Miscellanea

Kernel density estimation and marginalization consistency

BY MIKE WEST

Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27706, U.S.A.

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 marginalization consistency is considered here. It is shown that most commonly used kernel functions violate this condition. It is also shown that marginalization consistency holds only for classes of kernel estimates based on Laplacian, or double-exponential kernels whose window width parameters are appropriately structured. The practical relevance and implications of this result are discussed.

Some key words: Kernel density estimation; Laplacian kernel; Marginalization consistency; Predictive distribution.

1. Introduction

Kernel density estimation is one of the most widely used nonparametric 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 focused on practical issues of how to tailor kernel techniques to particular data sets 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 realized values of a random sample \( x^{(n)} = \{x_1, \ldots, x_n\} \) drawn from that distribution. An alternative formulation, that does not involve consideration of an underlying, true distribution, assumes the random quantities \( x^{(n+1)} = \{x^{(n)}, x_{n+1}\} \) to be exchangeable, and poses the problem as one of estimating, or approximating, the conditional density \( p(x_{n+1}|x^{(n)}) \), given the observed values of the \( n \) quantities \( x^{(n)} \). This is a natural predictive formulation, with obvious Bayesian overtones.

Although kernel density estimates are often used in a single analysis, with a fixed sample size and data set, there are good reasons to ask whether they can be justified as densities, or sensible approximations to densities, that are derived within some formal modelling framework. Foundational issues aside, a pragmatic view is that, if such a framework can be identified, the resulting techniques can be placed in the context of the associated modelling assumptions. Preliminary judgements about the appropriateness of the techniques in any application can then be based on the suitability of such assumptions for the particular problem. Model based justification may also provide insight into problems of generalization 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 in a parametric framework, since a likelihood function for the smoothing parameter could then be constructed.

It is the case, however, that kernel density estimates are usually used without reference to these foundational considerations. The issues that arise with this, and other nonparametric constructions, have to do with consistency, or coherence. Marginalization consistency, the focus of this paper,
provides an illustration. Suppose an investigator views the \( x^{(n+1)} \) as exchangeable, and uses a specified rule, such as a kernel technique, to derive densities for inference conditional on observing some of the \( x^{(n+1)} \). Unless these densities satisfy the conditions necessary to define a coherent joint distribution for the \( x^{(n+1)} \), the following type of problem arises. Consider inference about \( x_{n+1} \) conditional on \( x^{(n-1)} \). The rule applied directly, conditional on the \( n - 1 \) quantities observed, leads to some density estimate \( f_{n-1}(x_{n+1}|x^{(n-1)}) \). However, were \( x_n \) observed, the same rule delivers a density \( f_n(x_{n+1}|x^{(n)}) \), and since, under the exchangeability assumption, the rule also provides \( f_{n-1}(x_n|x^{(n-1)}) \) for \( x_n \), then the goal of inference may alternatively be calculated via marginalization with respect to \( x_n \). The question then arises as to whether the two routes lead to the same result, as they must in any formal model based framework.

This issue is studied below. It is shown that commonly-used kernel density estimates violate this condition of marginalization consistency. It is further shown that classes of kernel estimates based on Laplacian, or double-exponential kernels, uniquely satisfy the marginalization consistency condition, but only if the window width sequence is inversely proportional to the square root of the sample size. Some discussion of this result appears in § 3.

2. Marginalization 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 \( p(.,.|.) \), the relevant density being made explicit through the arguments and conditioning quantities. The collection of such densities satisfies the usual rules of probability calculus. In particular, any direct specification of conditional densities \( p(x_i|x) \), for \( i = 1, \ldots, n + 1 \), and \( x \subseteq \{x^{(n+1)} - x_i\} \), must cohere. Kernel density estimates are directly specified densities of the form

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

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, 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 the laws of probability, at least approximately. In particular, a joint density must satisfy the marginalization conditions

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

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 \( p(x_{n+1}|x^{(n-1)}) = f_{n-1}(x_n|x^{(n-1)}) \), whilst the expression on the right-hand side is

\[
\int f_n(x_{n+1}|x^{(n)}) f_{n-1}(x_n|x^{(n-1)}) \, dx_n.
\]

The following result shows that commonly-used kernel approaches violate the equality in (2), whilst identifying a particular class of kernel estimates for which (2) does holds.

**Theorem.** Marginalization 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 (2) 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 (2) gives, for \( n > 1 \),

\[
(n-1)^{-1} \sum_{j=1}^{n} g_{n-1}(x_{n+1} - x_j) = \int_{-\infty}^{\infty} \left\{ n^{-1} \sum_{j=1}^{n} g_n(x_{n+1} - x_j) \right\} \left\{ (n-1)^{-1} \sum_{j=1}^{n-1} g_{n-1}(x_n - x_j) \right\} \, dx_n.
\]

Now only the final component of the first bracketed sum in the integrand depends on \( x_n \). As a result, the expression simplifies to

\[
n \sum_{j=1}^{n-1} g_{n-1}(x_{n+1} - x_j) = \left\{ (n-1)^{-1} \sum_{j=1}^{n} g_n(x_{n+1} - x_j) \right\} + \int_{-\infty}^{\infty} g_n(x_{n+1} - x_n) \left\{ \sum_{j=1}^{n-1} g_{n-1}(x_n - x_j) \right\} \, dx_n,
\]

so that

\[
\sum_{j=1}^{n-1} \left\{ ng_{n-1}(x_{n+1} - x_j) - (n-1)g_n(x_{n+1} - x_j) - \int_{-\infty}^{\infty} g_n(x_{n+1} - x_n) g_{n-1}(x_n - x_j) \, dx_n \right\} = 0.
\]

To satisfy marginalization 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 \( j \neq n \),

\[
ng_{n-1}(x_{n+1} - x_j) = (n-1)g_n(x_{n+1} - x_j) + \int_{-\infty}^{\infty} g_n(x_{n+1} - x_n) g_{n-1}(x_n - x_j) \, dx_n.
\]

With no loss of generality, set \( x = x_{n+1} - x_j \) and \( y = x_n - x_j \). This expression now reads

\[
n g_{n-1}(x) = (n-1)g_n(x) + \int_{-\infty}^{\infty} g_n(x-y) g_{n-1}(y) \, dy \tag{3}
\]

for all real \( x \).

To explore solutions to this equation, we use the equivalent expression (3) in terms of densities. For each \( n \), the characteristic function of the density \( g_n(.) \), \( \gamma_n(t) = \int e^{itx} g_n(x) \, dx \), exists for all real \( t \). Then the convolution appearing in (3) is a density for \( x \) which has characteristic function \( \gamma_n(t) \gamma_{n-1}(t) \). Hence, from (3), we have

\[
n \gamma_{n-1}(t) = (n-1) \gamma_n(t) + \gamma_n(t) \gamma_{n-1}(t). \tag{4}
\]

This equation is more easily handled than the equivalent expression (3) 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 \( f(.) \). Now (4) 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 \( f(.) \) 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 \( k \),

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

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 (5) at \( k = p \) implies

\[
\sigma_n^2 = b n^{-1/p}.
\]
For the kernel function, note that the characteristic function of \( f(.) \) is now given by
\[
\gamma(t) = \frac{1}{(1 + a_p t^{2p})}
\] (7)
for some constant \( a_p \).

Hence, amongst classes of densities of the form (1) with symmetric kernels \( f(.) \), marginalization consistency (2) is satisfied only under (6) and (7). The question now arises as to when (7) 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, p. 486.) For \( \gamma(.) \) given by (6) for some nonnegative integer \( p \), it can be verified that
\[
\gamma''(t) = 2p a_p \gamma(t)^2 (p-1)(a_p (1+2p)t^{2p} + 1 - 2p).
\]
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 marginalization consistency condition. In addition, at \( p=1 \) eqn (6) implies that \( \sigma_n^2 = b/n \), and the result follows.

3. Discussion

As a corollary to this result note that marginalization 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, p. 43.) The extent to which other such kernel estimates may approximately cohere in this sense is unclear, apart from the obvious, but practically unimportant, comment that all is well asymptotically as \( n \) tends to infinity. The Laplacian kernel is interesting in that it is rather peaked relative to more standard kernels, and also heavier-tailed. Note also that the window width inversely proportional to the square root of the sample size implies a faster decay rate than traditional recommendations.

Marginalization is only one relevant consistency condition. Having narrowed the class of kernel estimates to those with Laplacian kernels, further conditions can be checked. Thus, for example, it is of interest to ask whether such kernel estimates satisfy Bayes’s theorem, namely
\[
f_n(x_n|x_{n+1}, x^{(n-1)}) \propto f_{n-1}(x_n|x^{(n-1)})f_n(x_{n+1}|x^{(n)})
\]
for all values of \( x^{(n+1)} \), for a given \( n \). Unfortunately, restricting attention to marginalization consistent densities based on Laplacian kernels with the appropriate window width sequence, straightforward, though tedious, calculations show that this cannot be true. Thus Bayes’s theorem defeats kernel density estimates. Despite this, it is still of interest to consider variations on kernel estimates if they may be shown to approximately cohere. The extent and nature of the violation of consistency conditions such as Bayes’s theorem, is unclear. Further, it may be that constructive modifications to the basic kernel technique could provide coherence, or approximate coherence, though whether simple and direct modifications to ensure consistency can easily be found remains an open question.

From a formal Bayesian viewpoint, techniques such as kernel estimation 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. More pragmatically, there are good reasons to consider kernel techniques and to explore their rationale. One reason is their wide application and current popularity. Also, 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.
Kernel techniques are being used in this area. Furthermore, it is fair to say that existing Bayesian approaches to density estimation outside standard parametric models (Ferguson, 1973, 1983; 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. What is clear is that some such modelling approaches naturally lead to kernel-like approximations to predictive distributions. Some possibilities include well-known approximate maximum likelihood estimates in mixture models (Laird, 1978; Lindsay, 1983), and, particularly, justifications using mixtures of Dirichlet processes, as by Ferguson (1983). Modelling developments in these areas might be expected to lead to increased interest in underlying rationales for kernel techniques, amongst Bayesians and non-Bayesians, and may shed light on some of the practical issues currently associated with their application.

Acknowledgements

The author acknowledges the comments of the referees, the editor, and Robert Wolpert, of ISDS, on an earlier version of this paper. The research reported here was partially financed by the National Science Foundation.

References


[Received October 1989. Revised November 1990]