KERNEL DENSITY ESTIMATION AND
MARGINALISATION (IN-) CONSISTENCY

MIKE WEST
Institute of Statistics and Decision Sciences
DP #89-10
Kernel density estimation and marginalisation (in-)consistency

By MIKE WEST

Institute of Statistics and Decision Sciences
Duke University, Durham NC 27706, USA.

SUMMARY

Kernel density estimates, as commonly applied, generally have no exact, model based interpretation since they violate conditions that define coherent joint distributions. The issue of marginalisation consistency is considered here. It is shown that most commonly used kernel functions violate this condition. It is also shown that marginalisation consistency holds only for classes of kernel estimates based on Laplacian, or double-exponential kernels whose window width parameters are appropriately structured.

Some key words: Kernel density estimation; Laplacian kernels; Marginalisation consistency; Predictive distributions
1. GENERAL COMMENTS

Kernel density estimation is one of the most widely used non-parametric techniques in data analysis. Kernel methods have strong intuitive appeal, conceptual simplicity, and current computing standards make them inexpensive to implement. Much recent research has focussed on practical issues of how to tailor kernel techniques to particular datasets through choice of smoothing parameters. Silverman (1986) provides excellent coverage and review of the subject from an applied viewpoint.

The problem considered is that usually referred to as estimation of an uncertain distribution based on the realised values of a random sample drawn from that distribution. From a Bayesian viewpoint, if random quantities $x_{n+1} = \{x_1, \ldots, x_n, x_{n+1}\}$ are judged to be exchangeable, the Bayesian density estimation problem is that of calculating the density $p(x_{n+1}|x_n)$, given the observed values of the $n$ quantities $x_n$. Standard parametric models solve this problem with specific forms for predictive densities. Non-parametric models provide flexibility in modelling distributional forms, but Bayesian non-parametric approaches (Ferguson, 1973, and Leonard, 1978) have to date had very limited impact on statistical practice. Since important reasons for this lack of success relate to the technical and computational difficulties of implementation of these methods, it is somewhat surprising that simpler approaches, such as kernel estimation, have been rather ignored by Bayesians. Recent work in Bayesian analysis using simulation techniques, and in other areas, has led to increased interest in simple and automatic techniques of estimating posterior distributions based on random samples drawn from those posteriors, and so there is reason to consider application of kernel techniques in such context.

A formal Bayesian view of such use would be that ad-hoc density estimates are of interest only to the extent that they can be justified as predictive
densities, or sensible approximations to predictive densities, derived within some formal modelling framework. A more pragmatic view is that if such a framework can be identified, then the techniques can be placed in the context of the assumptions underlying the framework, and their appropriateness in any application judged on the suitability of such assumptions in the practical problem. Additionally, model based justification of ad-hoc techniques may provide insight into problems of generalisation and extension through variations in the basic model structure. For example, the problem of choosing a suitable window-width smoothing parameter in kernel estimation would reduce to a standard problem of parameter estimation if kernel estimates were actually Bayesian predictive densities, for then a likelihood function for the smoothing parameter could be constructed.

Generally, commonly used kernel density estimates violate basic consistency conditions and hence cannot be derived within a model based framework without modification. Here the specific issue of marginalisation consistency is considered. It is shown that, whilst most commonly used kernel functions violate this condition, there exists a class of kernel functions for which the condition is satisfied, so long as the sequence of window width parameters is appropriately structured. It is shown that classes of kernel estimates based on Laplacian, or double-exponential kernels, uniquely satisfy the marginalisation consistency condition, but only if the window width sequence is inversely proportional to the square root of the sample size. Questions relating to whether approaches using this, or any other, form provide useful approximations to Bayesian predictive distributions, involving study of the implications of violation of other consistency conditions, remain open.
2. MARGINALISATION CONSISTENCY

Let $x_{n+1} = \{x_1, \ldots, x_{n+1}\}$ be exchangeably distributed random quantities whose joint, marginal and conditional densities are denoted by $p(\cdot \mid \cdot)$, the relevant density being made explicit through the arguments and conditioning quantities. The collections of such densities satisfies the usual rules of probability calculus. In particular, any direct specification of conditional densities $p(x_i | x), (i = 1, \ldots, n+1,$ and $x \subseteq \{x_{n+1} - x_i\}$, must cohere. Kernel density estimates are directly specified densities of the form

$$f_k(x | x_1, \ldots, x_k) = \frac{1}{k\sigma_k} \sum_{i=1}^{k} f((x - x_i) / \sigma_k)$$

(1)

where $f(.)$ is a chosen kernel, a symmetric density function on the real line, and, for each $k$, $\sigma_k$ is a positive scale parameter, known as the window width, that acts as a smoothing parameter. One commonly used kernel is the standard normal density; for others, see Silverman (1986), for example. If kernel density estimates are to cohere, then they must satisfy basic laws such as Bayes' theorem, rules for marginalisation, and so forth. It is clear that Bayes' theorem defeats kernel density estimates. For example, it is not generally possible to find a kernel function and smoothing parameter sequence so that, for a given $n$,

$$f_n(x_n | x_{n+1}, x_{n-1}) \propto f_{n-1}(x_n | x_{n-1}) f_n(x_{n+1} | x_n)$$

(2)

holds for all values of $x_{n+1}$.

Despite this, it is still of interest to consider the use of kernel estimates (for the reasons stated earlier) if they may be shown to approximately cohere. Further consistency conditions are also important, the key being that of marginalisation consistency. In particular, a joint density must satisfy the marginalisation conditions

$$p(x_{n+1} | x_{n-1}) = \int_{-\infty}^{\infty} p(x_{n+1} | x_n) p(x_n | x_{n-1}) dx_n,$$

(3)
for all \( n \) and all values of the quantities \( x_{n+1} \). Under kernel density estimation, the density of the left hand side of this equation is given by
\[ p(x_{n+1}|x_n) = f_{n-1}(x_n|x_{n-1}) \]
whilst the expression on the right hand side is
\[ \int_{-\infty}^{\infty} f_{n}(x_{n+1}|x_n) f_{n-1}(x_n|x_{n-1})dx_n \]. The following result shows that commonly used kernel approaches violate the equality in (3), whilst identifying a particular class of kernel estimates for which (3) does hold.

**Theorem.** Marginalisation consistency is satisfied if, and only if, the kernel function is Laplacian, or double-exponential, with
\[ f(x) = \frac{1}{2}e^{-|x|}, \quad (-\infty < x < \infty), \]
and if the window width \( \sigma_n \) is defined, for each \( n \), by
\[ \sigma_n^2 = b/n \]
where \( b \) is some positive constant.

**Proof.** In (3) write \( g_n(u) = f(u/\sigma_n)/\sigma_n \) for all \( u \) and \( n \). Substituting the relevant kernel density estimates (1) into the identity (3) gives, for \( n > 1 \),
\[
(n - 1)^{-1} \sum_{j=1}^{n-1} g_{n-1}(x_{n+1} - x_j) = \\
\int_{-\infty}^{\infty} \left\{ \sum_{j=1}^{n} g_n(x_{n+1} - x_j) \right\} \left\{ (n - 1)^{-1} \sum_{j=1}^{n-1} g_{n-1}(x_n - x_j) \right\} dx_n.
\]
Now, only the final component of the first bracketed sum in the integrand depends on \( x_n \). As a result, the expression simplifies to
\[
(n - 1)^{-1} \sum_{j=1}^{n-1} g_{n-1}(x_{n+1} - x_j) = \\
\left\{ (n - 1) \sum_{j=1}^{n-1} g_n(x_{n+1} - x_j) \right\} + \int_{-\infty}^{\infty} g_n(x_{n+1} - x) \left\{ \sum_{j=1}^{n-1} g_{n-1}(x_n - x_j) \right\} dx_n.
\]
so that
\[ \sum_{j=1}^{n-1} \{ n g_{n-1}(x_{n+1} - x_j) - (n - 1) g_n(x_{n+1} - x_j) \} \]
\[ \int_{-\infty}^{\infty} g_n(x_{n+1} - x_n) g_{n-1}(x_n - x_j) dx_n \] = 0.

To satisfy marginalisation consistency, this identity must hold for all possible values of the elements of \( x_{n+1} \). By letting subsets of these elements tend to infinity, so that the corresponding density values tend to zero, it follows that the identity holds if, and only if, the summands are each equal to zero for all values of their particular arguments. Thus, for all \( n \) and \( j \neq n \),
\[ n g_{n-1}(x_{n+1} - x_j) = (n - 1) g_n(x_{n+1} - x_j) + \]
\[ \int_{-\infty}^{\infty} g_n(x_{n+1} - x_n) g_{n-1}(x_n - x_j) dx_n. \]

With no loss of generality, set \( x = x_{n+1} - x_j \) and \( y = x_n - x_j \). This expression now reads
\[ n g_{n-1}(x) = (n - 1) g_n(x) + \int_{-\infty}^{\infty} g_n(x - y) g_{n-1}(y) dy \]  \hspace{1cm} (4)
for all real \( x \).

To explore solutions to this equation, the convolution term on the right suggests the use of characteristic functions. For each \( n \), the characteristic function of the density \( g_n(.) \), \( \gamma_n(t) = \int_{-\infty}^{\infty} e^{itx} g_n(x) dx \), exists for all real \( t \). Then the convolution appearing in (4) is a density for \( x \) which has characteristic function \( \gamma_n(t) \gamma_{n-1}(t) \). Hence, from (4), we have
\[ n \gamma_{n-1}(t) = (n - 1) \gamma_n(t) + \gamma_n(t) \gamma_{n-1}(t). \]  \hspace{1cm} (5)

This equation is more easily handled than the equivalent expression (4) in terms of densities. Since \( \sigma_n \) is a scale parameter for the density \( g_n(.) \), we
can write $\gamma_n(t) = \gamma(\sigma_n t)$, for all $n$, where $\gamma(t)$ is the characteristic function of the standard kernel density $f(\cdot)$. Now (5) may be rewritten as

$$n\gamma(\sigma_n t)^{-1} = 1 + (n - 1)\gamma(\sigma_{n-1} t)^{-1}.$$  

Suppose now that $\gamma(t)^{-1}$ has a Taylor series expansion about $t = 0$ of the form $\gamma(t)^{-1} = 1 + a_1 t^2 + a_2 t^4 + \ldots$, the constant term being unity by virtue of the fact that $\gamma(0) = 1$. Note that the odd powers of $t$ do not appear since the kernel density $f(\cdot)$ is symmetric about zero, so that $\gamma(t)$ and hence $\gamma(t)^{-1}$ are functions only of $t^2$. Then, matching coefficients of $t^2$, $t^4$, and so forth, we are led to the identities

$$n a_k \sigma_n^{2k} = (n - 1) a_k \sigma_{n-1}^{2k},$$

for each $k = 1, 2, \ldots$, from which it follows that, for all such $k$,

$$a_k \sigma_n^{2k} = n^{-1} a_k \sigma_1^{2k} = n^{-1} a_k b^k,$$  

(6)

where $b = \sigma_1^2 > 0$. Clearly, these identities can hold simultaneously only under the circumstances that all but one of the coefficients $a_k$ are zero. So, for a particular positive integer $p$, $a_p > 0$, all the other coefficients being zero. Then equation (6) at $k = p$ implies

$$\sigma_n^2 = b n^{-1/p}.$$  

(7)

For the kernel function, note that the characteristic function of $f(\cdot)$ is now given by

$$\gamma(t) = \frac{1}{(1 + a_p t^2)^p}$$

for some constant $a_p$.

Hence, amongst classes of densities of the form (1) with symmetric kernels $f(\cdot)$, marginalisation consistency (3) is satisfied only under (7) and (8).
The question now arises as to when (8) defines a valid characteristic function. It is easily shown that this is true only for \( p = 1 \). To see this, recall that, if the second derivative of \( \gamma(.) \) vanishes at zero, then \( \gamma(.) \) cannot be a characteristic function since the corresponding distribution would have a vanishing second moment (Feller, 1966, page 486.) For \( \gamma(.) \) given by (7) for some non-negative integer \( p \), it can be verified that

\[
\gamma''(t) = 2pa_p \gamma(t)^3 t^{2(p-1)} \{ a_p (1 + 2p) t^{2p} + 1 - 2p \}.
\]

If \( p > 1 \), \( \gamma''(0) = 0 \) so that \( \gamma(.) \) is not a characteristic function. If \( p = 1 \),
\( \gamma(t) = (1 + a_1 t^2)^{-1} \). Since \( \sigma_n \) is the scale parameter of the kernel, \( a_1 = 1 \) and \( \gamma(t) = (1 + t^2)^{-1} \) is just the characteristic function the double-exponential density

\[
f(x) = \frac{1}{2} e^{-|x|}, \quad (-\infty < x < \infty).
\]

This is therefore the only kernel function that satisfies the marginalisation consistency condition. In addition, at \( p = 1 \) equation (7) implies that \( \sigma_n^2 = b/n \), and the result follows.

\( \diamond \)

As a corollary to this result note that marginalisation consistency is violated by most commonly used kernels. This is true, in particular, for the popular normal, biweight, rectangular, Epanechnikov and triangular kernels (Silverman, 1986, p43.)
ACKNOWLEDGEMENT

The author is grateful to Robert Wolpert, of ISDS, for discussion and suggestions on an earlier version of this paper.

REFERENCES


