Statistics and Probability Letters 83 (2013) 916–922

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

# Entropic kernels for data smoothing

Roger Bowden *

*Kiwicap Research Ltd, 11 Waiteata Road, Kelburn, Wellington 6012, New Zealand*

ARTICLE INFO

*Article history:*
Received 9 May 2012
Received in revised form 4 December 2012
Accepted 4 December 2012
Available online 10 December 2012

*MSC:*
62J02
62F05
94A17

*Keywords:*
Bandwidth
Data smoothing
Kernel regression
Locational entropy
Subjective probabilities

ABSTRACT

Data smoothing or regression kernels based on locational entropy embody the principle that observations towards the extremes of the chosen data window should provide less information than those at the midpoint. Weight patterns can be flexible, depending on the choice of prior information density.



## 1. Introduction

Data smoothing using local windows is commonly executed with kernel functions that attach diminishing weight to observations indexed further from the current data point, taken as the window centre (Savitzky and Golay, 1964). A related context is non parametric regression, where the index set consists of sequential values of an independent variable and the object is to estimate the conditional expectation of a dependent variable (Jennen-Steinmetz and Gasser, 1988). The Nadaraya–Watson (Nadaraya, 1964; Watson, 1964), Priestley–Chao (Priestley and Chao, 1972) and Gasser–Muller (Gasser and Müller, 1984; Gasser et al., 1984) regression kernels are all instances. Li and Racine (2007) is an extended discussion in an econometric context, also Richard et al. (2009) for time series prediction. A variety of kernel specifications are in use, with profiles often based on common distribution functions, including the uniform as a kernel representation of a fixed window unweighted moving average. The Epanechnikov kernel (Epanechnikov, 1969) is widely cited as an efficiency standard, as it minimises the asymptotic mean integrated square error (AMISE), where the data are independently drawn from a common underlying probability distribution (e.g. Wand and Jones, 1995).

Smoothing kernels in common use vary widely in their weighting profiles (Simonoff, 1996). The Epanechnikov kernel is concave, while continuous kernels originating from continuous distributions, such as the Gaussian, have mixed concave–convex profiles and points of inflexion. Optimality properties are commonly established in terms of loss functions adapted to specific data generation structures. Thus the Epanechnikov kernel is AMISE in contexts with additive i.i.d. disturbances. In data smoothing, the disturbance properties are commonly unknown, and the objective may simply be to aid pattern comprehension ('eyeballing'), much as in the lower order details of wavelet analysis.

* Tel.: +64 4 472 4984; fax: +64 4 472 4983.
*E-mail address:* roger.bowden@kiwicap.co.nz.

0167-7152/$ – see front matter 
doi:10.1016/j.spl.2012.12.006

The present note suggests a different vantage point for the derivation of fixed window weights, originating in information theory. The resulting kernel is strictly concave with a general resemblance to the Epanechnikov kernel, though less severe on weight censoring away from the centre. For expositional convenience, the context adopted is data smoothing based on discrete time data, but with some qualifications, the same framework can also apply to kernel regression (Section 4).

An informational approach to data smoothing can be subjectively rationalised in the following terms. In many data smoothing contexts, very little prior information is available about the form of any underlying signal. Making an observation $x_t$ at time $t$ is resolving some, but not all, of the intrinsic uncertainty about its nature or reality. But informational collapse is not complete because residual doubts remain—the observation may be unrepresentative. An observation on neighbouring time points $t \pm \tau$ should also contribute to the collapse of entropy at $t$, but progressively less so towards the edges of any chosen window, suggesting that the weight profile should be uniformly concave. With such considerations in mind, the choice of an information based weighting system is based on the following criteria:

(a) The smoothing weights should be preassigned and invariant. In particular, they should not be affected at any point by the realised value of the current (dependent variable) observation, though the parameter choice can legitimately depend upon the degree of random variability of the series as a whole.
(b) A rationalisation should exist in terms of a formal informational framework, such that observations closer to the smoothing centre should be recognised as collapsing uncertainty to a greater extent than more distant. The weighting scheme should assign progressively less weight to neighbouring observations further out from the centre of the bandwidth. Moreover, the weight pattern within any given window should diminish at a greater rate, i.e. the weight pattern is concave.
(c) The weight formula should be flexible, able to handle multiple bandwidth lengths in a simple one-parameter framework.

The suggested information theory framework is that of locational entropy (Bowden, 2012). Locational entropy is always higher around the median or midpoint, corresponding here to the current data point, but diminishes away from the centre. The approach requires the user to nominate a starting distribution, analogous to a prior in Bayesian statistics, though the end result is the locational entropy weighting system, in place of the Bayesian posterior. The techniques are illustrated with a uniform distribution, corresponding to the Bayesian uninformative prior, defined by the width of the window as a single parameter.

The scheme of development is as follows. The entropic weighting procedure is described on a general level in Section 2. Section 3 implements the procedure in terms of a maximum entropy discrete time kernel. Section 4 extends to the metric window approach commonly used in regression. An illustrative application is the returns on onshore versus offshore Chinese RMB bonds.

## 2. The generic entropy based kernel

As earlier noted, the proposed kernel is to reflect the contribution of neighbouring time points to the collapse of uncertainty concerning the current observation, as the centre of the kernel. The uncertainty measure adapted for the purpose is locational (or partition) entropy (Bowden, 2012) with a suitably chosen probability density as the point of departure. The kernel weights are then obtained by normalising the locational entropy ordinates to add to unity within the chosen window. The development in this section outlines the generic procedure.

Given a random variable $\upsilon$ with distribution function $F$, the locational entropy function $h$ at any given point $\tau$ has value:

$$h_\tau = -[F(\tau)\ln F(\tau) + (1 - F(\tau))\ln(1 - F(\tau))]. \tag{1}$$

For a given value $\tau$, expression (1) measures the Shannon entropy[1] associated with a dichotomous random variable that takes value unity if $\upsilon > \tau$, and zero otherwise; its entropy can be regarded as a measure of the uncertainty of position attached to the chosen point $\tau$. The nonnegative function $h$ has a maximum of $\ln 2$ at the median of $F$, and values of zero at the end points of the range. Its expected value with respect to the distribution $F$ is $1/2$. The locational entropy function $h$ does not necessarily mirror the density $f$ associated with $F$. Thus for long tailed densities such as the logistic, the tails are overemphasised, which have a greater amount of entropy. A uniform distribution for $F$ results in a concave $h$ and hence a weighting function with an internal maximum.

In what follows, the support of the locational entropy function $h$ is taken as discrete to correspond to discrete data. On the other hand, the parent $F$ can have either discrete or continuous support, even the infinite real line. All that is necessary in either case is to renormalise to ensure that $\sum_\tau h_\tau = 1$. Starting with a continuous parent distribution has advantages in a greater menu of choice of kernel function shapes for discrete time kernel windows. Likewise, a continuous time uniform distribution as a choice for $F$ is the Shannon entropy maximising distribution, but this is not automatically true in discrete time.

In the present implementation, $F$ will be taken as a uniform distribution, whose midpoint (median) and range correspond to the chosen window. This is equivalent to saying that each point inside the window starts with an uninformative weight, until the user forms an attitude as to what the weight should be. In forming ideas about relative uncertainty to be assigned, the user acts as though he or she adopts locational entropy as the basis. To this extent, the chosen approach reflects criteria (a)–(c) of Section 1. However, similar techniques can be utilised starting with any alternative distribution function.

[1] For example $\sum_i f(\tau_i)\ln f(\tau_i)$ or $\int f(\tau)\ln f(\tau)d\tau$.