Optimal Recovery of Local Truth

Carlos C. Rodríguez
Department of Mathematics and Statistics
University at Albany, SUNY
Albany NY 12222, USA
carlos@math.albany.edu
http://omega.albany.edu:8008/

Contents

Abstract

Probability mass curves the data space with horizons!. Let f be a multivariate probability density function with continuous second order partial derivatives. Consider the problem of estimating the true value of $f(z) > 0$ at a single point z, from n independent observations. It is shown that, the fastest possible estimators (like the k-nearest neighbor and kernel) have minimum asymptotic mean square errors when the space of observations is thought as conformally curved. The optimal metric is shown to be generated by the Hessian of f in the regions where the Hessian is definite. Thus, the peaks and valleys of f are surrounded by singular horizons when the Hessian changes signature from Riemannian to pseudo-Riemannian. Adaptive estimators based on the optimal variable metric show considerable theoretical and practical improvements over traditional methods. The formulas simplify dramatically when the dimension of the data space is 4. The similarities with General Relativity are striking but possibly illusory at this point. However, these results suggest that nonparametric density estimation may have something new to say about current physical theory.

1 Introduction

During the past thirty years the theory of Nonparametrics has been dominating the scene in mathematical statistics. Parallel to the accelerating discovery of new technical results, a consensus has been growing among some researchers in the area, that we may be witnessing a promising solid road towards the elusive Universal Learning Machine (see e.g. [1, 2]).

The queen of nonparametrics is density estimation. All the fundamental ideas for solving the new problems of statistical estimation in functional spaces (smoothing, generalization, optimal minimax rates, etc.) already appear in the problem of estimating the probability density (i.e. the model) from the observed data. More over, it is now well known that a solution for the density estimation problem automatically implies solutions for the problems of pattern recognition and nonparametric regression as well as for most problems that can be expressed as a functional of the density.

In this paper I present a technical result, about optimal nonparametric density estimation, that shows at least at a formal level, a surprising similarity between nonparametrics and General Relativity. Simply put,

probability mass curves the data space with horizons.

What exactly it is meant by this is the subject of this paper but before proceeding further a few comments are in order. First of all, let us assume that we have a set $\{x_1, \ldots, x_n\}$ of data. Each observation x_j consisting of p measurements that are thought as the p coordinates of a vector in $\mathbb{R}^p$. To make the data space into a probability space we endow $\mathbb{R}^p$ with the field of Borelians but nothing beyond that. In particular no a priori metric structure on the data space is assumed. The n observations are assumed to be n independent realizations of a given probability measure P on $\mathbb{R}^p$. By the Lebesgue decomposition theorem, for every Borel set B we can write,

$$P(B) = \int_B f(x)\lambda(dx) + \nu(B) \tag{1}$$

where ν is the singular part of P that assigns positive probability mass to Borel sets of zero Lebesgue volume. Due to the existence of pathologies like the Cantor set in one dimension and its analogies in higher dimensions, the singular part ν cannot be empirically estimated (see e.g. [3]). Practically all of the papers on density estimation rule out the singular part of P a priori. The elimination of singularities by fiat has permitted the construction of a

rich mathematical theory for density estimation, but it has also ruled out a priori the study of models of mixed dimensionality (see [4]) that may be necessary for understanding point masses and spacetime singularities coexisting with absolutely continuous distributions.

We assume further that in the regions where $f(x) > 0$ the density f is of class $\mathcal{C}^2$ i.e., it has continuous second order partial derivatives.

1.1 Nonparametrics with the World in Mind

The road from Classical Newtonian Physics to the physics of today can be seen as a path paved by an increasing use of fundamental concepts that are statistical in nature. This is obvious for statistical mechanics, becoming clearer for quantum theory, and appearing almost as a shock in General Relativity. Not surprisingly there have been several attempts to take this trend further (see e.g. [5, 6, 7, 8]) in the direction of *Physics as Inference*.

Now suppose for a moment that in fact some kind of restatement of the foundations of physics in terms of information and statistical inference will eventually end up providing a way to advance our current understanding of nature. As of today, that is either already a solid fact or remains a wild speculation, depending on who you ask. In any case, for the trend to take over, it will have to be able to reproduce all the successes of current science and make new correct predictions. In particular it would have to reproduce General Relativity. Recall that the main lesson of General Relativity is that space and time are not just a passive stage on top of which the universe evolves. General Relativity is the theory that tells (through the field equation) how to build the stage (left hand side of the equation) from the data (right hand side of the equation). The statistical theory that tells how to build the stage of inference (the probabilistic model) from the observed data is: *Nonparametric Density Estimation*. It is therefore reassuring to find typical signatures of General Relativity in density estimation as this paper does. Perhaps Physics is not just a special case of statistical inference and all these are only coincidences of no more relevance than for example the fact that multiplication or the logarithmic function appear everywhere all the time. That may be so, but even in that case I believe it is worth noticing the connection for undoubtedly GR and density estimation have a common goal: *The dynamic building of the stage*.

More formally. Let f be a multivariate probability density function with continuous second order partial derivatives. Consider the problem of esti-

mating the true value of $f(z) > 0$ at a single point z, from n independent observations. It is shown that, fastest possible estimators (including the k-nearest neighbor and kernel as well as the rich class of estimators in [9, theorem3.1]) have minimum asymptotic mean square errors when the space of observations is thought as conformally curved. The optimal metric is shown to be generated by the Hessian of f in the regions where the Hessian is definite. Thus, the peaks and valleys of f are surrounded by horizons where the Hessian changes signature from Riemannian to pseudo-Riemannian.

The result for the case of generalized k-nearest neighbor estimators [9] has circulated since 1988 in the form of a technical report [10]. Recently I found that a special case of this theorem has been known since 1972 [11] and undergone continuous development in the Pattern Recognition literature, (see e.g. [12, 13, 14, 15]).

2 Estimating Densities from Data

The canonical problem of density estimation at a point $z \in \mathbb{R}^p$ can be stated as follows: *Estimate $f(z) > 0$ from n independent observations of a random variable with density f.*

The most successful estimators of $f(z)$ attempt to approximate the density of probability at z by using the observations $x_1, \ldots, x_n$ to build both, a small volume around z and, a weight for this volume in terms of probability mass. The density is then computed as the ratio of the estimated mass over the estimated volume. The two classical examples are the k-nearest neighbor (knn) and the kernel estimators.

2.1 The knn

The simplest and historically the first example of a nonparametric density estimator is [16] the knn. The knn estimator of $f(z)$ is defined for $k \in \{1, 2, \ldots, n\}$ as,

$$h_n(z) = \frac{k/n}{\lambda_k} \tag{2}$$

where λ_k is the volume of the sphere centered at the point $z \in \mathbb{R}^p$ of radius $R(k)$ given by the distance from z to the kth-nearest neighbor observation. If λ denotes the Lebesgue measure on $\mathbb{R}^p$ we have,

$$\lambda_k = \lambda(S(R(k))) \tag{3}$$

where,

$$S(r) = \{x \in \mathbb{R}^p : \|x - z\| \leq r\} \tag{4}$$

The sphere $S(r)$ and the radius $R(k)$ are defined relative to a given norm, $\|\cdot\|$ in $\mathbb{R}^p$. The stochastic behavior of the knn depends on the specific value of the integer k chosen in (2). Clearly, in order to achieve consistency (e.g. stochastic convergence of $h_n(z)$ as $n \to \infty$ towards the true value of $f(z) > 0$) it is necessary to choose $k = k(n)$ as a function of n. The volumes λ_k must shrink, to control the bias, and consequently we must have $k/n \to 0$ for $h_n(z)$ to be able to approach a strictly positive number. On the other hand, we must have $k \to \infty$ to make the estimator dependent on an increasing number k of observations and in this way to control its variance. Thus, for the knn to be consistent, we need k to increase with n but at a rate slower than n itself.

The knn estimator depends not only on k but also on a choice of norm. The main result of this paper follows from the characterization of the $\|\cdot\|$ that, under some regularity conditions, produces the best asymptotic (as $n \to \infty$) performance for density estimators.

2.2 The kernel

If we consider only regular norms $\|\cdot\|$, in the sense that for all sufficiently small values of $r > 0$,

$$\lambda(S(r)) = \lambda(S(1))r^p \equiv \beta r^p \tag{5}$$

then, the classical kernel density estimator can be written as:

$$g_n(z) = \frac{M_\mu}{\lambda(S(\mu))} \tag{6}$$

where,

$$M_\mu = \frac{1}{n}\left(\sum_{x_j \in S(\mu)} K_{\mu^{-1}}(x_j - z)\right) \tag{7}$$

The smoothing parameter $\mu = \mu(n)$ is such that $k = [n\mu^p]$ satisfies the conditions for consistency of the knn, $K_{\mu^{-1}}(x) = K(\mu^{-1}x)$ where the kernel

function K is a non negative bounded function with support on the unit sphere (i.e. $K(x) = 0$ for $\|x\| > 1$) and satisfying,

$$\int_{\|x\|\leq 1} K(x)dx = \beta \tag{8}$$

Notice that for the constant kernel (i.e. $K(x) = 1$ for $\|x\| \leq 1$) the estimator (6) approximates $f(z)$ by the proportion of observations inside $S(\mu)$ over the volume of $S(\mu)$. The general kernel function K acts as a weight function allocating different weights $K_{\mu^{-1}}(x_j - z)$ to the x_j's inside $S(\mu)$. To control bias (see (32) below) the kernel K is usually taken as a decreasing radially symmetric function in the metric generated by the norm $\|\cdot\|$. Thus, $K_{\mu^{-1}}(x_j - z)$ assigns a weight to x_j that decreases with its distance to z. This has intuitive appeal, for the observations that lie closer to z are less likely to fall off the sphere $S(\mu)$, under repeated sampling, than the observations that are close to the boundary of $S(\mu)$.

The performance of the kernel as an estimator for $f(z)$ depends first and foremost on the value of the smoothness parameter μ. The numerator and the denominator of $g_n(z)$ depend not only on μ but also on the norm $\|\cdot\|$ chosen and the form of the kernel function K. As it is shown in theorem (8) these three parameters are inter-related.

2.3 Double Smoothing Estimators

The knn (2) and the kernel (6) methods are two extremes of a continuum. Both, $h_n(z)$ and $g_n(z)$ estimate $f(z)$ as *probability-mass-per-unit-volume*. The knn fixes the mass to the deterministic value k/n and lets the volume λ_k to be stochastic, while the kernel method fixes the volume $\lambda(S(\mu))$ and lets the mass M_μ to be random. The continuum gap between (2) and (6) is filled up by estimators that stochastically estimate mass and volume by smoothing the contribution of each sample point with different smoothing functions for the numerator and denominator (see [9]).

Let $b \geq 1$ and assume, without loss of generality that bk is an integer. The double smoothing estimators with deterministic weights are defined as,

$$f_n(z) = \frac{\frac{1}{n}\sum_{i=1}^{n} K\left(\frac{z-x_i}{R(k)}\right)}{\frac{1}{cb}\sum_{i=1}^{bk}\omega_i\lambda_i} \tag{9}$$

where,

$$\omega_i = \int_{(i-1)/bk}^{i/bk} \omega(u)du \tag{10}$$

and $\omega(\cdot)$ is a probability density on $[0, 1]$ with mean c.

3 The Truth as $n \to \infty$?

In nonparametric statistics, in order to assess the quality of an estimator $f_n(z)$ as an estimate for $f(z)$, it is necessary to choose a criterion for judging how far away is the estimator from what it tries to estimate. This is sometimes regarded as revolting and morally wrong by some Bayesian Fundamentalists. For once you choose a loss function and a prior, logic alone provides you with the Bayes estimator and the criterion for judging its quality. That is desirable, but there is a problem in high dimensional spaces. In infinite dimensional hypothesis spaces (i.e. in nonparametric problems) almost all priors will convince you of the wrong thing! (see e.g. [17, 18] for a non-regular way out see [19]). These kind of Bayesian nonparametric results provide a mathematical proof that: *almost all fundamental religions are wrong,* (more data can only make the believers more sure that the wrong thing is true!). An immediate corollary is that: *Subjective Bayesians can't go to Heaven*. Besides, the choice of goodness of fit criterion is as ad-hoc (an equivalent) to the choice of a loss function.

3.1 The Natural Invariant Loss Function and Why the MSE is not that Bad

The most widely studied goodness of fit criterion is the Mean Square Error (MSE) defined by,

$$(\text{MSE}) = E|f_n(z) - f(z)|^2 \tag{11}$$

where the expectation is over the joint distribution of the sample $x_1, \ldots, x_n$. By adding and subtracting $T = Ef_n(z)$ and expanding the square, we can express the MSE in the computationally convenient form,

$$\begin{aligned} (\text{MSE}) &= E|f_n(z) - T|^2 + |T - f(z)|^2 \\ &= (\text{variance}) + (\text{bias})^2 \end{aligned} \tag{12}$$

By integrating (11) over the $z \in \mathbb{R}^p$ and interchanging E and $\int$ (OK by Fubbini's theorem since the integrand ≥ 0) we obtain,

$$(\text{MISE}) = E \int |f_n(z) - f(z)|^2 dz \tag{13}$$

The Mean Integrated Square Error (MISE) is just the expected L^2 distance of f_n from f. Goodness of fit measures based on the (MSE) have two main advantages: They are often easy to compute and they enable the rich Hilbertian geometry of L^2. On the other hand the (MSE) is unnatural and undesirable for two reasons: Firstly, the (MSE) is only defined for densities in L^2 and this rules out a priori all the densities in $L^1 \setminus L^2$ which is unacceptable. Secondly, even when the $(MISE)$ exists, it is difficult to interpret (as a measure of distance between densities) due to its lack of invariance under relabels of the data space. Many researchers see the expected L^1 distance between densities as the natural loss function in density estimation. The L^1 distance does in fact exist for all densities and it is easy to interpret but it lacks the rich geometry generated by the availability of the inner product in L^2. A clean way out is to use the expected L^2 distance between the wave functions $\psi_n = \sqrt{f_n}$ and $\psi = \sqrt{f}$.

Theorem 1 *The L^2 norm of wave functions is invariant under relabels of the data space, i.e.,*

$$\int |\psi_n(z) - \psi(z)|^2 dz = \int |\tilde{\psi}_n(z') - \tilde{\psi}(z')|^2 dz' \tag{14}$$

where $z = h(z')$ with h any one-to-one smooth function.

Proof: Just change the variables. From, the change of variables theorem the pdf of z' is,

$$\tilde{f}(z') = f(h(z')) \left| \frac{\partial(h)}{\partial(z')} \right| \tag{15}$$

from where the wave function of z' is given by,

$$\tilde{\psi}(z') = \psi(h(z')) \left| \frac{\partial(h)}{\partial(z')} \right|^{1/2} \tag{16}$$

Thus, making the substitution $z = h(z')$ we get,

$$\begin{aligned}\int |\psi_n - \psi|^2 dz &= \int |\psi_n(h(z')) - \psi(h(z'))|^2 \left|\frac{\partial(h)}{\partial(z')}\right| dz' \\ &= \int |\psi_n \left|\frac{\partial(h)}{\partial(z')}\right|^{1/2} - \psi \left|\frac{\partial(h)}{\partial(z')}\right|^{1/2} |^2 dz' \\ &= \int |\tilde{\psi}_n - \tilde{\psi}|^2 dz' \end{aligned} \tag{17}$$

Q.E.D.

The following theorem shows that a transformation of the MSE of a consistent estimator provides an estimate for the expected L^2 norm between wave functions.

Theorem 2 *Let $f_n(z)$ be a consistent estimator of $f(z)$. Then,*

$$E \int |\psi_n - \psi|^2 dz = \frac{1}{4} \int \frac{E|f_n(z) - f(z)|^2}{f(z)} dz + \textit{(smaller order terms)} \tag{18}$$

Proof: A first order Taylor expansion of $\sqrt{x}$ about x_0 gives,

$$\sqrt{x} - \sqrt{x_0} = \frac{1}{2} \frac{(x - x_0)}{\sqrt{x_0}} + o((x - x_0)^2) \tag{19}$$

Substituting $x = f_n(z), x_0 = f(z)$ into (19) squaring both sides and taking expectations we obtain,

$$E|\psi_n(z) - \psi(z)|^2 = \frac{1}{4} \frac{E|f_n(z) - f(z)|^2}{f(z)} + o(E|f_n(z) - f(z)|^2) \tag{20}$$

integrating over z and interchanging E and $\int$ we arrive at (18).

Q.E.D.

Proceeding as in the proof of theorem 1 we can show that

$$\int \frac{|f_n - f|^2}{f} dz = \int \frac{|\tilde{f}_n - \tilde{f}|^2}{\tilde{f}} dz' \tag{21}$$

where, as before, $z \leftrightarrow z'$ is any one-to-one smooth transformation of the data space and $\tilde{f}$ is the density of z'. Thus, it follows from (21) that the leading term on the right hand side of (18) is also invariant under relabels of the data space. The nice thing about the L^2 norm of wave functions, unlike (21), is that it is defined even when $f(z) = 0$.

4 Some Classic Asymptotic Results

We collect here the well known Central Limit Theorems (CLT) for the knn and kernel estimators together with some remarks about nonparametric density estimation in general. The notation and the formulas introduced here will be needed for computing the main result about optimal norms in the next section.

Assumption 1 *Let f be a pdf on $\mathbb{R}^p$ of class $\mathcal{C}^2$ with non singular Hessian, $H(z)$ at $z \in \mathbb{R}^p$, and with $f(z) > 0$, i.e., the matrix of second order partial derivatives of f at z exists, it is non singular and its entries are continuous at z.*

Assumption 2 *Let K be a bounded non negative function defined on the unit sphere, $S_0 = \{x \in \mathbb{R}^p : \|x\| \leq 1\}$ and satisfying,*

$$\int_{\|x\|\leq 1} K(x)dx = \lambda(S_0) \equiv \beta \tag{22}$$

$$\int_{\|x\|\leq 1} xK(x)dx = 0 \in \mathbb{R}^p \tag{23}$$

Theorem 3 (CLT for knn) *Under assumption 1, if $k = k(n)$ is taken in the definition of the knn (2) in such a way that for some $a > 0$*

$$\lim_{n\to\infty} n^{-4/(p+4)}k = a \tag{24}$$

then, if we let $Z_n = \sqrt{k}(h_n(z) - f(z))$ we have,

$$\lim_{n\to\infty} P(Z_n \leq t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - B(z))^2}{2V(z)}\right) dy \tag{25}$$

where,

$$B(z) = \left(\frac{a^{\frac{p+4}{2p}}}{2f^{2/p}(z)}\right)\left\{\beta^{-1-2/p}\int_{\|x\|\leq 1} x^T H(z)xdx\right\} \tag{26}$$

and,

$$V(z) = f^2(z) \tag{27}$$

Proof: This is a special case of [9, theorem3.1].

Theorem 4 (CLT for kernel) *Under assumptions 1, and 2 if $\mu = \mu(n)$ is taken in the definition of the kernel (6) in such a way that for some $a > 0$, $k = [n\mu^p]$ satisfies (24) then, if we let $Z_n = \sqrt{k}(g_n(z) - f(z))$ we have (25) where now,*

$$B(z) = \left(\frac{a^{\frac{p+4}{2p}}}{2}\right)\left\{\beta^{-1}\int_{\|x\|\leq 1} x^T H(z)xK(x)dx\right\} \tag{28}$$

and,

$$V(z) = f(z)\left\{\beta^{-2}\int_{\|x\|\leq 1} K^2(x)dx\right\} \tag{29}$$

Proof: The sample $x_1, \ldots, x_n$ is assumed to be iid f and therefore the kernel estimator $g_n(z)$ given by (6) and (7) is a sum of iid random variables. Thus, the classic CLT applies and we only need to verify the rate (24) and the asymptotic expressions for the bias (28) and variance (29). We have,

$$\begin{aligned}
E\left[g_n(z)\right] &= \frac{1}{\beta\mu^p}\frac{1}{n}\sum_{j=1}^{n}\int K\left(\frac{x_j - z}{\mu}\right)f(x_j)dx_j && (30)\\
&= \frac{1}{\beta\mu^p}\int K(y)f(z+\mu y)\mu^p dy && (31)\\
&= \int \frac{K(y)}{\beta}\left\{f(z) + \mu\nabla f(z)\cdot y + \frac{\mu^2}{2}y^T H(z)y + o(\mu^2)\right\}dy && (32)\\
&= f(z) + \frac{\mu^2}{2\beta}\int y^T H(z)yK(y)dy + o(\mu^2) && (33)
\end{aligned}$$

where we have changed the variables of integration to get (31), used assumption 1 and Taylor's theorem to get (32) and used assumption 2 to obtain (33). For the variance we have,

$$\begin{aligned}
\mathrm{var}(g_n(z)) &= \frac{1}{n\beta^2\mu^{2p}}\mathrm{var}\left(K((X-z)/\mu)\right) && (34)\\
&= \frac{1}{n\beta^2\mu^{2p}}\left\{\int_{\|y\|\leq 1} f(z+\mu y)K^2(y)\mu^p dy\right.
\end{aligned}$$

$$- \left(\int_{\|y\| \leq 1} f(z + \mu y) K(y) \mu^p dy \right)^2 \Bigg\} \tag{35}$$

$$= \frac{f(z)}{n\beta^2 \mu^p} \int_{\|y\| \leq 1} K^2(y) dy + o\left(\frac{1}{n\mu^p} \right) \tag{36}$$

where we have used $\text{var}(K) = EK^2 - (EK)^2$ and changed the variables of integration to get (35), used assumption 1 and (0th order) Taylor's theorem to get (36). Hence, the theorem follows from (33) and (36) after noticing that (24) and $k = n\mu^p$ imply,

$$\sqrt{k}\mu^2 = k^{\frac{4+p}{2p}} n^{-2/p} = (n^{-\frac{4}{p+4}} k)^{\frac{p+4}{2p}} \longrightarrow a^{\frac{p+4}{2p}} \tag{37}$$

$$\frac{k}{n\mu^p} = \frac{k}{k} = 1 \tag{38}$$

Q.E.D.

Theorem 5 (CLT for double smoothers) *Consider the estimator $f_n(z)$ defined in (9). Under assumptions 1, 2, and (24) if we let $Z_n = \sqrt{k}(f_n(z) - f(z))$ we have (25) where now,*

$$B(z) = \left(\frac{a^{\frac{p+4}{2p}}}{2[\beta f(z)]^{2/p}} \right) \beta^{-1} \left\{ \int_{\|x\| \leq 1} x^T H(z) x \left[K(x) + \lambda_0 \right] dx \right\} \tag{39}$$

and,

$$V(z) = f^2(z) \left\{ \beta^{-1} \int_{\|x\| \leq 1} K^2(x) dx - \lambda_1 \right\} \tag{40}$$

with,

$$\lambda_0 = \frac{b^{2/p}}{c} \int_0^1 u^{1+\frac{2}{p}} \omega(u) du - 1 \tag{41}$$

$$\lambda_1 = 1 - c^{-2} b^{-1} \int_0^1 \left\{ \int_y^1 \omega(x) dx \right\}^2 dy \tag{42}$$

Proof: See [9, theorem3.1]. Remember to substitute K by $\beta^{-1}K$ since in the reference the Kernels are probability densities and in here we take them as weight functions that integrate to β.

4.1 Asymptotic Mean Square Errors

Let f_n be an arbitrary density estimator and let $Z_n = \sqrt{k}(f_n(z) - f(z))$. Now suppose that $f_n(z)$ is asymptotically normal, in the sense that when $k = k(n)$ satisfies (24) for some $a > 0$, we have (25) true. Then, all the moments of Z_n will converge to the moments of the asymptotic Gaussian. In particular the mean and the variance of Z_n will approach $B(z)$ and $V(z)$ respectively. Using, (12) and (24) we can write,

$$\lim_{n\to\infty} n^{4/(p+4)} E|f_n(z) - f(z)|^2 = \frac{V(z)}{a} + \frac{B^2(z)}{a} \tag{43}$$

We call the right hand side of (43) the asymptotic mean square error (AMSE) of the estimator $f_n(z)$. The value of a can be optimized to obtain a global minimum for the (AMSE) but it is well known in nonparametrics that the rate $n^{-4/(p+4)}$ is best possible (in a minimax sense) under the smoothness asumption 1 (see e.g. [20]). We can take care of the knn, the kernel, and the double smoothing estimators simultaneously by noticing that in all cases,

$$(\text{AMSE}) = \alpha_1 a^{-1} + \alpha_2 a^{4/p} \tag{44}$$

has a global minimum of,

$$(\text{AMSE})^* = \left\{ (1 + 4/p) \left(\frac{p}{4}\right)^{\frac{4}{p+4}} \right\} \alpha_1^{\frac{4}{p+4}} \alpha_2^{\frac{p}{p+4}} \tag{45}$$

achieved at,

$$a^* = \left(\frac{p\alpha_1}{4\alpha_2}\right)^{\frac{p}{p+4}} \tag{46}$$

Replacing the corresponding values for α_1 and α_2 for the knn, for the kernel, and for the double smoothing estimators, we obtain that in all cases,

$$(\text{AMSE})^* = (\text{const. indep. of } f) \left\{ f(z) \left(\frac{\Delta^2}{f(z)}\right)^{\frac{p}{p+4}} \right\} \tag{47}$$

where,

$$\Delta = \int_{\|x\|\le 1} x^T H(z) x G(x) dx \tag{48}$$

$$= \sum_{j=1}^{p} \rho_j \frac{\partial^2 f}{\partial z_j^2}(z) \tag{49}$$

with $G(x) = 1$ for the knn, $G(x) = K(x)$ for the kernel, $G(x) = K(x) + \lambda_0$ for the double smoothers (see (39) and (41)) and, if e_j denotes the jth canonical basis vector (all zeroes except a 1 at position j),

$$\rho_j = \int_{\|x\| \leq 1} (x \cdot e_j)^2 G(x) dx \tag{50}$$

Notice that (49) follows from (48), (23) and the fact that $H(z)$ is the Hessian of f at z. The generality of this result shows that (47) is typical for density estimation. Thus, when f_n is either the knn, the kernel, or one of the estimators in ([9, theorem3.1]), we have:

$$\lim_{n \to \infty} n^{4/(p+4)} E|f_n(z) - f(z)|^2 \geq c f(z) \left(\frac{\Delta^2}{f(z)} \right)^{\frac{p}{p+4}} \tag{51}$$

The positive constant c may depend on the particular estimator but it is independent of f. Dividing both sides of (51) by $f(z)$, integrating over z, using theorem 2 and interchanging E and $\int$ we obtain,

$$\lim_{n \to \infty} n^{4/(p+4)} E \int |\psi_n(z) - \psi(z)|^2 dz \geq 4c \int \left| \frac{\Delta}{\psi(z)} \right|^{\frac{2p}{p+4}} dz \tag{52}$$

The worst case scenario is obtained by the model $f = \psi^2$ that maximizes the action given by the right hand side of (52),

$$\mathcal{L} = \int \left| \frac{1}{\psi(z)} \sum_{j=1}^{p} \rho_j \frac{\partial^2 \psi^2}{\partial z_j^2}(z) \right|^{\frac{2p}{p+4}} dz \tag{53}$$

This is a hard variational problem. However, it is worth noticing that the simplest case is obtained when the exponent is 1, i.e. when the dimension of the data space is $p = 4$. Assuming we were able to find a solution, this solution would still depend on the p parameters $\rho_1, \ldots, \rho_p$. A choice of ρ_j's is equivalent to the choice of a global metric for the data space. Notice also, that the exponent becomes 2 for $p = \infty$ and that for $p \geq 3$ (but not for $p = 1$ or 2) there is the possibility of non trivial (i.e. different from uniform) super-efficient models for which estimation can be done at rates higher than $n^{-4/(p+4)}$. These super-efficient models are characterized as the non negative

solutions of the Laplace equation in the metric generated by the ρ_j's, i.e., non negative ($f(z) \geq 0$) solutions of,

$$\sum_{j=1}^{p} \rho_j \frac{\partial^2 f}{\partial z_j^2}(z) = 0 \tag{54}$$

Recall that there are no non trivial (different from constant) non negative super-harmonic functions in dimensions one or two but there are plenty of solutions in dimension three and higher. For example the Newtonian potentials,

$$f(z) = c\|z\|_\rho^{-(p-2)} \tag{55}$$

with the norm,

$$\|z\|_\rho^2 = \sum_{j=1}^{p} \left(\frac{z_j}{\sqrt{\rho_j}} \right)^2 \tag{56}$$

will do, provided the data space is compact. The existence of (hand picked) super-efficient models is what made necessary to consider best rates only in the minimax sense. Even though we can estimate a Newtonian potential model at faster than usual nonparametric rates, in any neighborhood of the Newtonian model the worst case scenario is at best estimated at rate $n^{-4/(p+4)}$ under second order smoothness conditions.

5 Choosing the Optimal Norm

All finite ($p < \infty$) dimensional Banach spaces are isomorphic (as Banach spaces) to $\mathbb{R}^p$ with the euclidian norm. This means, among other things, that in finite dimensional vector spaces all norms generate the same topology. Hence, the spheres $\{x \in \mathbb{R}^p : \|x\| \leq r\}$ are Borelians so they are Lebesgue measurable and thus, estimators like the knn (2) are well defined for arbitrary norms. It is possible, in principle, to consider norms that are not coming from inner products but in this paper we look only at Hilbert norms $\|\cdot\|_A$ of the form,

$$\|z\|_A^2 = z^T A^T A z \tag{57}$$

where $A \in \Lambda$ with Λ defined as the open set of real non-singular $p \times p$ matrices. For each $A \in \Lambda$ define the unit sphere,

$$S_A = \{x \in \mathbb{R}^p : x^T A^T A x \leq 1\} \tag{58}$$

its volume,

$$\beta_A = \lambda(S_A) = \int_{S_A} \lambda(dx) \tag{59}$$

and the A-symmetric (i.e. $\|\cdot\|_A$ radially symmetric) kernel, K_A,

$$K_A(x) = (K \circ A)(x) = K(Ax) \tag{60}$$

where K satisfies assumption 2 and it is I-symmetric, i.e., radially symmetric in the euclidian norm. This means that $K(y)$ depends on y only through the euclidian length of y, i.e. there exists a function F such that,

$$K(y) = F(y^T y) \tag{61}$$

The following simple theorem shows that all A-symmetric functions are really of the form (60).

Theorem 6 *For any $A \in \Lambda$, $\tilde{K}$ is A-symmetric if and only if we can write*

$$\tilde{K}(x) = K(Ax) \text{ for all } x \in \mathbb{R}^p \tag{62}$$

for some I-symmetric K.

Proof: $\tilde{K}(x)$ is A-symmetric iff $\tilde{K}(x) = F(\|x\|_A^2)$ for some function F. Choose $K(x) = \tilde{K}(A^{-1}x)$. This K is I-symmetric since $K(x) = F\left((AA^{-1}x)^T(AA^{-1}x)\right) = F(x^T x)$. More over, $\tilde{K}(x) = \tilde{K}(A^{-1}(Ax)) = K(Ax)$. Thus, (62) is necessary for A-symmetry. It is also obviously sufficient since the assumed I-symmetry of K in (62) implies that $\tilde{K}(x) = F((Ax)^T(Ax)) = F(\|x\|_A^2)$ so it is A-symmetric.
Q.E.D.

An important corollary of theorem 6 is,

Theorem 7 *Let $A, B \in \Lambda$. Then, $\tilde{K}$ is AB-symmetric if and only if $\tilde{K}_{B^{-1}}$ is A-symmetric.*

Proof: By the first part of theorem 6 we have that $\tilde{K} = K \circ A \circ B$ with K some I-symmetric. Thus, $\tilde{K} \circ B^{-1} = K \circ A$ is A-symmetric by the second part of theorem 6.
Q.E.D.

Let us denote by $\beta(A, K)$ the total volume that a kernel K assigns to the unit A-sphere S_A, i.e.,

$$\beta(A, K) = \int_{S_A} K(x)dx \tag{63}$$

In order to understand the effect of changing the metric on a density estimator like the kernel (6), it is convenient to write g_n explicitly as a function of everything it depends on; The point z, the metric A, the A-symmetric kernel function $\tilde{K}$, the positive smoothness parameter μ and, the data set $\{x_1, \ldots, x_n\}$. Hence, we define,

$$g_n(z; A, \tilde{K}, \mu, \{x_1, \ldots, x_n\}) = \frac{\frac{1}{n}\sum_{j=1}^{n} \tilde{K}\left(\frac{x_j - z}{\mu}\right)}{\beta(A, \tilde{K})\mu^p} \tag{64}$$

The following result shows that kernel estimation with metric AB is equivalent to estimation of a transformed problem with metric A. The explicit form of the transformed problem indicates that a perturbation of the metric should be regarded as composed of two parts: Shape and volume. The shape is confounded with the form of the kernel and the change of volume is equivalent to a change of the smoothness parameter.

Theorem 8 *Let $A, B \in \Lambda$, $\mu > 0$, and $\tilde{K}$ an AB-symmetric kernel. Then, for all $z \in \mathbb{R}^p$ and all data sets $\{x_1 \ldots, x_n\}$ we have,*

$$g_n(z; AB, \tilde{K}, \mu, \{x_1, \ldots, x_n\}) = g_n(\hat{B}z; A, \tilde{K} \circ B^{-1}, |B|^{-1/p}\mu, \{\hat{B}x_1, \ldots, \hat{B}x_n\}) \tag{65}$$

where $|B|$ denotes the determinant of B and $\hat{B} = |B|^{-1/p}B$ is the matrix B re-scaled to have unit determinant.

Proof: To simplify the notation let us denote,

$$\mu_B = \frac{\mu}{|B|^{1/p}} \tag{66}$$

Substituting AB for A in (64) and using theorem 6 we can write the left hand side of (65) as,

$$\frac{\frac{1}{n}\sum_{j=1}^{n} K\left(AB\left(\frac{x_j-z}{\mu}\right)\right)}{\beta(AB,\tilde{K})\mu^p} = \frac{\frac{1}{n}\sum_{j=1}^{n}(K\circ A)\left(\frac{\hat{B}x_j-\hat{B}z}{\mu_B}\right)}{\beta(A,K\circ A)(\mu_B)^p}$$

where K is some I-symmetric kernel and we have made the change of variables $x = B^{-1}y$ in $\beta(AB,\tilde{K})$ (see (63)). The last expression is the right hand side of (65). Notice that, $K\circ A = \tilde{K}_{B^{-1}}$ is A-symmetric.
Q.E.D.

5.1 Generalized knn Case with Uniform Kernel

In this section we find the norm of type (57) that minimizes (47) for the estimators of the knn type with uniform kernel which include the double smoothers with $K(x) = 1$. As it is shown in theorem 8 a change in the determinant of the matrix that defines the norm is equivalent to changing the smoothness parameter. The quantity (57) to be minimized was obtained by fixing the value of the smoothness parameter to the best possible, i.e. the one that minimizes the expression of the (AMSE) (43). Thus, to search for the best norm at a fix value of the smoothness parameter we need to keep the determinant of the matrix that defines the norm constant. We have,

Theorem 9 *Consider the problem,*

$$\min_{|A|=1}\left(\int_{S_A} x^T H(z)x dx\right)^2 \tag{67}$$

where the minimum is taken over all $p\times p$ matrices with determinant one, S_A is the unit A-ball and $H(z)$ is the Hessian of the density $f\in\mathcal{C}^2$ at z which is assumed to be nonsingular.

When $H(z)$ is indefinite, i.e. when $H(z)$ has both positive and negative eigenvalues, the objective function in (67) achieves its absolute minimum value of zero when A is taken as,

$$A = c^{-1}diag(\sqrt{\frac{\xi_1}{p-m}},\ldots,\sqrt{\frac{\xi_m}{p-m}},\sqrt{\frac{\xi_{m+1}}{m}},\ldots,\sqrt{\frac{\xi_p}{m}})M \tag{68}$$

where the ξ_j are the absolute value of the eigenvalues of $H(z)$, m is the number of positive eigenvalues, M is the matrix of eigenvectors and c is a

normalization constant to get $\det A = 1$ *(see the proof for more detailed definitions).*

If $H(z)$ *is definite, i.e. when* $H(z)$ *is either positive or negative definite, then for all* $p \times p$ *real matrices* A *with* $\det A = 1$ *we have,*

$$\left| \int_{S_A} x^T H(z) x dx \right| \geq \frac{2^p \pi}{p+3} \; p \left| \det H(z) \right|^{1/p} \tag{69}$$

with equality if and only if,

$$A = \frac{H_+^{1/2}(z)}{|H_+^{1/2}(z)|^{1/p}} \tag{70}$$

where $H_+^{1/2}(z)$ denotes the positive definite square-root of $H(z)$ (see the proof below for explicit definitions).

Proof: Since $f \in \mathcal{C}^2$ the Hessian is a real symmetric matrix and we can therefore find an orthogonal matrix M that diagonalizes $H(z)$, i.e. such that,

$$H(z) = M^T D M \quad \text{with} \quad M^T M = I \tag{71}$$

where,

$$D = \operatorname{diag}(\xi_1, \xi_2, \ldots, \xi_m, -\xi_{m+1}, \ldots, -\xi_p) \tag{72}$$

where all the $\xi_j > 0$ and we have assumed that the columns of M have been ordered so that all the m positive eigenvalues appear first and all the negative eigenvalues $-\xi_{m+1}, \ldots, -\xi_p$ appear last so that (71) agrees with (72). Split D as,

$$\begin{aligned} D &= \operatorname{diag}(\xi_1, \ldots, \xi_m, 0, \ldots, 0) - \operatorname{diag}(0, \ldots, 0, \xi_{m+1}, \ldots, \xi_p) \\ &= D_+ - D_- \end{aligned} \tag{73}$$

and use (71) and (73) to write,

$$\begin{aligned} H(z) &= M^T D_+ M - M^T D_- M \\ &= \left(D_+^{1/2} M\right)^T \left(D_+^{1/2} M\right) - \left(D_-^{1/2} M\right)^T \left(D_-^{1/2} M\right) \\ &= \Sigma_+^T \Sigma_+ - \Sigma_-^T \Sigma_- \end{aligned} \tag{74}$$

Using (74) and the substitution $y = Ax$ we obtain,

$$\begin{aligned} \int_{S_A} x^T H(z) x dx &= \int_{y^T y \leq 1} y^T \left(A^{-1}\right)^T \left(\Sigma_+^T \Sigma_+ - \Sigma_-^T \Sigma_-\right) A^{-1} y dy \\ &= \int_{y^T y \leq 1} \left\langle \Sigma A^{-1} y, \Sigma A^{-1} y \right\rangle dy \end{aligned} \tag{75}$$

where,

$$\Sigma = \Sigma_+ + \Sigma_- = (D_+ + D_-)^{1/2} M \tag{76}$$

and $< .,. >$ denotes the pseudo-Riemannian inner product,

$$\langle x, y \rangle = \sum_{i=1}^{m} x^i y_i - \sum_{i=m+1}^{p} x^i y_i \tag{77}$$

By letting $G = \mathrm{diag}(1, \ldots, 1, -1, \ldots, -1)$ (i.e. m ones followed by $p - m$ negative ones) be the metric with the signature of $H(z)$ we can also write (77) as,

$$\langle x, y \rangle = x^T G y \tag{78}$$

Let,

$$B = [b_1 | b_2 | \ldots | b_p] = \Sigma A^{-1} \tag{79}$$

where $b_1, \ldots, b_p$ denote the columns of B. Substituting (79) into (75), using the linearity of the inner product and the symmetry of the unit euclidian sphere we obtain,

$$\begin{aligned} \int_{S_A} x^T H(z) x dx &= \int_{y^T y \leq 1} \langle By, By \rangle dy \\ &= \sum_j \sum_k \langle b_j, b_k \rangle \int_{S_I} y^j y^k dy \qquad (80) \\ &= \sum_j \sum_k \langle b_j, b_k \rangle \delta^{jk} \rho \\ &= \rho \sum_{j=1}^{p} \langle b_j, b_j \rangle \qquad (81) \end{aligned}$$

where ρ stands for,

$$\rho = \int_{S_I} \left(y^1\right)^2 dy = \frac{2^p \pi}{p+3} \tag{82}$$

From (79) and (81) it follows that problem (67) is equivalent to,

$$\min_{|B|=|\Sigma|} \left(\sum_{j=1}^{p} \langle b_j, b_j \rangle \right)^2 \tag{83}$$

When $H(z)$ is indefinite, i.e. when $m \notin \{0, p\}$ it is possible to choose the columns of B so that $\sum_j \langle b_j, b_j \rangle = 0$ achieving the global minimum. There are many possible choices but the simplest one is,

$$B = c \cdot \mathrm{diag}(\underbrace{\sqrt{p-m}, \sqrt{p-m}, \ldots, \sqrt{p-m}}_{m}, \underbrace{\sqrt{m}, \sqrt{m}, \ldots, \sqrt{m}}_{p-m}) \tag{84}$$

since,

$$\sum_{j=1}^{p} \langle b_j, b_j \rangle = c^2 m(\sqrt{p-m})^2 - c^2(p-m)(\sqrt{m})^2 = 0. \tag{85}$$

The scalar constant c needs to be adjusted to satisfy the constraint that $\det B = \det \Sigma$. From (79), (84) and (76) we obtain that the matrix for the optimal metric can be written as,

$$A = B^{-1}\Sigma = \frac{c^{-1}}{\sqrt{p-m}}\Sigma_+ + \frac{c^{-1}}{\sqrt{m}}\Sigma_- \tag{86}$$

From (86) we get,

$$A^T A = \frac{c^{-2}}{p-m}\Sigma_+^T\Sigma_+ + \frac{c^{-2}}{m}\Sigma_-^T\Sigma_- \tag{87}$$

Finally from (74) we can rewrite (87) as,

$$\begin{aligned} A^T A &= c^{-2} M^T \left(\frac{1}{p-m} D_+ + \frac{1}{m} D_- \right) M & (88) \\ &= c^{-2} M^T \mathrm{diag}(\frac{\xi_1}{p-m}, \ldots, \frac{\xi_m}{p-m}, \frac{\xi_{m+1}}{m}, \ldots, \frac{\xi_p}{m}) M & (89) \end{aligned}$$

Notice that when $p - m = m$ (i.e. when the number of positive equals the number of negative eigenvalues of $H(z)$) the factor $1/m$ can be factorized out from the diagonal matrix in (89) and in this case the optimal A can be expressed as,

$$A = \frac{H_+^{1/2}(z)}{|H_+^{1/2}(z)|^{1/p}} \tag{90}$$

where we have used the positive square-root of $H(z)$ defined as,

$$H_+^{1/2}(z) = \mathrm{diag}(\sqrt{\xi_1}, \ldots, \sqrt{\xi_p}) M \tag{91}$$

In all the other cases for which $H(z)$ is indefinite, i.e. when $m \notin \{0, p/2, p\}$ we have,

$$A = c^{-1}\mathrm{diag}(\sqrt{\frac{\xi_1}{p-m}}, \ldots, \sqrt{\frac{\xi_m}{p-m}}, \sqrt{\frac{\xi_{m+1}}{m}}, \ldots, \sqrt{\frac{\xi_p}{m}}) M \tag{92}$$

The normalization constant c is fixed by the constraint that $\det A = 1$ as,

$$c = (p-m)^{-\frac{m}{2p}} m^{-\frac{(p-m)}{2p}} |\det H(z)|^{\frac{1}{2p}} \tag{93}$$

This shows (68).

Let us now consider the only other remaining case when $H(z)$ is definite, i.e. either positive definite ($m = p$) or negative definite ($m = 0$). Introducing λ_0 as the Lagrange multiplier associated to the constraint $\det B = \det \Sigma$ we obtain that the problem to be solved is,

$$\min_{b_1,\dots b_p,\lambda_0} \mathcal{L}(b_1, b_2, \dots, b_p, \lambda_0) \tag{94}$$

where the Lagrangian $\mathcal{L}$ is written as a function of the columns of B as,

$$\mathcal{L}(b_1, b_2, \dots, b_p, \lambda_0) = \left(\sum_{j=1}^{p} \langle b_j, b_j \rangle \right)^2 - 4\lambda_0 (\det(b_1, \dots, b_p) - \det \Sigma) \tag{95}$$

The $-4\lambda_0$ instead of just λ_0 is chosen to simplify the optimality equations below. The optimality conditions are,

$$\frac{\partial \mathcal{L}}{\partial b_j} = 0 \quad \text{for} \quad j = 1, \dots, p \quad \text{and} \quad \frac{\partial \mathcal{L}}{\partial \lambda_0} = 0 \tag{96}$$

where the functional partial derivatives are taken in the Fréchet sense with respect to the column vectors b_j. The Fréchet derivatives of quadratic and multi linear forms are standard text-book exercises. Writing the derivatives as linear functions of the vector parameter h we have,

$$\frac{\partial}{\partial b_j} \langle b_j, b_j \rangle (h) = 2 \langle b_j, h \rangle \tag{97}$$

$$\frac{\partial}{\partial b_j} \det(b_1, \dots, b_p)(h) = \det(b_1, \dots, \underbrace{h}_{\text{j-th col.}}, \dots, b_p) \tag{98}$$

Thus, using (97) and (98) to compute the derivative of (95) we obtain that for all h and all $j = 1, \dots, p$ we must have,

$$\frac{\partial \mathcal{L}}{\partial b_j}(h) = 2 \left\{ \sum_{k=1}^{p} \langle b_k, b_k \rangle \right\} 2 \langle b_j, h \rangle - 4\lambda_0 \det(b_1, \dots, h, \dots, b_p) = 0 \tag{99}$$

When $\sum_k < b_k, b_k > \neq 0$ we can rewrite (99) as,

$$\langle b_j, h\rangle = c_0^{-1} \det(b_1, \ldots, h, \ldots, b_p) \tag{100}$$

But now we can substitute $h = b_i$ with $i \neq j$ into (100) and use the fact that the determinant of a matrix with two equal columns is zero, to obtain,

$$\langle b_j, b_i\rangle = 0 \quad \text{for all} \quad i \neq j. \tag{101}$$

In a similar way, replacing $h = b_j$ into (100), we get

$$\langle b_j, b_j\rangle = c_0^{-1} \det B = c \tag{102}$$

where c is a constant that needs to be fixed in order to satisfy the constraint that $\det B = \det \Sigma$. We have shown that the optimal matrix B must have orthogonal columns of the same length for the G-metric. This can be expressed with a single matrix equation as,

$$B^T G B = cI \tag{103}$$

Substituting (79) into (103) and re-arranging terms we obtain,

$$\begin{aligned} A^T A &= c^{-1}\Sigma^T G \Sigma & (104)\\ &= c^{-1}(\Sigma_+^T + \Sigma_-^T) G (\Sigma_+ + \Sigma_-) \\ &= c^{-1}(\Sigma_+^T \Sigma_+ - \Sigma_-^T \Sigma_-) \\ A^T A &= c^{-1} H(z) & (105) \end{aligned}$$

From (105), (103), (82) and (81) we obtain,

$$\left| \int_{S_A} x^T H(z) x \; dx \right| \geq \rho p |c| \tag{106}$$

and replacing the values of ρ and c we obtain (69).
Q.E.D.

5.2 Yet Another Proof When The Hessian is Definite

Consider the following lemma.

Lemma 1 *Let A, B be two $p \times p$ non-singular matrices with the same determinant. Then*

$$\int_{S_A} \|x\|_B^2 dx \geq \int_{S_B} \|y\|_B^2 dy \tag{107}$$

Proof: Just split S_A and S_B as,

$$S_A = (S_A S_B) \cup (S_A S_B^c) \tag{108}$$
$$S_B = (S_B S_A) \cup (S_B S_A^c) \tag{109}$$

and write,

$$\int_{S_A} \|x\|_B^2 dx = \int_{S_B} \|x\|_B^2 dx - \int_{S_A^c S_B} \|x\|_B^2 dx + \int_{S_A S_B^c} \|x\|_B^2 dx \tag{110}$$

Now clearly,

$$\min_{x \in S_A S_B^c} \|x\|_B^2 \geq 1 \geq \max_{y \in S_A^c S_B} \|y\|_B^2 \tag{111}$$

from where it follows that,

$$\int_{S_A S_B^c} \|x\|_B^2 dx \geq \min_{x \in S_A S_B^c} \|x\|_B^2 \int_{S_A S_B^c} dx \tag{112}$$
$$\geq \max_{y \in S_A^c S_B} \|y\|_B^2 \int_{S_A^c S_B} dy \tag{113}$$
$$\geq \int_{S_A^c S_B} \|y\|_B^2 dy \tag{114}$$

where (112) and (114) follow from (111). To justify the middle inequality (113) notice that from (108), (109) and the hypothesis that $|A| = |B|$ we can write,

$$\int_{S_A S_B^c} dx + \int_{S_A S_B} dx = \int_{S_A^c S_B} dy + \int_{S_A S_B} dy \tag{115}$$

The conclusion (107) follows from inequality (114) since that makes the last two terms in (110) non-negative.
Q.E.D.

If B is a nonsingular matrix we define,

$$\hat{B} = \frac{B}{|\det B|^{1/p}} \tag{116}$$

An immediate consequence of lemma 1 is,

Theorem 10 *If $H(z)$ is definite, then for all $p \times p$ matrices with $|A| = 1$ we have,*

$$\Delta = \int_{S_A} \|x\|_{H^{1/2}(z)}^2 dx \geq \int_{S_{\hat{H}^{1/2}(z)}} \|x\|_{H^{1/2}(z)}^2 dx \tag{117}$$

Proof:

$$\begin{aligned} \Delta &= |H(z)|^{1/p} \int_{S_A} \|x\|^2_{\hat{H}^{1/2}(z)} dx && (118) \\ &\geq |H(z)|^{1/p} \int_{S_{\hat{H}^{1/2}(z)}} \|x\|^2_{\hat{H}^{1/2}(z)} dx && (119) \\ &= \int_{S_{\hat{H}^{1/2}(z)}} \|x\|^2_{H^{1/2}(z)} dx && (120) \end{aligned}$$

where we have used lemma 1 to deduce the middle inequality (119).
Q.E.D.

5.3 Best Norm With General Kernels

In this section we solve the problem of finding the optimal norm in the general class of estimators (9).

Before we optimize the norm we need to state explicitly what it means to do estimation with different kernels and different norms. First of all a general kernel function is a nonnegative bounded function defined on the unit sphere generated by a given norm. Hence, the kernel only makes sense relative to the given norm. To indicate this dependence on the norm we write K_A for the kernel associated to the norm generated by the matrix A. We let

$$K_A = K \circ A \tag{121}$$

where $K = K_I$ is a fix mother kernel defined on the euclidian unit sphere. Equation (121) provides meaning to the notion of changing the norm without changing the kernel. What this means is not that the kernel is invariant under changes of A but rather equivariant in the form specified by (121). Recall also that a proper kernel must satisfy (22). To control bias we must also require the kernels to satisfy (23). It is easy to see (just change the variables) that if the mother kernel K has these properties so do all its children K_A with the only proviso that $|A| = 1$ in order to get (22). Notice also that radial symmetry of K is a sufficient but not a necessary condition for (23).

The optimization of the norm with general kernels looks more complicated than when the kernel is uniform since the best $(AMSE)^*$ also depends on $\int_{S_A} K_A^2(x)dx$. Consider the double smoothing estimators, which are the most general case treated in this paper. From, (39), (40) and (45) we have,

$$(AMSE)^* = (\text{const.}) \left\{ \beta^{-1} \int_{S_A} K_A^2(x)dx - \lambda_1 \right\}^{\frac{4}{p+4}} f(z) \left(\frac{\Delta^2}{f(z)} \right)^{\frac{p}{p+4}} \tag{122}$$

where the constant depends only on the dimension of the space. Even though the dependence of (122) on A looks much more complicated than (47) this is only apparently so. In fact the two expressions define very similar optimization problems as we now show.

First notice that the search for best A must be done within the class of matrices with a fix determinant. For otherwise we will be changing the value of the smoothness parameter that was fixed to the best possible value in order to obtain (122). If we let $|A| = 1$ we have,

$$\int_{S_A} K(x)\,dx = \beta = \int_{S_A} dx = \lambda(S_I) \tag{123}$$

We also have that,

$$\int_{S_A} K_A^2(y)\,dy = \int_{S_A} K^2(Ay)\,dy = \int_{S_I} K^2(y)\,dx \tag{124}$$

From (123) and (124) we deduce that the term in (122) within cursive brackets is the same for all matrices A and it depends only on the fix kernel K. Finally notice that the value of Δ in (122) is given by

$$\Delta = \int_{S_A} x^T H(z) x\, G(Ax)\,dx \tag{125}$$

where $G(x) = K(x) + \lambda_0$ in the general case. By retracing again the steps that led to (81) we can write,

$$\begin{aligned} \Delta &= \sum_j \sum_k \langle b_j, b_k \rangle \int_{S_I} y^j y^k G(y)\,dy & (126)\\ &= \sum_j \sum_k \langle b_j, b_k \rangle\, \delta^{jk} \rho_k(G) \\ &= \sum_{j=1}^{p} \langle b_j, b_j \rangle\, \rho_j(G) & (127) \end{aligned}$$

where now,

$$\rho_j(G) = \int_{S_I} \left(x^j\right)^2 G(x)\,dx \tag{128}$$

There are three cases to be considered.

1. All the $\rho_j(G) = \rho$ for $j = 1, \ldots, p$. The optimization problem reduces to the case when the kernel is uniform and therefore it has the same solution.

2. All the $\rho_j(G)$ have the same sign, i.e. they are all positive or all negative. If e.g. all $\rho_j > 0$ just replace b_j with $\sqrt{\rho_j} b_j$ and use the formulas obtained for the uniform kernel case.

3. Some of the $\rho_j(G)$ are positive and some are negative. This case can be handled like the previous one after taking care of the signs for different indices j.

The first case is the most important for it is the one implied when the kernels are radially symmetric. The other two cases are only included for completeness. Clearly if we do estimation with a non radially symmetric kernel the optimal norm would have to correct for this arbitrary builtin asymmetry, effectively achieving at the end the same performance as when radially symmetric kernels are used. The following theorem enunciates the main result.

Theorem 11 *In the general class of estimators (9) with radially symmetric (mother) kernels, best possible asymptotic performance (under second order smoothness conditions) is achieved when distances are measured with the best metrics obtained when the kernel is uniform.*

6 Asymptotic Relative Efficiencies

The practical advantage of using density estimators that adapt to the form of the optimal metrics can be measured by computing the Asymptotic Relative Efficiency (ARE) of the optimal metric to the euclidian metric. Let us denote by $AMSE(I)$ and $AMSE(H(z))$ the expressions obtained from (122) when using the euclidian norm and the optimal norm respectively. For the Euclidean norm we get,

$$AMSE(I) = (\text{const.}) \left\{ \beta^{-1} \int_{S_I} K^2(x) dx - \lambda_1 \right\}^{\frac{4}{p+4}} f(z) \left(\frac{(\rho \,\text{tr}\; H(z))^2}{f(z)} \right)^{\frac{p}{p+4}} \tag{129}$$

where tr stands for the trace since,

$$\Delta = \int_{S_I} x^T H(z) x \, G(x) \, dx = \sum_{i,j} h_{ij}(z) \int_{S_I} x^i x^j G(x) \, dx = \rho \,\text{tr}\; H(z) \tag{130}$$

Using (123), (124) and (69) we obtain that when $H(z)$ is definite,

$$AMSE(H(z)) = \tag{131}$$
$$(\text{const.}) \left\{ \beta^{-1} \int_{S_I} K^2(x)dx - \lambda_1 \right\}^{\frac{4}{p+4}} f(z) \left(\frac{(\rho\, p\, |\det H(z)|^{1/p})^2}{f(z)} \right)^{\frac{p}{p+4}}$$

Hence, when $H(z)$ is definite the ARE is,

$$ARE = \frac{AMSE(I)}{AMSE(H(z))} = \left(\frac{\text{tr } H(z)}{p\, |\det H(z)|^{1/p}} \right)^{\frac{2p}{p+4}} \tag{132}$$

If $\xi_1, \ldots, \xi_p$ are the absolute value of the eigenvalues of $H(z)$ then we can write,

$$ARE = \left(\frac{\frac{1}{p} \sum_j \xi_j}{\left(\prod_j \xi_j \right)^{1/p}} \right)^{\frac{2p}{p+4}} = \left(\frac{\text{arith. mean of } \{\xi_j\}}{\text{geom. mean of } \{\xi_j\}} \right)^{\frac{2p}{p+4}} \tag{133}$$

It can be easily shown that the arithmetic mean is always greater or equal than the geometric mean (take logs, use the strict concavity of the logarithm and Jensen's inequality) with equality if and only if all the ξ_j's are equal. Thus, it follows from (133) that the only case in which the use of the optimal metric will not increase the efficiency of the estimation of the density at a point where the Hessian is definite is when all the eigenvalues of $H(z)$ are equal. It is also worth noticing that the efficiency increases with p, the dimension of the data space. There is of course infinite relative efficiency in the regions where the $H(z)$ is indefinite.

7 An Example: Radially Symmetric Distributions

When the true density $f(z)$ has radial symmetry it is possible to find the regions where the Hessian $H(z)$ is positive and negative definite. These models have horizons defined by the boundary between the regions where $H(z)$ is definite. We show also that when and only when the density is linear in the radius of symmetry, the Hessian is singular in the interior of a solid sphere. Thus, at the interior of these spheres it is impossible to do estimation with the best metric.

Let us denote simply by L the log likelihood, i.e.,

$$f(z) = \exp(L) \tag{134}$$

If we also denote simply by L_j the partial derivative of L with respect to z_j then,

$$\frac{\partial f}{\partial z_j} = f(z)\, L_j \tag{135}$$

and also,

$$\frac{\partial^2 f}{\partial z_i \partial z_j} = \frac{\partial f}{\partial z_i} L_j + f(z)\, L_{ij} = f(z)\, \{L_i L_j + L_{ij}\} \tag{136}$$

where we have used (135) and the definition $L_{ij} = \frac{\partial L_j}{\partial z_i}$. It is worth noticing, by passing, that (136) implies a notable connection with the so called nonparametric Fisher information $\mathcal{I}(f)$ matrix,

$$\int H(z)\, dz = \mathcal{I}(f) - \mathcal{I}(f) = 0 \tag{137}$$

our main interest here however, is the computation of the Hessian when the density is radially symmetric. Radial symmetry about a fix point $\mu \in \mathbb{R}^p$ is obtained when L (and thus f as well) depends on z only through the norm $\|z-\mu\|_{V^{-1}}$ for some symmetric positive definite $p \times p$ matrix V. Therefore we assume that,

$$L = L(-\frac{1}{2}(z-\mu)^T V^{-1}(z-\mu)) \tag{138}$$

from where we obtain,

$$L_i = \left(-v^{i\cdot}(z-\mu)\right) L' \tag{139}$$

$$L_{ij} = L'' v^{i\cdot}(z-\mu) v^{j\cdot}(z-\mu) - L' v^{ij} \tag{140}$$

where $v^{i\cdot}$ and v^{ij} denote the i-th row and ij-th entries of V^{-1} respectively. Replacing (139) and (140) into (136), using the fact that V^{-1} is symmetric and that $v^{j\cdot}(z-\mu)$ is a scalar and thus, equal to its own transpose $(z-\mu)^T v^{\cdot j}$, we obtain

$$H(z) = f(z) L' \left\{ \left(L' + \frac{L''}{L'} \right) V^{-1}(z-\mu)(z-\mu)^T - I \right\} V^{-1} \tag{141}$$

We have also assumed that L' is never zero. With the help of (141) we can now find the conditions for $H(z)$ to be definite and singular. Clearly $H(z)$

will be singular when the determinant of the matrix within curly brackets in (141) is zero. But that determinant being zero means that $\lambda = 1$ is an eigenvalue of

$$(L' + L''/L')V^{-1}(z-\mu)(z-\mu)^T \tag{142}$$

and since this last matrix has rank one its only nonzero eigenvalue must be equal to its own trace. Using the cyclical property of the trace and letting

$$y = -\frac{1}{2}(z-\mu)^T V^{-1}(z-\mu)$$

we can write,

Theorem 12 *The Hessian of a radially symmetric density is singular when and only when either $L' = 0$ or*

$$L' + \frac{d}{dy}\log L' = -\frac{1}{2y} \tag{143}$$

Notice that theorem 12 provides an equation in y after replacing a particular function $L = L(y)$. Theorem 12 can also be used to find the functions $L(y)$ that will make the Hessian singular. Integrating (143) we obtain,

$$L(y) + \log L'(y) = -\frac{1}{2}\log(|y|) + c \tag{144}$$

and solving for L', separating the variables and integrating we get,

$$L(y) = \log\left(a\sqrt{|y|} + b\right) \tag{145}$$

where a and b are constants of integration. In terms of the density equation (145) translates to,

$$f(z) = a\|z-\mu\|_{V^{-1}} + b \tag{146}$$

Hence, in the regions where the density is a straight line as a function of $r = \|z-\mu\|_{V^{-1}}$ the Hessian is singular and estimation with best metrics is not possible. Moreover, from (141) we can also obtain the regions of space where the Hessian is positive and where it is negative definite. When $L' > 0$, $H(z)$ will be negative definite provided that the matrix,

$$I - (L' + L''/L')V^{-1}(z-\mu)(z-\mu)^T \tag{147}$$

is positive definite. But a matrix is positive definite when and only when all its eigenvalues are positive. It is immediate to verify that ξ is an eigenvalue for the matrix (147) if and only if $(1-\xi)$ is an eigenvalue of the matrix (142). The matrix (142) has rank one and therefore its only nonzero eigenvalue is its trace so we arrive to,

Theorem 13 *When,*

$$L' + \frac{d}{dy}\log L' < -\frac{1}{2y} \tag{148}$$

$H(z)$ is negative definite when $L' > 0$ and positive definite when $L' < 0$. When,

$$L' + \frac{d}{dy}\log L' > -\frac{1}{2y} \tag{149}$$

$H(z)$ is indefinite.

For example when $f(z)$ is multivariate Gaussian $L(y) = y + c$ so that $L' = 1$ and the horizon is the surface of the V^{-1}-sphere of radius one i.e., $(z-\mu)^T V^{-1}(z-\mu) = 1$. Inside this sphere the Hessian is negative definite and outside the sphere the Hessian is indefinite. The results in this section can be applied to any other class of radially symmetric distributions, e.g. multivariate T which includes the Cauchy.

8 Conclusions

We have shown the existence of optimal metrics in nonparametric density estimation. The metrics are generated by the Hessian of the underlying density and they are unique in the regions where the Hessian is definite. The optimal metric can be expressed as a continuous function of the Hessian in the regions where it is indefinite. The Hessian varies continuously from point to point thus, associated to the general class of density estimators (9) there is a Riemannian manifold with the property that if the estimators are computed based on its metric the best asymptotic mean square error is minimized. The results are sufficiently general to show that these are absolute bounds on the quality of statistical inference from data.

The similarities with General Relativity are evident but so are the differences. For example, since the Hessian of the underlying density is negative definite at local maxima, it follows that there will be a horizon boundary

where the Hessian becomes singular. The cross of the boundary corresponds to a change of signature in the metric. These horizons almost always are null sets and therefore irrelevant from a probabilistic point of view. However, when the density is radially symmetric changing linearly with the radius we get solid spots of singularity. There is a qualitative change in the quality of inference that can be achieved within these dark spots. But unlike GR, not only around local maxima but also around local minima of the density we find horizons. Besides, it is not necessary for the density to reach a certain threshold for these horizons to appear. Nevertheless, I believe that the infusion of new statistical ideas into the foundations of Physics, specially at this point in history, should be embraced with optimism. Only new data will (help to) tell.

There are many unexplored promising avenues along the lines of the subject of this paper but one that is obvious from a GR point of view. What is missing is the connection between curvature and probability density, i.e. the field equation. I hope to be able to work on this in the near future.

The existence of optimal metrics in density estimation is not only of theoretical importance but of significant practical value as well. By estimating the Hessian (e.g. with kernels that can take positive and negative values, see [21]) we can build estimators that adapt to the form of the optimal norm with efficiency gains that increase with the number of dimensions. The antidote to the curse of dimensionality!

9 Acknowledgments

I would like to thank my friends in the Maximum Entropy community specially Gary Erickson for providing a stimulating environment for this meeting. I am also in debt to Ariel Caticha for many interesting conversations about life, the universe, and these things.

References

[1] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, Inc., 1998.

[2] L. G. L. Devroye and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer, New York, 1996.

[3] L. Devroye and L. Györfi, "No empirical probability measure can converge in the total variation sense for all distributions," *Annals of Statistics*, **18**, (3), pp. 1496–1499, 1990.

[4] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Math. Acad. Sci. Hungar.*, **10**, pp. 193–215, 1959.

[5] E. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, **106**, p. 620, 1957. Part II; ibid, vol 108,171.

[6] B. R. Frieden, *Physics from Fisher Information, a Unification*, Cambridge University Press, 1998.

[7] C. C. Rodríguez, "Are we cruising a hypothesis space?," in *Maximum Entropy and Bayesian Methods*, R. F. W. von der Linden, V. Dose and R. Preuss, eds., vol. 18, (Netherlands), pp. 131–140, Kluwer Academic Publishers, 1998. Also at xxx.lanl.gov/abs/physics/9808009.

[8] A. Caticha, "Change, time and information geometry," in *Maximum Entropy and Bayesian Methods*, A. Mohammad-Djafari, ed., vol. 19, Kluwer Academic Publishers, 2000. too appear. Also at math-ph/0008018.

[9] C. C. Rodriguez, "On a new class of multivariate density estimators," tech. rep., Dept. of Mathematics and Statistics, The University at Albany, 1986. (http://omega.albany.edu:8008/npde.ps).

[10] C. C. Rodriguez, "The riemannian manifold induced by a density estimator," tech. rep., Dept. of Mathematics and Statistics, The University at Albany, 1988. (http://omega.albany.edu:8008/rmide.html).

[11] K. Fukunaga and L. D. Hostetler, "Optimization of k-nearest-neighbor density estimates," *IEEE Trans. on Information Theory*, **IT-19**, pp. 320–326, May 1972.

[12] R. D. Short and K. Fukunaga, "The optimal distance measure for nearest neighbor classification," *IEEE Trans. on Information Theory*, **IT-27**, pp. 622–637, September 1981.

[13] K. Fukunaga and T. Flick, "An optimal global nearest neigbor metric," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **PAMI-6**, pp. 314–318, May 1984.

[14] K. Fukunaga and D. M. Hummels, "Bayes error estimation using parzen and k-nn procedures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **PAMI-9**, pp. 634–643, September 1987.

[15] J. P. Myles and D. J. Hand, "The multi-class metric problem in nearest neighbour discrimination rules," *Pattern Recognition*, **23**, (11), pp. 1291–1297, 1990.

[16] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," Tech. Rep. 4 Project number 21-49-004, USAF School of Aviation Medicine, Randolph Field, Tx., 1951.

[17] L. M. Le-Cam and G. Lo-Yang, *Asymptotics in Statistics: Some Basic Concepts*, Springer series in statistics, Springer-Verlag, 1990.

[18] P. Diaconnis and D. Freedman, "On the consistency of bayesian estimates (with discussions)," *Ann. Stat.*, **14**, (1), pp. 1–67, 1986.

[19] C. C. Rodríguez, "Cv-np bayesianism by mcmc," in *Maximum Entropy and Bayesian Methods*, G. J. Erickson, ed., vol. 17, Kluwer Academic Publishers, 1997. (physics/9712041).

[20] I. Ibragimov and R. Has'minskii, *Statistical Estimation*, vol. 16 of *Applications of Mathematics*, Springer-Verlag, 1981.

[21] R. S. Singh, "Nonparametric estimation of mixed partial derivatives of a multivariate density," *Journal of Multivariate Analysis*, **6**, pp. 111–122, 1976.