int\int \mathbf{1}\{r \in A\}\, g(r)\, |J_T(r,\theta)|\, d\theta dr \\
&= \int_A g(r)\, \mathcal{L}_D(r)\, dr
\end{aligned}$$

where $\mathcal{L}_D(r) = \int |J_T(r,\theta)|\, d\theta$. Thus, **H3** implies that the random variable $D(X_1)$ has the density $f_D = g\mathcal{L}_D$.

Theorem 1 *If **H0-H3** hold,*

$$\sqrt{nh}\, \left\{\hat{f}(x; D, h) - f(x)\right\} \xrightarrow{\mathcal{L}}$$
$$N\left(\tfrac{1}{2}\, \beta\, \int u^2 K(u)\, du\, \frac{f_D''(D(x))}{\mathcal{L}_D(D(x))}, \int K^2(u)\, du\, \frac{f(x)}{\mathcal{L}_D(D(x))}\right).$$

The asymptotic distribution of $\hat{f}(x; D, h)$ should be compared to that of the multivariate kernel density estimator $\hat{f}(x; h)$ which is well known to be

$$\sqrt{nh^d}\left\{\hat{f}(x;h) - f(x)\right\} \xrightarrow{\mathcal{L}} N\left(\frac{1}{2}\beta \int u^2 K(u)\, du\, \nabla^2 f(x), \int K^2(u)\, du\, f(x)\right).$$

Table 1 provides the asymptotic bias and variance for both $\hat{f}(x; h)$ and $\hat{f}(x; D, h)$. The most striking difference is the rate of the convergence to 0 for the asymptotic variance. The asymptotic bias has the same rate of the convergence to 0, but the constants depend on f in different ways. In the table, $\alpha = \int u^2 K(u)\, du$ and $\beta = \int K^2(u)\, du$.

Table 1: The asymptotic bias and variance for $\hat{f}(x; h)$ and $\hat{f}(x; D, h)$.

Estimator	Asymptotic Bias	Asymptotic Variance
$\hat{f}(x; h)$	$\frac{1}{2}\alpha h^2\, \nabla^2 f(x)$	$\frac{\beta}{nh^d}\, f(x)$
$\hat{f}(x; D, h)$	$\frac{1}{2}\alpha h^2\, \frac{f''_D(D(x))}{\mathcal{L}_D(D(x))}$	$\frac{\beta}{nh}\, \frac{f(x)}{\mathcal{L}_D(D(x))}$
$D(x) = \lvert x - \mu \rvert \quad (p = 1)$	$\frac{1}{4}\alpha h^2\, f''_D(D(x))$	$\frac{\beta}{nh}\, \frac{f(x)}{2}$

Our proof of the asymptotic normality for $\hat{f}(x; \hat{D}_n, h)$ uses a three term Taylor series approximation for $\hat{f}(x; \hat{D}_n, h)$. We prove that provided $\sqrt{nh}$ $\|\hat{D}_n - D\|_\infty^3 = o_P(h_f^4)$,

$$\sqrt{nh}\left\{\hat{f}(x; \hat{D}_n, h) - \hat{f}(x; D, h)\right\} \xrightarrow{\mathcal{P}} 0.$$

Note that for the kernel density estimator, the optimal bandwidth is of the order $n^{-1/5}$ so that if $\|\hat{D}_n - D\|_\infty = O_P(1/\sqrt{n})$, the condition $\sqrt{nh}\, \|\hat{D}_n - D\|_\infty^3 = o_P(h_f^4)$ amounts to $h/(n^2 h_f^8) \to 0$ and can be satisfied for the optimal bandwidth h and a slightly faster h_f. A higher order Taylor series approximation for $\hat{f}(x; \hat{D}_n, h)$ would result in weaker restrictions on the bandwidths h and h_f but at the cost of additional smoothness requirements on K, T and J_T.

Theorem 2 *Assume **H0-H4** hold and define*

$$H_{nk}(s,t) = n\, \mathrm{E}\left\{A_n^k(x, X_1) A_n^k(x, X_2) | X_1 = s, X_2 = t\right\}$$

and

$$G_{nk}(s,t) = n\, \mathrm{E}\, A_n^k(x,s) A_n^k(x,t)$$

where $A_n(x,y) = ((\hat{D}_n(x) - \hat{D}_n(y)) - (D(x) - D(y)))$. *If* H_{nk} *and* G_{nk} *have two countinuous and uniformly bounded derivatives for* $k = 1, 2$ *and if* $\sqrt{nh}\, \|\hat{D}_n - D\|_\infty^3 = o_P(h_f^4)$, $\hat{f}(x; \hat{D}_n, h)$ *and* $\hat{f}(x; D, h)$ *have the same asymptotic distribution.*

Besides the smoothness assumption (**H2**) made on f and g, the application of Theorem 2 requires the verification of **H3**, the existence of a smooth change of variable transformation T such that $D(T(r,\theta)) = r$. The existence of such a transformation is guaranteed whenever D has a unique global maximum M (such a global maximum can be regarded as the deepest point or a median) and is decreasing along every ray originating from M. In such circumstances, we can define T as the inverse of the "polar transformation" (D, Θ) where D is replacing the usual norm and where the angles Θ are determined about the maximum M. The smoothness of the resulting T is equivalent to the smoothness of D. Thus, we can reformulate all the smoothness assumptions in terms of the smoothness of g and D.

3 Simulation

In this section, we present a small simulation that compares $\hat{f}(x; \hat{D}_n, h)$ to the kernel density estimator $\hat{f}(x, h)$ in cases where the underlying density f is of the form $f = g \circ D$. The simulation involves the Mahalanobis depth. It involves 100 samples of 50 observations i.i.d. with distribution

$$N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1.0 & 0.9\\0.9 & 1.0\end{pmatrix}\right).$$

For the Mahalanobis depth, $D(x,y) = (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)$ and $\hat{D}_n(x,y) = (x-\hat{\mu})^{\mathrm{T}}\hat{\Sigma}^{-1}(x-\hat{\mu})$ with the usual sample moment estimators $\hat{\mu}$ and $\hat{\Sigma}$. In this case we know $\mathcal{L}_D(r)$ up to a normalizing factor and we use $h_f = 0$. We evaluate both $\hat{f}(x; \hat{D}_n, h)$ and $\hat{f}(x, h)$ with their respective ASE–optimal bandwidth determined by minimizing over h

$$ASE(\hat{f}(x; \hat{D}_n, h), h) = \frac{1}{50} \sum_{i=1}^{50} \left(\hat{f}(X_i; \hat{D}_n, h) - f(X_i)\right)^2$$

and

$$ASE(\hat{f}(x; h), h) = \frac{1}{50} \sum_{i=1}^{50} \left(\hat{f}(X_i; h) - f(X_i)\right)^2$$

respectively.

Note that the numerator of $\hat{f}(x; \hat{D}_n, h)$ is in fact a kernel density estimate for the data $\hat{D}_n(X_1), \ldots, \hat{D}_n(X_n)$ evaluated at $\hat{D}_n(x)$. In practice, a boundary correction kernel must be used because $\hat{D}_n$ has a bounded range. It is also possible to use a transformation to avoid having to make a boundary correction. The simulation uses a boundary correction kernel.

Table 2 below summarizes the results. With the Mahalanobis depth, $\hat{f}(x; \hat{D}_n, h)$ clearly outperforms $\hat{f}(x, h)$. The average ASE for $\hat{f}(x; \hat{D}_n, h)$ is

about three times smaller than the average ASE for $\hat{f}(x,h)$. According to the theoretical results of the previous section, for such elliptical densities, the difference between the performance of $\hat{f}(x;\hat{D}_n,h)$ and $\hat{f}(x,h)$ should grow with the dimension.

Table 2: ASE of $\hat{f}(x,h)$ and $\hat{f}(x;\hat{D}_n,h)$.

		Mahalanobis
	$\hat{f}(x,h)$	$\hat{f}(x;\hat{D}_n,h)$
Average	0.96	0.32
S.D.	0.28	0.19

Figure 1 illustrates a very typical outcome of the above simulation. Figure 1 a) displays the 50 observations, Figure 1 b) displays the true density, Figure 1 c) displays the kernel density estimate $\hat{f}(x,h)$ and Figure 1 d) displays $\hat{f}(x;\hat{D}_n,h)$ for the Mahalanobis depth. The estimate $\hat{f}(x;\hat{D}_n,h)$ is by construction perfectly ellipsoidal and the spurious bumps and dips found in $\hat{f}(x,h)$ have disappeared. Also the bandwidth minimizing the ASE for the kernel density estimate is much smaller than the bandwidth minimizing the ASE for the Mahalanobis depth density estimate.

4 Conclusion

In cases where the unknown density f satisfies $f = g \circ D$, our theoretical results show that, in high dimension, $\hat{f}(x;\hat{D}_n,h)$ has better asymptotic performance than the usual kernel density estimator $\hat{f}(x,h)$. Our simulations suggest this is true in two dimension, even for small samples.

The theoretical results also make smoothness assumptions that excludes cases where the unknown density f is multimodal. Even though the smoothness assumptions can be reduced to include such cases, an important practical problem remains. For multimodal densities, f_D (estimated by the numerator of $\hat{f}(x;\hat{D}_n,h)$) and $\mathcal{L}_D$ (estimated by the denominator of $\hat{f}(x;\hat{D}_n,h)$) are discontinuous functions. Depth based estimation for such densities would require the development of a reasonably good density estimator for discontinuous densities.

Appendix: Proofs of Theorems 1 and 2

This section is devoted to the proof of Theorems 1 and 2.

Proof of Theorem 1: Define

$$\hat{f}_D(r;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(r - D(X_i)),$$

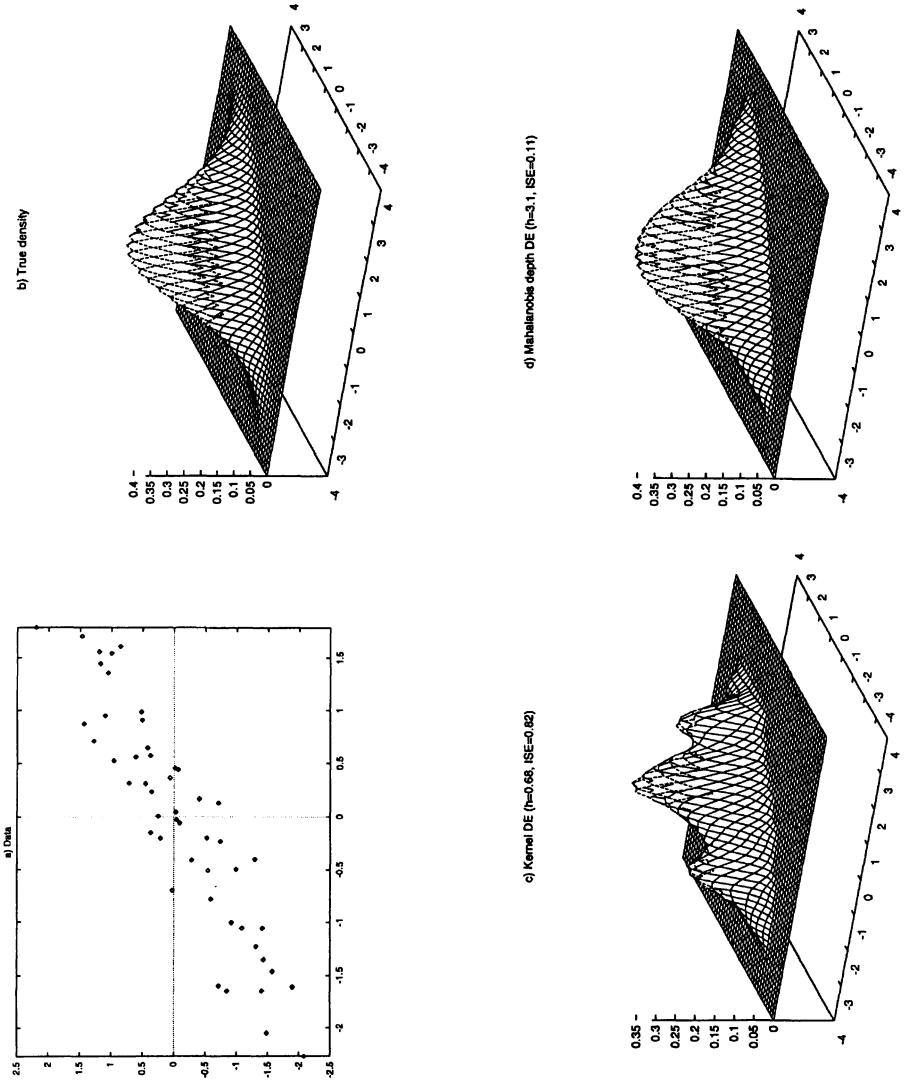

Figure 1: Outcome of the simulation.

$$\overline{\mathcal{L}}_D(r) = \int K_{h_f}(r - D(t))\, dt = \int K(u)\, \mathcal{L}_D(r - h_f u)\, du$$

and

$$\overline{f}_D(r) = \mathrm{E}\, K_h(r - D(X_1)) = \int K(u)\, f_D(r - hu)\, du.$$

With this notation,

$$\hat{f}(x; D, h) = \frac{\widehat{f}_D(D(x); h)}{\overline{\mathcal{L}}_D(D(x))}$$

and $\widehat{f}_D(r; h)$ is clearly a kernel density estimator of $f_D(r)$. The asymptotic normality of the kernel density estimator is well know so that

$$\sqrt{nh}\, \left\{\widehat{f}_D(r; h) - \overline{f}_D(r)\right\} \xrightarrow{\mathcal{L}} N\left(0, f_D(r) \int K^2(u)\, du\right).$$

Therefore (since $g = f_D/\mathcal{L}_D$),

$$\sqrt{nh}\, \frac{\widehat{f}_D(r; h) - \overline{f}_D(r)}{\overline{\mathcal{L}}_D(r)} \xrightarrow{\mathcal{L}} N\left(0, \frac{g(r)}{\mathcal{L}_D(r)} \int K^2(u)\, du\right).$$

The stated asymptotic normality for $\hat{f}(x; D, h)$ follows because

$$\begin{aligned}
&\sqrt{nh}\{\frac{\overline{f}_D(r)}{\overline{\mathcal{L}}_D(r)} - \frac{f_D(r)}{\mathcal{L}_D(r)}\} = \sqrt{nh}\frac{\overline{f}_D(r)\mathcal{L}_D(r) - f_D(r)\overline{\mathcal{L}}_D(r)}{\overline{\mathcal{L}}_D(r)\mathcal{L}_D(r)} \\
&= \sqrt{nh}\frac{\int K(u)\{f_D(r-hu)\mathcal{L}_D(r) - f_D(r)\mathcal{L}_D(r-h_f u)\}du}{\overline{\mathcal{L}}_D(r)\mathcal{L}_D(r)} \\
&\to \tfrac{1}{2}\beta\frac{f_D''(r)}{\mathcal{L}_D(r)} \int u^2 K(u) du.
\end{aligned}$$

Lemma A Let γ_n be some sequence that converges to infinity and assume that

$$\gamma_n \left\{\frac{a_n}{b_n} - c\right\} \xrightarrow{\mathcal{L}} G.$$

If $\gamma_n(A_n - a_n) \xrightarrow{\mathcal{P}} 0$, $\gamma_n(B_n - b_n) \xrightarrow{\mathcal{P}} 0$ and $B_n \xrightarrow{\mathcal{P}} b > 0$, then

$$\gamma_n \left\{\frac{A_n}{B_n} - c\right\} \xrightarrow{\mathcal{L}} G.$$

Proof: Simply write

$$\gamma_n \left\{\frac{A_n}{B_n} - c\right\} = \gamma_n \left\{\frac{A_n}{B_n} - \frac{a_n}{b_n}\right\} + \gamma_n \left\{\frac{a_n}{b_n} - c\right\},$$

note that

$$\gamma_n \left\{\frac{A_n}{B_n} - \frac{a_n}{b_n}\right\} = \frac{\gamma_n}{B_n}(A_n - a_n) - \frac{\gamma_n}{B_n}\,(B_n - b_n)\,\frac{a_n}{b_n}$$

and apply Slutsky Theorem. □

Lemma B If $\int \left(K^{(2)}(u)\right)^2 \, du < \infty$ and if the density f_D of $D(X_1)$ is bounded,

$$\sup_h h^{2k} \mathrm{E} h \left(K_h^{(k)}(r - D(X_1))\right)^2 < \infty.$$

Proof: Note that

$$\begin{aligned}
&h^{2k} \mathrm{E} h \left(K_h^{(k)}(r - D(X_1))\right)^2 = h^{2k+1} \int \left(K_h^{(k)}(r-t)\right)^2 f_D(t)\, dt \\
&= h^{2k+1} \int h^{-2(k+1)} \left(K^{(k)}\left(\tfrac{r-t}{h}\right)\right)^2 f_D(t)\, dt \\
&= \int \tfrac{1}{h} \left(K^{(k)}\left(\tfrac{r-t}{h}\right)\right)^2 f_D(t)\, dt \\
&= \int \left(K^{(k)}(u)\right)^2 f_D(r - hu)\, dt \\
&\leq \|f_D\|_\infty \int \left(K^{(k)}(u)\right)^2 dt \quad \square
\end{aligned}$$

Lemma 1 Assume that **H3** holds. If $G_n : \mathcal{R}^2 \to \mathcal{R}$ has two continuous and uniformly bounded derivatives, then for $k \in \{1, 2\}$,

$$\sup_n \left| \int \int G_n(x, y)\, K_h^{(k)}(u - D(x))\, K_h^{(k)}(v - D(y))\, dxdy \right| < \infty.$$

Proof: By virtue of **H3**,

$$\begin{aligned}
&\int \int G_n(x, y)\, K_h^{(k)}(u - D(x))\, K_h^{(k)}(v - D(y))\, dxdy \\
&= \int \int \int \int G_n(T(s, \theta_1), T(t, \theta_2))\, K_h^{(k)}(u - D(T(s, \theta_1)))\, K_h^{(k)}(v - D(T(t, \theta_2))) \\
&\qquad\qquad |J_T(s, \theta_1)|\, |J_T(t, \theta_2)|\, d\theta_1 d\theta_2\, dsdt \\
&= \int \int \overline{G}_n(s, t)\, K_h^{(k)}(u - s)\, K_h^{(k)}(v - t)\, dsdt
\end{aligned}$$

with

$$\overline{G}_n(s, t) = \int \int G_n(T(s, \theta_1), T(t, \theta_2))\, |J_T(s, \theta_1)|\, |J_T(t, \theta_2)|\, d\theta_1 d\theta_2.$$

Since T, J_T and G_n have two continuous and uniformly bounded derivatives, so does $\overline{G}_n$ and the result easily follows from

$$\begin{aligned}
&\int \int G_n(x, y)\, K_h^{(k)}(u - D(x))\, K_h^{(k)}(v - D(y))\, dxdy \\
&= \int \int \overline{G}_n(s, t)\, K_h^{(k)}(u - s)\, K_h^{(k)}(v - t)\, dsdt \\
&= \int \int \tfrac{\partial^k}{\partial u^k} \tfrac{\partial^k}{\partial v^k} \overline{G}_n(s, t)\, K_h(u - s)\, K_h(v - t)\, dsdt \\
&= \int \int \tfrac{\partial^k}{\partial u^k} \tfrac{\partial^k}{\partial v^k} \overline{G}_n(u - s, v - t)\, K_h(s)\, K_h(t)\, dsdt \\
&\leq \left\| \tfrac{\partial^k}{\partial u^k} \tfrac{\partial^k}{\partial v^k} \overline{G}_n(u, v) \right\|_\infty . \quad \square
\end{aligned}$$

Lemma 2 Assume that **H3** holds. If $H_n : \mathcal{R}^2 \to \mathcal{R}$ has two continuous and uniformly bounded derivatives, and if f has two continuous and bounded derivatives, then for $k \in \{1, 2\}$,

$$\sup_n \left| \mathrm{E}\, H_n(X_1, X_2)\, K_h^{(k)}(u - D(X_1))\, K_h^{(k)}(v - D(X_2)) \right| < \infty.$$

Proof: Note that

$$\begin{aligned} &\mathrm{E}\, H_n(X_1, X_2)\, K_h^{(k)}(u - D(X_1))\, K_h^{(k)}(v - D(X_2)) \\ &= \int\int H_n(x, y)\, K_h^{(k)}(u - D(x))\, K_h^{(k)}(v - D(y))\, f(x) f(y)\, dxdy \end{aligned}$$

and apply Lemma 1 to $G_n(x, y) = H_n(x, y) f(x) f(y)$. □

Proof of Theorem 2: First note that $|A_n(u, v)| \leq 2\|\hat{D}_n - D\|_\infty$ and that $\sqrt{nh}\, \|\hat{D}_n - D\|_\infty^3 = \mathrm{o_P}(h_f^4)$ implies $\sqrt{nh}\, \|\hat{D}_n - D\|_\infty^3 = \mathrm{o_P}(h^4)$ and $\|\hat{D}_n - D\|_\infty = \mathrm{O_P}(h)$. Using Lemma A, we can prove that $\hat{f}(x; \hat{D}_n, h)$ and $\hat{f}(x; D, h)$ have the same asymptotic distribution by showing

$$\sqrt{nh}\{\hat{f}_{\widehat{D}}(\hat{D}_n(x); h) - \hat{f}_D(D(x); h)\} \xrightarrow{\mathcal{P}} 0 \tag{1}$$

and

$$\sqrt{nh}\{\overline{\mathcal{L}}_{\widehat{D}}(\hat{D}_n(x)) - \overline{\mathcal{L}}_D(D(x))\} \xrightarrow{\mathcal{P}} 0 \tag{2}$$

We use Taylor series approximations for the numerator (1) and the denominator (2) separately. For the numerator,

$$\begin{aligned} \hat{f}_{\hat{D}}(\hat{D}_n(x); h) \;=\; &\sum_{k=0}^{2} \frac{1}{n} \sum A_n^k(x, X_i)\, K_h^{(k)}(D(x) - D(X_i)) \\ &+ \frac{1}{n} \sum A_n^3(x, X_i)\, K_h^{(3)}(\hat{\theta}) \end{aligned}$$

for some $\hat{\theta}$ on the segment joining $D(x) - D(X_i)$ to $\hat{D}_n(x) - \hat{D}_n(X_i)$. For the remainder, we have

$$A_n^3(x, X_i)\, K_h^{(3)}(\hat{\theta}) \leq 8\|\hat{D}_n - D\|_\infty^3\, \frac{\|K^{(3)}\|_\infty}{h^4}$$

and $\sqrt{nh}\, \|\hat{D}_n - D\|_\infty^3 = \mathrm{o_P}(h^4)$ ensures that the remainder is negligible. Thus, (1) follows from

$$nh\, \mathrm{E} \left(\frac{1}{n} \sum A_n^k(x, X_i)\, K_h^{(k)}(D(x) - D(X_i)) \right)^2 \to 0$$

for $k = 1, 2$. The above can be written as a double sum over all indices i and j. For $i = j$, we have

$$nh \, \tfrac{n}{n^2} \, \mathrm{E} \, A_n^{2k}(x, X_1) \left(K_h^{(k)}(D(x) - D(X_1)) \right)^2$$
$$\leq h \left(\tfrac{2\|\hat{D}_n - D\|_\infty}{h} \right)^{2k} \mathrm{E} h^{2k} \left(K_h^{(k)}(D(x) - D(X_1)) \right)^2$$

which converges to zero (invoke Lemma B) provided $\|\hat{D}_n - D\|_\infty = \mathrm{O_P}(h)$ (a consequence of $\sqrt{nh} \, \|\hat{D}_n - D\|_\infty^3 = \mathrm{o_P}(h^4)$). When $i \neq j$, we have

$$nh \, \tfrac{n(n-1)}{n^2} \, \mathrm{E} \, A_n^k(x, X_1) \, K_h^{(k)}(D(x) - D(X_1)) \, A_n^k(x, X_2) \, K_h^{(k)}(D(x) - D(X_2))$$
$$= h \, \mathrm{E} \, H_{nk}(X_1, X_2) \, K_h^{(k)}(D(x) - D(X_1)) K_h^{(k)}(D(x) - D(X_2))$$

which converges to zero as well because according to Lemma 2,

$$\sup_n \, |\mathrm{E} \, H_{nk}(X_1, X_2) K_h^{(k)}(D(x) - D(X_1)) K_h^{(k)}(D(x) - D(X_2))| < \infty.$$

For the denominator, we consider the expansion

$$\overline{\mathcal{L}}_{\hat{D}}(\hat{D}_n(x)) = \sum_{k=0}^{2} \int A_n^k(x, t) \, K_{h_f}^{(k)}(D(x) - D(t)) \, dt + \int A_n^3(x, t) \, K_{h_f}^{(3)}(\hat{\theta}_t) \, dt$$

for some $\hat{\theta}_t$ on the segment joining $D(x) - D(t)$ to $\hat{D}_n(x) - \hat{D}_n(t)$. Note that since K has a bounded support, **H**4 implies that for n large enough, $K_{h_f}^{(3)}(\hat{\theta}_t)$ almost surely vanishes outside of a bounded set so that (2) follows from

$$\sqrt{nh} \;\; A_n^3(x, t) \, K_{h_f}^{(3)}(\hat{\theta}_t) \xrightarrow{\mathcal{P}} 0$$

for almost all t and

$$nh \, \mathrm{E} \left(\int A_n^k(x, t) \, K_{h_f}^{(k)}(D(x) - D(t)) \, dt \right)^2 \to 0$$

for $k = 1, 2$. The proof of the above is similar to the one we made for the numerator but uses Lemma 1 instead of Lemma 2. □

Acknowledgements

This research was supported by NSF Grant DMS90-22126, CONICYT #37 and NSERC 5-81089.

References

[1] Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton Univ. Press.

[2] Brown, B. M. (1983). Statistical uses of the spatial median. *J. R. Statist. Soc.* B **45,** 25-30.

[3] Devroye, L. (1987). *A Course in Density Estimation.* Progress in Probability and Statistics, Vol. 14, Birkhauser.

[4] Fraiman, R. and Meloche, J. (1996). Multivariate L-Statistics. Technical Report, The University of British Columbia, Department of Statistics.

[5] Friedman, J.H. (1991). Multivariate Adaptive Regression Splines. *Ann. Statist.* **19,** 1-67.

[6] Friedman, J.H., Stuelzle, W. and Schroeder, A. (1984). Projection pursuit density estimation. *J. Am. Statist. Assoc.* **79,** 599-608.

[7] Grenander, U. (1956). On the theory of mortality measurements, part II. *Skandinavisk Aktuarietidskriff* **39,** 125-153.

[8] Huber, P.J. (1985). Projection pursuit. *Ann. Statist.* **13**, 435-475.

[9] Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18,** 405-414.

[10] Mahalanobis, P.C. (1936). On the generalized distance in statistics. In *Proc. of the National Academy of India,* Vol. 12, pp. 49-55.

[11] Oja, H., (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters* **1,** 327-332.

[12] Singh, K. (1991). A notion of majority depth. Technical Report, Rutgers University, Department of Statistics.

[13] Small, C.G. (1990). A survey of multidimensional medians. *Int. Statist. Rev.* **58,** 263-277.

[14] Stute, W. and Werner, U. (1991). Nonparametric estimation of elliptically contoured densities.In *Nonparametric Functional Estimation and Related Topics*, Ed. G. Roussas, pp. 173-190. NATO ASI Series.

[15] Tukey, J.W. (1975). Mathematics and picturing data. In *Proc. of the 1974 International Congress of Mathematicians,* Vol. 2, pp. 523-531. Vancouver.