Weak Conditions for Shrinking Multivariate Nonparametric Density Estimators

Alessio Sancetta*

August 24, 2012

Abstract

Nonparametric density estimators on $\mathbb{R}^K$ may fail to be consistent when the sample size n does not grow fast enough relative to reduction in smoothing. For example a Gaussian kernel estimator with bandwidths proportional to some sequence h_n is not consistent if nh_n^K fails to diverge to infinity. The paper studies shrinkage estimators in this scenario and shows that we can still meaningfully use - in a sense to be specified in the paper - a nonparametric density estimator in high dimensions, even when it is not asymptotically consistent. Due to the "curse of dimensionality", this framework is quite relevant to many practical problems. In this context, unlike other studies, the reason to shrink towards a possibly misspecified low dimensional parametric estimator is not to improve on the bias, but to reduce the estimation error.

Key words: Integrated Square Error, Kolmogorov Asymptotics, Nonparametric Estimation, Parametric Model, Shrinkage.

2000 Mathematics Subject Classifications: 62G07, 62G20.

1 Introduction

Suppose f is a density function (with respect to the Lebesgue measure) with support in $\mathbb{R}^K$, and $\hat{f}_n$ is a nonparametric density estimator derived from a sample of n independent identically distributed (iid) observations from f. When n goes to infinity, it is often the

*Address: Via Flaminia Nuova 213, 00191 Roma, Italy. E-mail: <asancetta@gmail.com>. URL: <http://sites.google.com/site/wwwsancetta/>.

case that a suitable choice of $\hat{f}_n$ converges to f in some mode of convergence (e.g. Scott, 1992, and Devroye and Gyorfi, 2002). However, the number of observations required for consistency of the estimator often needs to grow exponentially with respect to K (though, exceptions may exist for some problems, e.g. Barron 1994).

Hence, in a finite sample, the performance of the nonparametric estimator might be disappointing especially if K is large. Moreover, the performance often deteriorates in the tails of the distribution. This poor finite sample behaviour can be mimicked asymptotically by saying that the estimator fails to be consistent: it is too localised relative to the sample size. This is the framework used in this paper, where no assumption is made about the consistency of the nonparametric estimator. In such cases, one could assume that $K \to \infty$ with the sample size.

In an effort to mitigate the "curse of dimensionality", many authors have studied shrunk estimators of one form or the other (e.g. Hjort and Glad, 1995, Hjort and Jones, 1996, Fan and Ullah, 1999, Mays et al., 2001, Gonzalo and Linton, 2000, Naito, 2004, Hagmann and Scaillet, 2007, El Ghouch and Genton, 2009). These papers assume consistency and derive shrunk estimators that may improve on the bias. Here the point of view is different, as the dimensionality problem can easily lead to such a poor finite sample performance that it makes sense to study the effect of shrinkage when consistency may not be obtained as a result of a nonvanishing estimation error. Hence, the present goal is to improve on the estimation error. It is worth mentioning that in this framework, the only explicit requirement on the true density is square integrability. Depending on the nonparametric density estimator that is used, other restrictions are implicitly needed: integrability of the cube of the density appears to be a sufficient requirement in most circumstances. This differs substantially from the number of regularity conditions imposed on the true unknown density as well as the nonparametric estimator in order to derive the results in the references above. For example, in the present context, K is not required to be fixed, but can grow with n.

Let $\hat{f}_n$ be a localised nonparametric estimator, so that its bias is low relative to the estimation error. Using the Gaussian kernel example with diagonal smoothing matrix proportional to h, we can have $nh^K \to c < \infty$ (using $h := h_n$ for ease of notation). Even for fix h (i.e. bias only growing linearly in K), we can think of what happens when both

K and n increase. For $c \to \infty$ we need n growing exponentially faster than K. Mutatis mutandis, this framework is conceptually similar to Kolmogorov asymptotics for vector valued statistics (e.g. Aivasian et al., 1989). In order to reduce the estimation error, we shrink $\hat{f}_n$ towards a parametric model g_θ indexed in a compact Euclidean set Θ. In this case the estimator becomes $\tilde{f}_n = \alpha g_\theta + (1-\alpha)\,\hat{f}_n$, $\alpha \in [0,1]$, $\theta \in \Theta$. Mutatis mutandis, this is similar to large dimensional covariance shrinkage problems (e.g. Ledoit and Wolf, 2004, Sancetta, 2008). The problems are related, as the nonparametric estimator can be made nearly unbiased, though very noisy in a finite sample when K is large. Shrinking $\hat{f}_n$ towards the parametric model $(g_\theta)_{\theta\in\Theta}$ will reduce the variability of the estimator at the cost of an increase in bias when $f \notin \{g_\theta : \theta \in \Theta\}$. This statement will be made precise below.

Olkin and Spiegelman (1987) have already studied a maximum likelihood estimator of $\tilde{f}_n$, though in a different context. Here, the estimation of α is not based on maximum likelihood, avoiding Olkin and Spiegelman (1987)'s restrictive conditions that, for example, would prevent g_θ from being a Gaussian density and would require the nonparametric estimator to be consistent, ruling out the large K dimensional problem addressed here. These restrictions are used by Olkin and Spiegelman (1987) because their goal is to devise a method that is robust against misspecification of the parametric model, hence as a way to reduce any possible bias. Here, the focus is on the nonparametric estimator being combined to a low dimensional - hence likely to be misspecified - parametric model to reduce the estimation error.

A simulation study in Section 3 shall also be used to highlight the behaviour of the estimator when the parametric model is highly biased. In this case, some of the conclusions are that the estimator $\tilde{f}_n$ is less sensitive to the choice of bandwidth than a kernel density estimator. Moreover, when we choose an "ideal" bandwidth for both $\tilde{f}_n$ and the kernel density, $\tilde{f}_n$ still compares favourably. Alternative semiparametric methods to improve on nonparametric density estimators have been considered in the last two decades (e.g. Hjort and Glad, 1995, Hjort and Jones, 1996, and Naito, 2004, who brought unity for the different methods by local L_2 fitting; more recently also Hagmann and Scaillet, 2007). These methods rely on a multiplicative correction term. To the author's experience, these estimators perform remarkably well in one dimension, while they deteriorate in higher

dimensions, occasionally performing worse than simple kernel smoothers and/or being sensitive to the choice of bandwidth. The simulation study of this paper will consider one of these estimators for comparison reasons.

We introduce some notation. The symbol $\mathbb{P}_n$ stands for the empirical measure, e.g. $\mathbb{P}_n X = n^{-1}\sum_{i=1}^n X_i$, where $X_1, ..., X_n$ are iid copies of X. The symbol $\lesssim$ stands for inequality up to a finite absolute constant, $\asymp$ implies equality in order of magnitude; $\wedge$ and $\vee$ are used for the minimum and maximum between left and right hand side, respectively. Finally, $\|\bullet\|_{2,\lambda}$ and $\|\bullet\|_{2,\mathbb{P}}$ are the norms with respect to the Lebesgue measure λ and the true measure $\mathbb{P}$.

2 Shrinking the Density Estimator

Given the sample $X_1, ..., X_n$, we estimate the nonparametric estimator $\hat{f}_n$. The best parametric fit from $(g_\theta)_{\theta\in\Theta}$ is denoted by g_{θ_0}. Clearly,

$$\min_{\alpha\in[0,1]} \left\|\alpha g_{\theta_0} + (1-\alpha)\,\hat{f}_n - f\right\|_{2,\lambda} \leq \left\|\hat{f}_n - f\right\|_{2,\lambda}. \tag{1}$$

The right hand side (r.h.s.) is the integrated square error (ISE) for the nonparametric density estimator. Hardle and Marron (1986) show that under reasonable assumptions, ISE and mean square error are asymptotically the same. In the present context, it is easier to work with the ISE. The r.h.s. of (1) cannot achieve the root-n parametric rate of convergence.

Example 1 *Suppose f has support in $\mathbb{R}^K$ and $\hat{f}_n$ is its estimator based on a first order kernel. Then, under regularity conditions,*

$$\left\|\hat{f}_n - f\right\|_{2,\lambda} \asymp n^{-2/(4+K)},$$

in probability (e.g. Scott, 1992). It is clear that if n is not exponentially larger than K, the estimator cannot be consistent, e.g. $K = 2\ln n - 4$ as $n \to \infty$ makes the ISE bounded away from zero for any sample size.

Shrinking towards the parametric model $(g_\theta)_{\theta\in\Theta}$ might improve on this slow rate of convergence. The ideal shrinking parameter α is given by the following:

Proposition 1 *Suppose* $\tilde{f}_n = \alpha g_\theta + (1-\alpha)\hat{f}_n$. *Then,*

$$[(\alpha_n \vee 0) \wedge 1] = \arg\min_{\alpha\in[0,1]} \left\| \tilde{f}_n - f \right\|_{2,\lambda},$$

where

$$\alpha_n := \frac{\int \left[g_{\theta_0}(x) - \hat{f}_n(x)\right] f(x)\,dx - \int \left[g_{\theta_0}(x) - \hat{f}_n(x)\right] \hat{f}_n(x)\,dx}{\int \left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 dx}.$$

Proof. Differentiating and factoring terms in α,

$$\begin{aligned} & \frac{d\left\|\alpha g_{\theta_0} + (1-\alpha)\hat{f}_n - f\right\|_{2,\lambda}^2}{d\alpha} \\ = \; & \alpha \int \left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 du + \int \left[\hat{f}_n(x) - f(x)\right]\left[g_{\theta_0}(x) - \hat{f}_n(x)\right] dx \\ = \; & 0. \end{aligned}$$

Solving for α, subject to the constraint, gives the result. ■

Remark 1 *To ease the notation, we shall assume* $\alpha_n \in [0,1]$ *so that* $\alpha_n = [(\alpha_n \vee 0) \wedge 1]$.

The result of Proposition 1 gives a random value for α because it depends on $\hat{f}_n$. However, by definition α_n satisfies

$$\left\|\alpha_n g_{\theta_0} + (1-\alpha_n)\hat{f}_n - f\right\|_{2,\lambda} \leq \left\|\hat{f}_n - f\right\|_{2,\lambda}. \tag{2}$$

If $\left\|\hat{f}_n - f\right\|_{2,\lambda} \to 0$ (in probability) the procedure can also lead to consistent estimation, but with possibly smaller ISE, as shown in the references cited in the Introduction.

Clearly, we do not know the best parametric approximation in $(g_\theta)_{\theta\in\Theta}$ and we do not know the integral of $\left[g_{\theta_0}(x) - \hat{f}_n(x)\right] f(x)$ with respect to x. Hence, we shall find sample estimators for these. In particular, θ_0 is replaced by an estimator, say $\hat{\theta}$, (e.g. the maximum likelihood estimator), while

$$\int g_{\theta_0}(x) f(x)\,dx = \mathbb{E} g_{\theta_0}(X)$$

can be approximated by its sample counterpart $\mathbb{P}_n g_\theta(X)$. However,

$$\int \hat{f}_n(x) f(x)\,dx = \mathbb{E}\hat{f}_n(X)$$

should not be replaced by $\mathbb{P}_n \hat{f}_n(X)$ because this quantity is biased and has poor variance properties. A suitable sample estimator can be found using classic leave out estimators.

Divide $\{1,...,n\}$ into $V \in \mathbb{N}$ blocks $A_1,...,A_V$ of mutually exclusive sets, with $1/V = q \in (0,1)$. Hence, $\#A_v = nq$ is the cardinality of A_v. Then, the problem is solved by using the leave out estimator

$$\mathbb{P}_n\left(\hat{f}_n|q\right) := \frac{1}{V}\sum_{v=1}^{V}\frac{1}{qn}\sum_{i\in A_v}\hat{f}_{n(1-q)}\left(X_i;(X_j)_{j\in A_v^c}\right) \tag{3}$$

where $\hat{f}_{n(1-q)}\left(X_i;(X_j)_{j\in A_v^c}\right)$ is the nonparametric estimator $\hat{f}_n$ based on $(X_j)_{j\in A_v^c}$ only, where A_v^c is the complement of A_v so that $\#A_v^c = n(1-q)$ (e.g. van der Laan and Dudoit, 2003). An explicit representation is given in Remark 6, below. In the case $nq = 1$, we have the usual leave one out estimator. However, leaving out a fraction of the sample n is often found to perform well, e.g. $q = .1$ (see discussion in van der Laan and Dudoit, op.cit.). In our framework, we will see that the leave one out estimator (i.e. $nq = 1$) is not a good idea.

We denote the feasible estimator of α_n by

$$\hat{\alpha}_n := \frac{\mathbb{P}_n g_{\hat{\theta}}(X) - \mathbb{P}_n\left(\hat{f}_n|q\right) - \int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]^2 dx}. \tag{4}$$

Remark 2 *Again, for notational convenience we shall assume* $\hat{\alpha}_n \in [0,1]$.

The following conditions are used to derive the results of the paper.

Condition 1 $\left(\hat{\theta}_n\right)_{n\in\mathbb{N}}$ *is a sequence of random elements (the estimators for the parameter of the model) with values inside a compact set* $\Theta \subset \mathbb{R}^S$ *such that*

$$\left|\hat{\theta}_n - \theta_0\right| = O_p\left(n^{-1/2}\right).$$

Condition 2 *There is an open ball* B_0 *centered at* θ_0*, and a* $q \in [1,2]$ *and a* $p \in [1,\infty]$ *with* $p^{-1} + q^{-1} = 1$*, such that,*

$$\sup_{\theta\in B_0}\|g_\theta\|_{p,\lambda} + \sup_{\theta\in B_0}\|\nabla_{\theta_s} g_\theta\|_{p,\lambda} < \infty \ \ (\forall s)\,, \ \left\|\hat{f}_n\right\|_{q,\lambda} < \infty \ a.s.$$

and

$$\sup_{\theta\in B_0}\|g_\theta\|_{2,\mathbb{P}} + \sup_{\theta\in B_0}\|\nabla_{\theta_s} g_\theta\|_{1,\mathbb{P}} < \infty \ \ (\forall s)$$

where $\nabla_{\theta_s} g_\theta$ *is the* s^{th} *element of the gradient of* g_θ *with respect to* θ*, evaluated at* θ.

Condition 3 *There exists a function* $\psi_n : \mathbb{R}^K \times \mathbb{R}^K \times \mathbb{N} \to \mathbb{R}$ *such that* $\hat{f}_n$ *admits the following representation*

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \psi_n(x, X_i),$$

where $\mathbb{E}|\psi_n(X_1, X_2)|^2 < \infty$ *for any fixed* n.

Condition 4 $g_{\theta_0}(x) \neq \hat{f}_n(x)$ *for any* n; $\|f\|_{2,\lambda} < \infty$.

Remark 3 *Condition 1 is the standard consistency of parametric estimators for the pseudo true value* θ_0.

Remark 4 *Condition 2 imposes smoothness restrictions on the parametric model around the pseudo true value. The required level of smoothness is a function of how localised is the nonparametric estimator. A very localised nonparametric estimator does require to shrink towards a smoother parametric model. While the* L_1 *and* L_2 *norm of the pseudo true parametric model with respect to the true measure is unknown, the user can choose* $(g_\theta)_{\theta\in\Theta}$ *such that Condition 2 is likely to be satisfied in practice.*

Example 2 *By Minkowski inequality*

$$\left\|\hat{f}_n\right\|_{q,\lambda} \leq \left\|(1-\mathbb{E})\hat{f}_n\right\|_{q,\lambda} + \left\|\mathbb{E}\hat{f}_n\right\|_{q,\lambda}.$$

Consider the r.h.s. of the above display. For the Gaussian kernel example, the second term is bounded if f *is in* L_2. *The first term is always bounded for* $q = 1$. *However, this requires a very smooth parametric model (i.e.* $p = \infty$ *in Condition 2). On the other hand, for* $q = 2$, *the first term in the r.h.s. of the display is almost surely bounded if* $\lim_n nh^K > 0$. *Under this condition, we can impose less restrictions on the parametric model (i.e.* $p = 2$*).*

Remark 5 *Condition 3 is satisfied by most nonparametric density estimators: kernels, orthogonal polynomials, Bernstein polynomials, etc.. Many estimators satisfy even stronger conditions. In the case of a bounded kernel density estimator,* ψ_n *is such that* $|\psi_n|_\infty := \sup_{x,y\in\mathbb{R}^K} |\psi_n(x,y)| \asymp h_n^{-K}$ *where* h_n *is the bandwidth in one dimension. For polynomials over compact intervals,* $|\psi_n|_\infty$ *is of the same order as the order of the polynomial.*

Remark 6 *By Condition 3, in (3) we have*

$$\hat{f}_{n(1-q)}\left(x;(X_i)_{i\in A_v^c}\right) := \frac{1}{n(1-q)}\sum_{i\in A_v^c}\psi_n(x,X_i).$$

Remark 7 *Condition 4 is technical. The first part is required for identification of α_n. Moreover, for obvious reasons f needs to be in L_2.*

To control the error in the foregoing approximation, we define the following:

$$\zeta_n := 1 + Var\left(\mathbb{E}_X\psi_n(X,X_1)\right) + Var\left(\mathbb{E}_X\psi_n(X_1,X)\right) + (nq)^{-1}Var\left(\psi_n(X_1,X_2)\right), \quad (5)$$

where X is an independent copy of X_1 and $\mathbb{E}_X$ stands for expectation with respect to X. For $\psi_n(x,y)$ symmetric, the above expression simplifies. Note that ζ_n is artificially defined adding a 1 to make sure that $\inf_n \zeta_n > 0$. This can be equivalently achieved by imposing a suitable lower bound condition on ψ_n uniformly in n to make ensure that $\inf_n Var\left(\mathbb{E}_X\psi_n(X,X_1)\right) > 0$. We have the following:

Theorem 1 *Under Conditions 1, 2, 3, and 4,*

$$\hat{\alpha}_n = \alpha_n + O_p\left(\sqrt{\zeta_n/n}\right),$$

and there is a finite positive constant C, independent of f_n, such that

$$\left\|\hat{\alpha}_n g_{\hat{\theta}} + (1-\hat{\alpha}_n)\hat{f}_n - f\right\|_{2,\lambda} \leq \left\|\alpha_n g_{\theta_0} + (1-\alpha_n)\hat{f}_n - f\right\|_{2,\lambda} + C\left(1 + \left\|\hat{f}_n - f\right\|_{2,\lambda}\right)\sqrt{\frac{\zeta_n}{n}}$$

in probability, which by (2) also implies

$$\left\|\hat{\alpha}_n g_{\hat{\theta}} + (1-\hat{\alpha}_n)\hat{f}_n - f\right\|_{2,\lambda} \leq \left\|\hat{f}_n - f\right\|_{2,\lambda} + C\left(1 + \left\|\hat{f}_n - f\right\|_{2,\lambda}\right)\sqrt{\frac{\zeta_n}{n}},$$

in probability

3 Discussion and Simulation Study

Theorem 1 shows that what would determine the success of the procedure is that $\zeta_n = o\left(n\left\|\hat{f}_n - f\right\|_{2,\lambda}^2\right)$, in which case, the ISE of the shrunk estimator is of smaller order of magnitude than the original ISE. Depending on the nonparametric estimator, this implies extra restrictions on f as we need $Var\left(\mathbb{E}_X\psi_n(X,X_1)\right) < \infty$. For a Gaussian kernel

density estimator, this requires f^3 to be integrable. In general, if $Var\left(\mathbb{E}_X \psi_n\left(X, X_1\right)\right) < \infty$ in (5), one should think about some very unnatural (non-consistent) estimators for $\zeta_n = o\left(n\left\|\hat{f}_n - f\right\|_{2,\lambda}^2\right)$ not to be true. A specific example can provide some further insights into this claim:

Example 3 *Suppose $\psi_n(x,y)$ is the K dimensional Gaussian kernel with smoothing matrix proportional to h. Under regularity conditions on f (including $\|f\|_{3,\lambda} < \infty$), if we leave out a fix fraction of the sample (i.e. q is fixed), direct calculations give $\zeta_n \lesssim 1 + n^{-1}h^{-K}$, and $\left\|\hat{f}_n - f\right\|_{2,\lambda}^2 \asymp h + n^{-1}h^{-K}$, so that $\zeta_n = o\left(n\left\|\hat{f}_n - f\right\|_{2,\lambda}^2\right)$ as long as $nh \to \infty$. Hence, the shrunk estimator is guaranteed to perform asymptotically as well if not better than $\hat{f}_n$. Note that in this example we can have $K \to \infty$ with n; all we need is $\lim_n nh = \infty$.*

From the above example, we deduce that the nonparametric estimator has a second order effect on $\hat{\alpha}_n$ and the ISE of the shrunk estimator. Moreover, we can see why the leave one out estimator is not the best choice: the last term in ζ_n is

$$\frac{Var\left(\psi_n\left(X_1, X_2\right)\right)}{nq} = O\left(\frac{h^{-K}}{nq}\right) = O\left(h^{-K}\right)$$

if $nq = O(1)$, so that in the previous example, we have $\zeta_n \lesssim 1 + h^{-K}$, instead. We shall still have $\zeta_n = o\left(n\left\|\hat{f}_n - f\right\|_{2,\lambda}^2\right)$, for $nh \to \infty$, but in a finite sample, the difference might not be negligible.

Under consistency, asymptotic normality of the shrunk estimator can be studied. Unfortunately, the present context is not amenable to such analysis: Theorem 1 does not say anything about the consistency of the nonparametric estimator $\hat{f}_n$, as all the analysis is relative to $\left\|\hat{f}_n - f\right\|_{2,\lambda}^2$ without any consistency condition on it. In fact, via Condition 2, Theorem 1 allows us to consider different situations where divergence is also possible.

The estimator should have an advantage over usual nonparametric estimators in terms of variance and not bias. The following experimental results show that α_n also allows us to counterbalance either oversmoothing or undersmoothing in the nonparametric density estimator $\hat{f}_n(x)$.

3.1 Experimental Results: Shrinking Towards a Very Biased Parametric Model

It is clear that if we choose a low dimensional parametric model whose bias is relatively low, the shrunk estimator will perform well even when K increases. Hence, it is of interest to understand the loss incurred in using the shrunk estimator when the parametric model is highly biased (misspecified). In this case, the reduction in estimation error is more than compensated by the increase in bias. Therefore, we cannot expect the shrunk estimator to perform better than a nonparametric estimator, but we still hope the performance to be reasonable even in these extreme cases. The goal of this experiment is to evaluate the estimator in some sort of worse case scenario. The relevant question is, how robust is the shrunk estimator to high levels of misspecification in the parametric term? As mentioned in the introduction, shrunk estimators have already been studied by other authors and numerous simulation results have been produced. Hence, here we are trying to look at the problem from a different point of view.

To this end, we simulate data from a mixtures of Gaussian and exponential density functions:

$$pdf_X(x) = p\phi(x) + (1-p)\{x \geq 0\}\exp\{-x\},$$

where $\phi(x)$ is the standard normal density and $\{x \geq 0\}\exp\{-x\}$ is the exponential density with mean one (and clearly positive support only). We also simulate data from the K-dimensional analog ($K = 2, 3$):

$$pdf_X(x_1, ..., x_K) = p\phi_K(x_1, ..., x_K; \rho) + (1-p)\prod_{k=1}^{K}\{x_k \geq 0\}\exp\{-x_k\},$$

where $\phi_K(x_1, x_2; \rho)$ is the K-dimensional standard Gaussian density with covariance matrix with diagonal entries equal to one and off diagonal entries equal to $\rho = .25$ (i.e. equal correlation between variables). We consider the following cases: $p = .25, .5, .75$ and samples of $n = 40, 80$ observations. A sample size of $n = 40$ is considered to be quite small for a three dimensional kernel density estimator. Recall that $nh^K \to \infty$ is needed, because the variance of the kernel density estimator is $O\left(n^{-1}h^{-K}\right)$. Therefore, what matters is not n but nh^K. For example, when $h = .1$, $n = 40$, and $K = 3$, we have $nh^K = .04$.

The density is estimated by kernel smoothing with Gaussian kernel (NP estimator) and by a Gaussian density with mean and variance matrix estimated by method of moments (P

estimator). For the latter estimator, misspecification is quite pronounced even for $p = .75$ becoming very pronounced for $p = .25$. Figure 1 shows the quite evident asymmetry of the density in one dimension, when $p = .75, .5, .25$. The Gaussian density ($p = 1$) is also plotted for reference.

[Figure 1 Here]

We shrink the NP estimator towards the biased P estimator using the shrinkage parameter in (4) and will refer to the shrunk estimator as the S estimator where, in (3), $q = .1$. We also shrink the NP estimator towards the P estimator for fixed $\alpha = 0, .1, .2, ..., 1$ and, for each p, n and K, report results for the best performing α referring to this as the D estimator. For comparison, we also compute the nonparametric estimator of Hjort and Glad (1995) with Gaussian parametric term and refer to it as the HG estimator. This is a special case of the L_2 fitting density estimators studied in Naito (2004). The latter density estimator usually improves on the asymptotic bias of the fully nonparametric estimator, but does not provide an improvement on the asymptotic variance. It appears that the only way to reduce variance in the nonparametric estimator is to shrink it towards a less variable constrained estimator. Estimators based on multiplicative correction do not possess this property. For this reason it is instructive to compare estimators that try to improve on fully nonparametric estimators but by different routes. Note that as p decreases we move even further away from the P estimator and the leading term in the HG estimator.

The bandwidth is chosen to be the standard deviation in the Gaussian kernel smoother and it is set equal to $h = .1, .3, .5, .7, .9$ times the identity matrix. To compute the ISE we used Monte Carlo integration based on 1000 simulated uniform random variables in $[-5, 5]$ when $K = 1$. When $K > 1$, the ISE is computed by Monte Carlo integration based on 10000 simulated uniform random variables in $[-5, 5]^K$. Results are in Tables 1-6, for the $K = 1, 2, 3$ dimensions, respectively. Tables 1-6 report the integrated square error, averaged over 1000 samples for the S, NP, HJ and D estimators, together with standard errors (rounded to second decimal place). The percentage relative improvement in average loss (PRIAL) of the estimators is also reported (rounded to first decimal place),

where

$$PRIAL\left(w\right) := 100\frac{\mathbb{E}\left\|\hat{f}_n - f\right\|_{2,\lambda}^2 - \mathbb{E}\left\|w - f\right\|_{2,\lambda}^2}{\mathbb{E}\left\|\hat{f}_n - f\right\|_{2,\lambda}^2},$$

and w is the estimator (i.e. the S, NP, HG and D estimator). Hence, $PRIAL\left(\text{NP}\right) = 0$ by definition, so that we measure the improvement relative to the NP estimator. All expectations are of course approximated using the mean over the 1000 simulated samples.

[Tables 1-6 Here]

The results show that the performance of the S estimator is often comparable to the NP and HG estimators. This is particularly so in high dimensions. In high dimensions, when n is small, as in here, nonparametric estimators perform poorly because of high variability, unless we oversmooth. The results confirm the theory in suggesting that the S estimator can be considered as a competitor to NP estimators, particularly in high dimensions and when a P estimator is useful to provide further structure for the data analysis. The PRIAL of the S estimator seem to confirm this, particularly when $p > .25$. When $K = 3$ the S estimator is usually superior to the HG estimator, which often performed worse than the NP estimator (as already anticipated in the Introduction). It is evident that the S estimator improves on the NP and HG estimators when nh^K is small even for the very misspecified parametric model (i.e. $p = .25$).

The performance of the HG estimator was very poor when $h = .9$. An explanation for negative outcomes when the bandwidth is large can be provided. Suppose that the kernel is bounded below by a constant c for all sample values when the bandwidth is large. In this case, the HG estimator is bounded below by

$$\phi\left(x\right)\frac{1}{n}\sum_{i=1}^{n}\frac{\psi_n\left(x, X_i\right)}{\phi\left(X_i\right)} \geq \phi\left(x\right)c\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\phi\left(X_i\right)}.$$

The right hand side can be particularly large in some occasions, as shown in Table 6 when $h = .9$ and $n = 80$. (Note that we used the same seed numbers for all computations, hence, in the sample when $n = 80$ there must have been at least an observation that led to the aforementioned phenomenon). Hjort and Glad (1995) suggest to trim the multiplicative term to avoid this instability. Since trimming involves and additional parameter to be tuned, for comparison reasons it was preferred to avoid this, as this problem only occurred

for $h = .9$. The goal of this experiments is to shed some light into the behaviour of these estimators in some special circumstances. Of course, the use of a less biased parametric model would have shown more substantial improvement in both the S and HG estimator relative to the NP estimator.

4 Further Remarks

The above experiment shows that the best estimator really depends on the situation and the extent of previous knowledge of the problem at hand. For high dimensional problems it is quite difficult to pick up a unique best model and/or estimation approach. Hence, a shrunk procedure could be considered as a relative safe option for difficult problems. The asymmetry in the true distribution was not captured at all by the parametric model. Nevertheless the increase in bias due to "wrongly choosing" the parametric estimator did not lead to considerable loss in performance in the S estimator.

The main feature of a shrunk estimator is robustness (also in terms of bandwidth selection, in this context). Indeed, a shrunk estimator is just a simple version of model combination and many of the insights of that literature can also be applied here (e.g. Timmermann, 2006, for a review). In the context of model combination, it is well known that combining models that are quite different might provide the highest benefit.

One may actually decide to shrink a nonparametric estimator towards multiple parametric models. This might be a more stable approach than selecting a single parametric model to shrink to. Indeed, it is well known that subset model selection tends to be noisier than model combination (e.g. Breiman, 1996). Some of these remarks will be considered in some future studies.

Finally, this paper was only concerned with estimation starting from a nonparametric estimator and not with inference. Indeed, one could utilise $\hat{\alpha}_n$ to check goodness of fit of the parametric model. This requires derivation of the asymptotic distribution of the shrinkage parameter. Under the null that the true density $f \in (g_\theta)_{\theta \in \Theta}$, then, $\alpha_n \to 1$, which is equivalent to a test of the true parameter at the boundary under the null. It is well known (e.g. Andrews, 1999) that in these case, the asymptotic distribution of the estimator is not normal. Analysis of this problem shall be the subject of future research.

5 Proof of Theorem 1

For ease of reference, we state the mean value theorem.

Lemma 1 *Suppose* $r : \Theta \to \mathbb{R}$. *Inside* Θ,

$$r\left(\hat{\theta}\right) = r\left(\theta_0\right) + \nabla_\theta r\left(\theta_*\right)'\left(\hat{\theta} - \theta_0\right),$$

where $\theta_* = \theta\left(\rho\right) = \rho\hat{\theta}_n + \left(1-\rho\right)\theta_0$, $\rho \in [0,1]$, *and* $\nabla_\theta r\left(\theta_*\right)$ *is the gradient of* $r\left(\theta\right)$ *evaluated at* θ_*, *and the prime is used for the transpose.*

We show that the estimated parametric leading term can be replaced by the best parametric approximation.

Lemma 2 *Under Conditions 1 and 2,*

$$\int\left[g_{\hat{\theta}}\left(x\right) - \hat{f}_n\left(x\right)\right]\hat{f}_n\left(x\right)dx = \int\left[g_{\theta_0}\left(x\right) - \hat{f}_n\left(x\right)\right]\hat{f}_n\left(x\right)dx + O_p\left(n^{-1/2}\right).$$

Proof. By Lemma 1

$$\int\left[g_{\hat{\theta}}\left(x\right) - \hat{f}_n\left(x\right)\right]\hat{f}_n\left(x\right)dx = \int\left[g_{\theta_0}\left(x\right) - \hat{f}_n\left(x\right)\right]\hat{f}_n\left(x\right)dx + \int\nabla_\theta g_{\theta_*}\left(x\right)'\left(\hat{\theta} - \theta_0\right)\hat{f}_n\left(x\right)dx.$$

By Holder and Minkowski inequalities,

$$\begin{aligned}\left|\int\nabla_\theta g_{\theta_*}\left(x\right)'\left(\hat{\theta} - \theta_0\right)\hat{f}_n\left(x\right)dx\right| &\leq \max_{s\in\{1,...,S\}}\left|\hat{\theta}_s - \theta_{0s}\right|\sum_{s=1}^{S}\left\|\nabla_{\theta_s} g_{\theta_*}\right\|_{p,\lambda}\left\|\hat{f}_n\right\|_{q,\lambda} \\ &= O_p\left(n^{-1/2}\right),\end{aligned}$$

by Conditions 1 and 2. ■

Lemma 3 *Under Conditions 1 and 2,*

$$\int\left[g_{\hat{\theta}}\left(x\right) - \hat{f}_n\left(x\right)\right]^2 dx = \int\left[g_{\theta_0}\left(x\right) - \hat{f}_n\left(x\right)\right]^2 dx + O_p\left(n^{-1/2}\right)$$

Proof. By Lemma 1

$$\begin{aligned}\int\left[g_{\hat{\theta}}\left(x\right) - \hat{f}_n\left(x\right)\right]^2 dx &= \int\left[g_{\theta_0}\left(x\right) - \hat{f}_n\left(x\right)\right]^2 dx \\ &\quad +2\int\left(\hat{\theta} - \theta_0\right)'\nabla_\theta g_{\theta_*}\left(x\right)\left[g_{\theta_*}\left(x\right) - \hat{f}_n\left(x\right)\right]dx \\ &\leq \int\left[g_{\theta_0}\left(x\right) - \hat{f}_n\left(x\right)\right]^2 dx \\ &\quad +2\max_{s\in\{1,...,S\}}\left|\hat{\theta}_s - \theta_{s0}\right|\sum_{s=1}^{S}\left\|\nabla_{\theta_s} g_{\theta_*}\right\|_{p,\lambda}\left\|g_{\theta_*} - \hat{f}_n\right\|_{q,\lambda}\end{aligned}$$

by similar arguments as in the proof of Lemma 2. Since

$$\left\|g_{\theta_*} - \hat{f}_n\right\|_{q,\lambda} \leq \sup_{\theta \in B_0} \|g_\theta\|_{q,\lambda} + \left\|\hat{f}_n\right\|_{q,\lambda},$$

then,

$$\max_{s \in \{1,\ldots,S\}} \left|\hat{\theta}_s - \theta_{s0}\right| \sum_{s=1}^{S} \|\nabla_{\theta_s} g_{\theta_*}\|_{p,\lambda} \left\|g_{\theta_*} - \hat{f}_n\right\|_{q,\lambda} = O_p\left(n^{-1/2}\right)$$

by Conditions 1 and 2. ■

Lemma 4 *Under Conditions 1 and 2,*

$$\mathbb{P}_n g_{\hat{\theta}}(X) = \int g_{\theta_0}(x) f(x)\, dx + O_p\left(n^{-1/2}\right).$$

Proof. By Lemma 1

$$\mathbb{P}_n g_{\hat{\theta}}(X) = \mathbb{P}_n g_{\theta_0}(X) + \sum_{s=1}^{S} \left(\hat{\theta}_s - \theta_{s0}\right) \mathbb{P}_n \nabla_{\theta_s} g_{\theta_*}(X).$$

Hence, by Condition 2 and Chebyshev's inequality

$$\mathbb{P}_n g_{\theta_0}(X) = \int g_{\theta_0}(x) f(x)\, dx + O_p\left(n^{-1/2}\right),$$

and

$$\sum_{s=1}^{S} \left(\hat{\theta}_s - \theta_{s0}\right) \mathbb{P}_n \nabla_{\theta_s} g_{\theta_*}(X) \leq \max_{s \in \{1,\ldots,S\}} \left|\hat{\theta}_s - \theta_{0s}\right| \sum_{s=1}^{S} \left|\mathbb{P}_n \nabla_{\theta_s} g_{\theta_*}(X)\right| = O_p\left(n^{-1/2}\right),$$

by Condition 1 and 2. ■

Finally we have the following consistency of the cross-validated estimator.

Lemma 5 *Suppose ζ_n is as in Theorem 1,*

$$\mathbb{P}_n\left(\hat{f}_n | q\right) = \int \hat{f}_n(x) f(x)\, dx + O_p\left(\sqrt{\zeta_n / n}\right).$$

Proof. To avoid trivialities in the notation, assume $V = 1/q \in \mathbb{N}$ and $qn \in \mathbb{N}$. With no loss of generality, assume that $\psi_n(x, y)$ is symmetric, as if not it can always be replaced by a symmetrised version (e.g. Arcones and Giné, 1992, eq 2.4). Note that

$$\begin{aligned}
\mathbb{P}_n\left(\hat{f}_n | q\right) &= \frac{1}{V} \sum_{v=1}^{V} \frac{1}{qn} \sum_{i \in A_v} \frac{1}{n(1-q)} \sum_{j \in A_v^c}^{n} \psi_n(X_i, X_j) \\
&= \frac{1}{V(V-1)} \sum_{1 \leq v_1 \neq v_2 \leq V} \left(\sum_{i \in A_{v_1}} \sum_{j \in A_{v_2}} \frac{\psi_n(X_i, X_j)}{n^2 q^2} \right),
\end{aligned}$$

which has a representation as a U-statistic of order 2 because the sets A_{v_1} and A_{v_2} do not overlap. Hence, computing the variance using the Hoeffding's decomposition of U statistics we have (e.g. Serfling, 1980, Lemma A, p.183)

$$Var\left(\mathbb{P}_n\left(\hat{f}_n|q\right)\right) \lesssim \frac{1}{V} Var\left(\sum_{i\in A_{v_1}}\sum_{j\in A_{v_2}} \frac{\psi_n\left(X_i, X_j\right)}{n^2q^2}\right).$$

By direct calculation (without assuming symmetrization) we have

$$\begin{aligned} & \frac{1}{V} Var\left(\sum_{i\in A_{v_1}}\sum_{j\in A_{v_2}} \frac{\psi_n\left(X_i, X_j\right)}{n^2q^2}\right) \\ = & \frac{1}{V}\left(\frac{Cov\left(\psi_n\left(X_1, X_2\right), \psi_n\left(X_1, X_3\right)\right) + Cov\left(\psi_n\left(X_1, X_2\right), \psi_n\left(X_3, X_2\right)\right)}{2nq} + \frac{Var\left(\psi_n\left(X_1, X_2\right)\right)}{\left(nq\right)^2}\right) \\ \lesssim & \frac{1}{n}\zeta_n, \end{aligned}$$

for ζ_n as defined in (5) and we deduce that

$$\mathbb{P}_n\left(\hat{f}_n|q\right) = \mathbb{E}\mathbb{P}_n\left(\hat{f}_n|q\right) + O_p\left(\sqrt{\zeta_n/n}\right) = \mathbb{E}\psi_n\left(X_1, X_2\right) + O_p\left(\sqrt{\zeta_n/n}\right)$$

Hence, it is sufficient to show that

$$\int \hat{f}_n\left(x\right) f\left(x\right) dx = \mathbb{E}\psi_n\left(X_1, X_2\right) + O_p\left(\sqrt{\zeta_n/n}\right).$$

Suppose X is a copy of X_1 independent of $X_1, ..., X_n$. Then, using $\mathbb{E}_X$ for expectation with respect to X only,

$$\begin{aligned} \int \hat{f}_n\left(x\right) f\left(x\right) dx &= \mathbb{E}_X \hat{f}_n\left(X\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}_X \psi_n\left(X, X_j\right). \end{aligned}$$

By the Chebyshev's inequality, $\mathbb{E}_X \hat{f}_n\left(X\right) = \mathbb{E}\psi_n\left(X_1, X_2\right) + O_p\left(\sqrt{Var\left(\mathbb{E}_X \psi_n\left(X, X_1\right)\right)/n}\right)$. Hence,

$$\mathbb{P}_n\left(\hat{f}_n|q\right) = \int \hat{f}_n\left(x\right) f\left(x\right) dx + O_p\left(\sqrt{\zeta_n/n}\right)$$

noting that

$$Var\left(\mathbb{E}_X \psi_n\left(X, X_1\right)\right) = Cov\left(\psi_n\left(X_1, X_2\right), \psi_n\left(X_3, X_2\right)\right) \leq Var\left(\psi_n\left(X_1, X_2\right)\right),$$

by stationarity. ■

The following two lemmata give Theorem 1. First, we show consistency of the shrinkage parameter.

Lemma 6 *Under the conditions of Theorem 1,*

$$\hat{\alpha}_n = \alpha_n + O_p\left(\sqrt{\zeta_n/n}\right).$$

Proof. We need to show

$$\begin{aligned}
&\frac{\mathbb{P}_n g_{\hat{\theta}}(X) - \mathbb{P}_n\left(\hat{f}_n|q\right) - \int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]^2 dx} \\
= &\frac{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right] f(x)\,dx - \int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 dx} + O_p\left(\sqrt{\zeta_n/n}\right).
\end{aligned}$$

By Lemma 3, the fact that $g_{\theta_0}(x) \neq \hat{f}_n(x)$ and that the numerator is $O_p(1)$, an application of the delta method gives

$$\begin{aligned}
&\frac{\mathbb{P}_n g_{\hat{\theta}}(X) - \mathbb{P}_n\left(\hat{f}_n|q\right) - \int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]^2 dx} \\
= &\frac{\mathbb{P}_n g_{\hat{\theta}}(X) - \mathbb{P}_n\left(\hat{f}_n|q\right) - \int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 dx} + O_p\left(n^{-1/2}\right).
\end{aligned}$$

Using again the fact that $g_{\theta_0}(x) \neq \hat{f}_n(x)$, Lemmata 2 and 5 gives

$$\begin{aligned}
&\frac{\mathbb{P}_n g_{\hat{\theta}}(X) - \mathbb{P}_n\left(\hat{f}_n|q\right) - \int\left[g_{\hat{\theta}}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 dx} \\
= &\frac{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right] f(x)\,dx - \int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]\hat{f}_n(x)\,dx}{\int\left[g_{\theta_0}(x) - \hat{f}_n(x)\right]^2 dx} + O_p\left(\sqrt{\zeta_n/n}\right),
\end{aligned}$$

proving the result. ■

To conclude, here is the proof of the last statement in Theorem1:

Proof. By the triangle inequality, we have the following chain of inequalities,

$$\begin{aligned}
&\left\|\hat{\alpha}_n g_{\hat{\theta}} + (1-\hat{\alpha}_n)\hat{f}_n - f\right\|_{2,\lambda} \\
\leq &\left\|\hat{\alpha}_n g_{\theta_0} + (1-\hat{\alpha}_n)\hat{f}_n - f\right\|_{2,\lambda} + \hat{\alpha}_n\left\|g_{\hat{\theta}} - g_{\theta_0}\right\|_{2,\lambda} \\
\leq &\left\|\alpha_n g_{\theta_0} + (1-\alpha_n)\hat{f}_n - f\right\|_{2,\lambda} + |\alpha_n - \hat{\alpha}_n|\left\|g_{\theta_0} - \hat{f}_n\right\|_{2,\lambda} + \hat{\alpha}_n\left\|g_{\hat{\theta}} - g_{\theta_0}\right\|_{2,\lambda} \quad (6)
\end{aligned}$$

and it is enough to bound the last two terms on the r.h.s.. To this end,

$$\begin{aligned}
\left\|g_{\theta_0} - \hat{f}_n\right\|_{2,\lambda} &\leq \left\|f - \hat{f}_n\right\|_{2,\lambda} + \|g_{\theta_0} - f\|_{2,\lambda} \\
&\lesssim 1 + \left\|f - \hat{f}_n\right\|_{2,\lambda},
\end{aligned}$$

as both g_{θ_0} and f are in L_2. Hence, an application of Lemma 6 gives the result, noting that the third term on the r.h.s. of (6) is $O_p\left(n^{-1/2}\right)$ by similar arguments as in Lemmata 2 and 3. ■

References

[1] Aivasian, S.A., V.M. Buchstaber, I.S. Yenyukov, L.D. Meshalkin (1989). Applied Statistics. Classification and Reduction of Dimensionality. Moscow (in Russian).

[2] Andrews, D. (1999) Estimation when a Parameter is on a Boundary. Econometrica 67, 1341-1383.

[3] Arcones, M.A. and E. Giné (1992) On the Bootstrap of U and V Statistics. Annals of Statistics 20, 655-674.

[4] Barron, A.R. (1994) Approximation and Estimation Bounds for Artificial Neural Networks. Machine Learning 14, 113-143.

[5] Breiman, L. (1996) Heuristics of Instability and Stabilization in Model Selection. Annals of Statistics 24, 2350-2383.

[6] Devroye L. and L. Gyorfi (2002) Distribution and Density Estimation. In L. Gyorfi (ed.) Principles of Nonparametric Learning, pp. 211-270, Vienna: Springer-Verlag.

[7] El Ghouch, A. and M. G. Genton (2009) Local Polynomial Quantile Regression With Parametric Features. Journal of the American Statistical Association 104, 1416-1429.

[8] Fan, Y., and A. Ullah (1999) Asymptotic Normality of a Combined Regression Estimator. Journal of Multivariate Analysis 71, 191-240.

[9] Gonzalo, P. and O. Linton (2000) Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression, Journal of Econometrics 99, 63-106.

[10] Hagmann, M. and O. Scaillet (2007) Local Multiplicative Bias Correction for Asymmetric Kernel Density Estimators. Journal of Econometrics 141, 213-249.

[11] Hjort, N.L. and I.K. Glad (1995) Nonparametric Density Estimation with a Parametric Start. The Annals of Statistics 23, 882–904.

[12] Hjort, N.L. and M.C. Jones(1996) Locally Parametric Nonparametric Density Estimation. Annals of Statistics 24, 1619-1647.

[13] Van der Laan, M. and S. Dudoit (2003) Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples. U.C. Berkeley Division of Biostatistics Working Paper 130. http://www.bepress.com/ucbbiostat/paper130.

[14] Ledoit, O. and M. Wolf (2004) A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Journal of Multivariate Analysis 88, 365-411.

[15] Marron, J.S. and W. Härdle (1986) Random Approximations to Some Measures of Accuracy in Nonparametric Curve Estimation. Journal of Multivariate Analysis 20, 91-113.

[16] Mays, J.E., J.B. Birch and B.A. Starnes (2001) Model Robust Regression: Combining Parametric, Nonparametric, and Semiparametric Methods. Journal of Nonparametric Statistics 13, 245-277.

[17] Naito, K. (2004) Semiparametric Density Estimation by Local L_2-Fitting. The Annals of Statistics 32, 1162-1191.

[18] Olkin, I. and C. Spiegelman (1987) A semiparametric Approach to Density Estimation. Journal of the American Statistical Association 82, 858-865.

[19] Sancetta, A. (2008) Sample Covariance Shrinkage for High Dimensional Dependent Data. Journal of Multivariate Analysis 99, 949-967.

[20] Scott, D.W. (1992) Multivariate Density Estimation. Theory, Practice and Visualization. New York: Wiley.

[21] Serfling, R.J. (1980) Approximation Theorems of Mathematical Statistics. New York: Wiley.

[22] Timmermann, A. (2006) Forecast Combinations. In G. Elliott, C.W.J. Granger and A. Timmermann, Handbook of Economic Forecasting. Amsterdam: North-Holland.

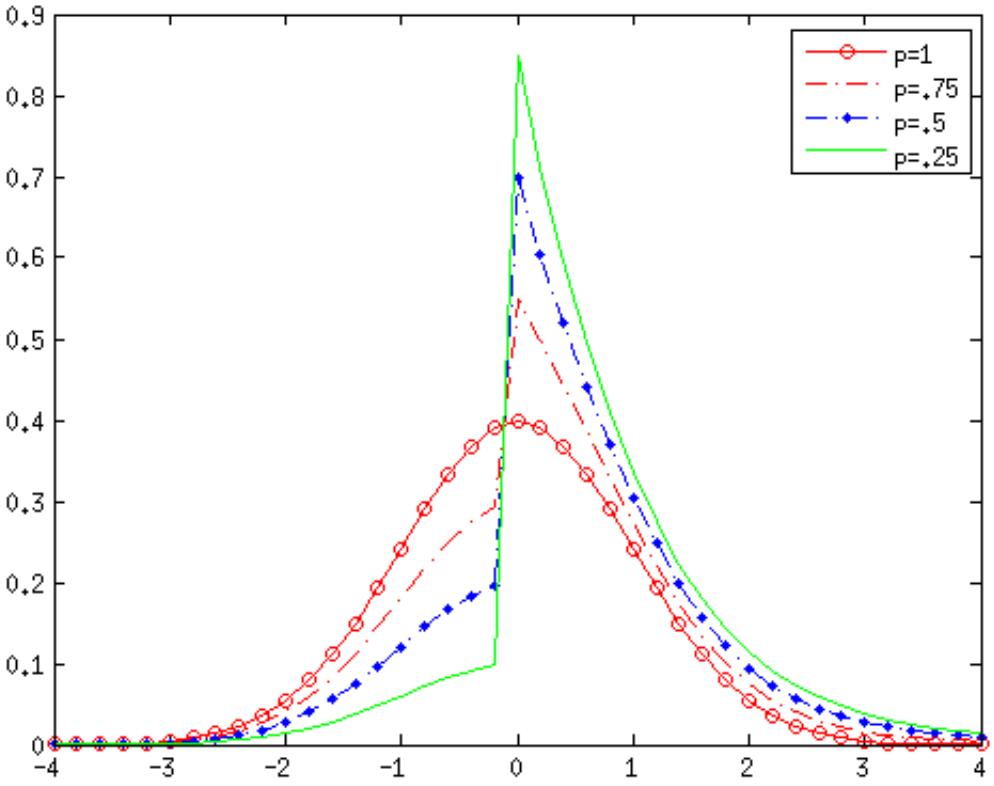

Figure 1: Densities for Different Values of p.

K=1		n=40			
h, p=.25		S	NP	HG	D
0.1	MEAN	6.81 **	7.44 **	7.41 ***	5.60
	SE	0.08	0.08	0.08	0.06
	PRIAL	8.5%	0.0%	0.4%	24.8%
0.3	MEAN	6.69 *	5.79 *	5.64 *	5.79
	SE	0.07	0.05	0.05	0.05
	PRIAL	-15.4%	0.0%	2.6%	0.0%
0.5	MEAN	8.00 ***	7.45 ***	7.14 **	7.45
	SE	0.06	0.04	0.05	0.04
	PRIAL	-7.5%	0.0%	4.1%	0.0%
0.7	MEAN	9.35	9.36	8.97	9.36
	SE	0.05	0.04	0.07	0.04
	PRIAL	0.1%	0.0%	4.3%	0.0%
0.9	MEAN	10.04	11.31	12.75	10.39
	SE	0.06	0.03	0.44	0.07
	PRIAL	11.2%	0.0%	-12.7%	8.1%
h, p=.5					
0.1	MEAN	4.07 ***	6.79	6.79	3.22
	SE	0.05	0.08	0.08	0.04
	PRIAL	40.0%	0.0%	-0.1%	52.6%
0.3	MEAN	3.68 *	3.27 *	3.23 *	3.16
	SE	0.04	0.04	0.04	0.03
	PRIAL	-12.4%	0.0%	1.2%	3.5%
0.5	MEAN	3.98 **	3.69 **	3.47 **	3.69
	SE	0.03	0.03	0.03	0.03
	PRIAL	-7.7%	0.0%	6.1%	0.0%
0.7	MEAN	4.46	4.74 ***	4.21 ***	4.63
	SE	0.03	0.03	0.04	0.03
	PRIAL	5.8%	0.0%	11.1%	2.3%
0.9	MEAN	4.73	6.02	6.04	4.87
	SE	0.04	0.03	0.34	0.05
	PRIAL	21.5%	0.0%	-0.2%	19.2%
h, p=.75					
0.1	MEAN	2.12	6.49	6.52	1.54
	SE	0.05	0.08	0.08	0.02
	PRIAL	67.4%	0.0%	-0.5%	76.3%
0.3	MEAN	1.70	2.14 ***	2.20 ***	1.45
	SE	0.03	0.04	0.03	0.02
	PRIAL	20.5%	0.0%	-2.9%	32.1%
0.5	MEAN	1.55 *	1.65 *	1.60 *	1.47
	SE	0.02	0.03	0.02	0.02
	PRIAL	6.3%	0.0%	3.4%	10.9%
0.7	MEAN	1.57 **	2.00 **	1.69 **	1.57
	SE	0.02	0.03	0.04	0.02
	PRIAL	21.6%	0.0%	15.6%	21.3%
0.9	MEAN	1.59 ***	2.79	2.64	1.61
	SE	0.02	0.02	0.34	0.03
	PRIAL	43.0%	0.0%	5.6%	42.4%

Table 1: Average Integrated Squared Errors, n= 40, K=1. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.

K=2		n=40			
h, p=.25		S	NP	HG	D
0.1	MEAN	8.17 ***	20.58	20.78	7.55
	SE	1.52	3.36	3.39	1.41
	PRIAL	60.3%	0.0%	-1.0%	63.3%
0.3	MEAN	6.80 *	6.25 *	6.31 *	6.25
	SE	0.05	0.04	0.04	0.04
	PRIAL	-8.7%	0.0%	-0.9%	0.0%
0.5	MEAN	7.77 **	7.48 **	7.34 **	7.48
	SE	0.03	0.03	0.03	0.03
	PRIAL	-3.9%	0.0%	1.8%	0.0%
0.7	MEAN	9.02	9.08 ***	9.12 ***	9.07
	SE	0.03	0.02	0.11	0.02
	PRIAL	0.7%	0.0%	-0.4%	0.1%
0.9	MEAN	9.47	10.51	13.51	9.54
	SE	0.03	0.02	0.88	0.04
	PRIAL	9.9%	0.0%	-28.6%	9.2%
h, p=.5					
0.1	MEAN	4.22 ***	20.29	20.48	3.92
	SE	0.03	0.10	0.10	0.02
	PRIAL	79.2%	0.0%	-0.9%	80.7%
0.3	MEAN	3.67 *	3.78 **	3.87 **	3.37
	SE	0.02	0.03	0.03	0.02
	PRIAL	2.9%	0.0%	-2.2%	10.9%
0.5	MEAN	3.90 **	3.75 *	3.61 *	3.74
	SE	0.02	0.02	0.02	0.02
	PRIAL	-4.2%	0.0%	3.6%	0.1%
0.7	MEAN	4.28	4.57 ***	4.27 ***	4.34
	SE	0.02	0.02	0.04	0.02
	PRIAL	6.2%	0.0%	6.4%	4.9%
0.9	MEAN	4.40	5.48	5.93	4.41
	SE	0.02	0.01	0.22	0.02
	PRIAL	19.8%	0.0%	-8.3%	19.5%
h, p=.75					
0.1	MEAN	1.60	19.95	20.17	1.40
	SE	0.02	0.08	0.08	0.01
	PRIAL	92.0%	0.0%	-1.1%	93.0%
0.3	MEAN	1.44	2.56	2.73	1.31
	SE	0.02	0.02	0.02	0.01
	PRIAL	43.7%	0.0%	-6.8%	48.6%
0.5	MEAN	1.37 *	1.61 *	1.61 *	1.33
	SE	0.01	0.02	0.01	0.01
	PRIAL	14.9%	0.0%	-0.3%	17.4%
0.7	MEAN	1.39 **	1.85 **	1.61 **	1.38
	SE	0.01	0.01	0.03	0.01
	PRIAL	24.6%	0.0%	12.5%	25.2%
0.9	MEAN	1.40 ***	2.41 ***	2.29 ***	1.40
	SE	0.01	0.01	0.22	0.01
	PRIAL	42.1%	0.0%	5.1%	42.1%

Table 2: Average Integrated Squared Errors, n= 40, K=2. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.

K=3		n=40			
h, p=.25		S	NP	HG	D
0.1	MEAN	6.81	61.41	62.32	6.46
	SE	0.04	0.77	0.80	0.03
	PRIAL	88.9%	0.0%	-1.5%	89.5%
0.3	MEAN	5.74 *	5.79 **	6.06 **	5.45
	SE	0.02	0.03	0.03	0.02
	PRIAL	0.8%	0.0%	-4.7%	5.9%
0.5	MEAN	5.95 **	5.79 *	5.91 *	5.79
	SE	0.02	0.02	0.03	0.02
	PRIAL	-2.9%	0.0%	-2.1%	0.0%
0.7	MEAN	6.58 ***	6.64 ***	7.34 ***	6.61
	SE	0.01	0.01	0.48	0.01
	PRIAL	0.8%	0.0%	-10.5%	0.5%
0.9	MEAN	6.77	7.37	13.72	6.78
	SE	0.01	0.01	4.59	0.02
	PRIAL	8.1%	0.0%	-86.2%	8.0%
h, p=.5					
0.1	MEAN	3.21	61.26	62.29	3.10
	SE	0.02	0.75	0.78	0.01
	PRIAL	94.8%	0.0%	-1.7%	94.9%
0.3	MEAN	2.88 *	3.70 ***	3.97 ***	2.73
	SE	0.01	0.02	0.02	0.01
	PRIAL	22.2%	0.0%	-7.6%	26.0%
0.5	MEAN	2.91 **	2.87 *	2.94 *	2.83
	SE	0.01	0.01	0.02	0.01
	PRIAL	-1.3%	0.0%	-2.5%	1.3%
0.7	MEAN	3.07 ***	3.25 **	3.30 **	3.08
	SE	0.01	0.01	0.08	0.01
	PRIAL	5.5%	0.0%	-1.6%	5.3%
0.9	MEAN	3.11	3.70	4.48	3.10
	SE	0.01	0.01	0.37	0.01
	PRIAL	16.1%	0.0%	-21.1%	16.1%
h, p=.75					
0.1	MEAN	1.05	62.02	63.11	0.92
	SE	0.02	0.74	0.75	0.00
	PRIAL	98.3%	0.0%	-1.8%	98.5%
0.3	MEAN	0.94	2.56	2.90	0.89
	SE	0.01	0.01	0.01	0.00
	PRIAL	63.4%	0.0%	-13.2%	65.2%
0.5	MEAN	0.91 *	1.14 *	1.27 **	0.89
	SE	0.00	0.01	0.01	0.00
	PRIAL	20.0%	0.0%	-11.8%	21.9%
0.7	MEAN	0.92 **	1.19 **	1.16	0.91
	SE	0.00	0.01	0.02	0.00
	PRIAL	22.7%	0.0%	1.8%	23.2%
0.9	MEAN	0.92 ***	1.45 ***	1.44 ***	0.92
	SE	0.00	0.00	0.07	0.00
	PRIAL	36.8%	0.0%	0.8%	36.9%

Table 3: Average Integrated Squared Errors, n= 40, K=3. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.

K=1		n=80			
h, p=.25		S	NP	HG	D
0.1	MEAN	4.70 *	4.45 *	4.42 *	4.06
	SE	0.05	0.04	0.04	0.04
	PRIAL	-5.6%	0.0%	0.7%	8.9%
0.3	MEAN	5.28 **	5.03 **	4.87 **	5.03
	SE	0.04	0.03	0.03	0.03
	PRIAL	-5.1%	0.0%	3.0%	0.0%
0.5	MEAN	7.21 ***	7.03 ***	6.65 ***	7.03
	SE	0.03	0.03	0.03	0.03
	PRIAL	-2.6%	0.0%	5.4%	0.0%
0.7	MEAN	9.08	9.10	9.30	9.10
	SE	0.03	0.03	0.25	0.03
	PRIAL	0.2%	0.0%	-2.2%	0.0%
0.9	MEAN	9.95	11.13	188.48	10.19
	SE	0.04	0.02	113.40	0.05
	PRIAL	10.5%	0.0%	-1593.7%	8.4%
h, p=.5					
0.1	MEAN	2.95 **	3.69 ***	3.68 ***	2.46
	SE	0.03	0.04	0.04	0.03
	PRIAL	19.9%	0.0%	0.1%	33.3%
0.3	MEAN	2.90 *	2.57 *	2.49 *	2.57
	SE	0.03	0.02	0.02	0.02
	PRIAL	-13.1%	0.0%	3.0%	0.0%
0.5	MEAN	3.56 ***	3.37 **	3.06 **	3.37
	SE	0.03	0.02	0.02	0.02
	PRIAL	-5.5%	0.0%	9.2%	0.0%
0.7	MEAN	4.26	4.54	4.06	4.44
	SE	0.03	0.02	0.08	0.03
	PRIAL	6.1%	0.0%	10.5%	2.2%
0.9	MEAN	4.51	5.88	23.44	4.55
	SE	0.03	0.02	7.65	0.03
	PRIAL	23.4%	0.0%	-298.5%	22.6%
h, p=.75					
0.1	MEAN	1.41	3.25	3.27	1.18
	SE	0.02	0.04	0.04	0.02
	PRIAL	56.7%	0.0%	-0.5%	63.7%
0.3	MEAN	1.22 *	1.30 **	1.30 **	1.11
	SE	0.01	0.02	0.02	0.01
	PRIAL	5.5%	0.0%	-0.3%	14.7%
0.5	MEAN	1.23 **	1.29 *	1.12 *	1.22
	SE	0.01	0.02	0.01	0.01
	PRIAL	4.4%	0.0%	13.3%	5.2%
0.7	MEAN	1.30 ***	1.82 ***	1.45 ***	1.34
	SE	0.01	0.02	0.06	0.02
	PRIAL	28.5%	0.0%	20.1%	26.5%
0.9	MEAN	1.32	2.69	5.22	1.34
	SE	0.02	0.02	1.28	0.02
	PRIAL	51.0%	0.0%	-93.8%	50.4%

Table 4: Average Integrated Squared Errors, n= 80, K=1. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.

K=2		n=80			
h, p=.25		S	NP	HG	D
0.1	MEAN	6.69 ***	10.99	11.08 ***	6.28
	SE	1.22	1.61	1.63	1.10
	PRIAL	39.1%	0.0%	-0.8%	42.8%
0.3	MEAN	5.46 *	5.26 *	5.22 *	5.26
	SE	0.03	0.03	0.03	0.03
	PRIAL	-3.7%	0.0%	0.9%	0.0%
0.5	MEAN	7.19	7.14 ***	6.92 **	7.14
	SE	0.02	0.02	0.03	0.02
	PRIAL	-0.7%	0.0%	3.1%	0.0%
0.7	MEAN	8.89 **	8.93 **	15.44	8.93
	SE	0.02	0.01	4.68	0.01
	PRIAL	0.5%	0.0%	-72.8%	0.0%
0.9	MEAN	9.43	10.43	474.87	9.46
	SE	0.03	0.01	372.03	0.03
	PRIAL	9.6%	0.0%	-4453.2%	9.3%
h, p=.5					
0.1	MEAN	3.60 ***	10.36	10.45	3.38
	SE	0.02	0.04	0.04	0.02
	PRIAL	65.2%	0.0%	-0.9%	67.4%
0.3	MEAN	3.03 *	2.82 *	2.80 *	2.80
	SE	0.02	0.02	0.02	0.02
	PRIAL	-7.5%	0.0%	0.9%	0.8%
0.5	MEAN	3.58 **	3.46 **	3.21 **	3.46
	SE	0.02	0.01	0.02	0.01
	PRIAL	-3.5%	0.0%	7.2%	0.0%
0.7	MEAN	4.16	4.44 ***	5.67 ***	4.24
	SE	0.01	0.01	0.74	0.02
	PRIAL	6.3%	0.0%	-27.6%	4.5%
0.9	MEAN	4.26	5.41	135.46	4.26
	SE	0.02	0.01	75.41	0.02
	PRIAL	21.4%	0.0%	-2402.5%	21.3%
h, p=.75					
0.1	MEAN	1.23	10.03	10.12	1.14
	SE	0.01	0.04	0.04	0.01
	PRIAL	87.7%	0.0%	-0.9%	88.6%
0.3	MEAN	1.13 *	1.51 **	1.55 **	1.06
	SE	0.01	0.01	0.01	0.01
	PRIAL	25.2%	0.0%	-2.9%	29.7%
0.5	MEAN	1.15 **	1.29 *	1.16 *	1.14
	SE	0.01	0.01	0.01	0.01
	PRIAL	10.6%	0.0%	10.3%	11.2%
0.7	MEAN	1.19 ***	1.71 ***	2.00 ***	1.19
	SE	0.01	0.01	0.28	0.01
	PRIAL	30.6%	0.0%	-16.7%	30.5%
0.9	MEAN	1.19	2.35	16.97	1.19
	SE	0.01	0.01	6.06	0.01
	PRIAL	49.2%	0.0%	-622.9%	49.3%

Table 5: Average Integrated Squared Errors, n= 80, K=2. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.

K=3		n=80			
h, p=.25		S	NP	HG	D
0.1	MEAN	6.34 ***	31.13	31.47	6.13
	SE	0.02	0.29	0.30	0.02
	PRIAL	79.6%	0.0%	-1.1%	80.3%
0.3	MEAN	4.98 *	4.76 *	4.84 *	4.76
	SE	0.02	0.02	0.02	0.02
	PRIAL	-4.6%	0.0%	-1.7%	0.0%
0.5	MEAN	5.60 **	5.58 **	5.62 **	5.58
	SE	0.01	0.01	0.04	0.01
	PRIAL	-0.4%	0.0%	-0.7%	0.0%
0.7	MEAN	6.54	6.57 ***	12.15 ***	6.57
	SE	0.01	0.01	2.69	0.01
	PRIAL	0.5%	0.0%	-84.9%	0.0%
0.9	MEAN	6.75	7.34	194.75	6.75
	SE	0.01	0.00	79.04	0.01
	PRIAL	8.1%	0.0%	-2554.1%	8.0%
h, p=.5					
0.1	MEAN	3.02	30.26	30.66	2.90
	SE	0.01	0.28	0.28	0.01
	PRIAL	90.0%	0.0%	-1.3%	90.4%
0.3	MEAN	2.57 *	2.67 **	2.76 **	2.45
	SE	0.01	0.01	0.01	0.01
	PRIAL	3.8%	0.0%	-3.3%	8.2%
0.5	MEAN	2.73 **	2.67 *	2.63 *	2.67
	SE	0.01	0.01	0.01	0.01
	PRIAL	-2.2%	0.0%	1.6%	0.0%
0.7	MEAN	3.01 ***	3.18 ***	4.87 ***	3.03
	SE	0.01	0.01	0.72	0.01
	PRIAL	5.5%	0.0%	-53.1%	4.9%
0.9	MEAN	3.03	3.67	180.73	3.03
	SE	0.01	0.00	112.32	0.01
	PRIAL	17.4%	0.0%	-4826.7%	17.4%
h, p=.75					
0.1	MEAN	0.87	29.96	30.42	0.82
	SE	0.01	0.27	0.28	0.00
	PRIAL	97.1%	0.0%	-1.5%	97.3%
0.3	MEAN	0.81 *	1.48	1.62 ***	0.78
	SE	0.00	0.01	0.01	0.00
	PRIAL	45.6%	0.0%	-9.5%	47.5%
0.5	MEAN	0.81 **	0.93 *	0.94 *	0.80
	SE	0.00	0.00	0.01	0.00
	PRIAL	13.0%	0.0%	-0.4%	13.8%
0.7	MEAN	0.82 ***	1.12 **	1.49 **	0.82
	SE	0.00	0.00	0.20	0.00
	PRIAL	26.5%	0.0%	-32.9%	26.6%
0.9	MEAN	0.83	1.43 ***	17.40	0.82
	SE	0.00	0.00	7.55	0.00
	PRIAL	42.2%	0.0%	-1117.6%	42.4%

Table 6: Average Integrated Squared Errors, n= 80, K=3. * Smallest Loss, ** Second Smallest Loss, *** Third Smallest Loss.