Adaptive kernel density estimation*

Philippe Van Kerm

CEPS/INSTEAD
G.-D. Luxembourg

9th UK Stata Users meeting, Royal Statistical Society, London
May 19-20, 2003

Note: This manual is forthcoming in a Stata Journal paper which provides practical examples and illustrations.

Abstract

This paper describes the Stata module `akdensity`. `akdensity` extends the official Stata command `kdensity` that estimates density functions by the kernel method. The extensions are of two types. Firstly, `akdensity` allows the use of an 'adaptive kernel' approach with varying, rather than fixed, bandwidths. Secondly, `akdensity` estimates pointwise variability bands around the estimated density functions.

Keywords: adaptive kernel density, local bandwidths, variability bands

1 Overview

Stata offers one official command for non-parametric estimation of density functions: `kdensity`; see [R] **kdensity**. Important user-written extensions have also been developed in Salgado-Ugarte et al. (1993), Salgado-Ugarte et al. (1995) and Salgado-Ugarte & Pérez-Hernández (2003) for bandwidth selection and estimation with adaptive kernel functions. The present insert describes `akdensity`, a module that further extends the possibilities offered for kernel density estimation in Stata. Extensions are of two types. Firstly, `akdensity` allows the use of a varying, rather than fixed, bandwidth as in Salgado-Ugarte et al. (1993) and Salgado-Ugarte & Pérez-Hernández (2003). The main improvement over existing modules in this regard is in computation speed. The algorithm implemented permits a much faster estimation when dealing with large datasets. `akdensity` is also more flexible in that it allows weights, user-defined grid points, and both Gaussian and Epanechnikov kernel functions. Secondly, `akdensity` provides estimation of pointwise variability bands. The new command is compatible with both Stata 7 and Stata 8, using the appropriate graphics engine under both versions.

2 Adaptive kernel density estimation and variability bands

Usefulness of varying (or local) bandwidths is widely acknowledged to estimate long-tailed or multi-modal density functions with kernel methods, when a fixed (or global) bandwidth approach may result in undersmoothing in areas with only sparse observations while oversmoothing in others. Varying the bandwidth along the support of the sample data gives flexibility to reduce the variance of the estimates in areas with few observations, and reducing the bias of the estimates in areas with many observations. Kernel density estimation methods relying on such varying bandwidths are generally referred to as 'adaptive kernel' density estimation methods. For an introductory exposition of such methods, see, e.g., Silverman (1986), Bowman & Azzalini (1997), or Pagan & Ullah (1999). Salgado-Ugarte et al. (1993), Salgado-Ugarte et al. (1995)

*This paper is a by-product of joint work with Stephen Jenkins whose comments and suggestions are gratefully acknowledged. Financial support was received from the European Commission under the Transnational Access to major Research Infrastructures contract HPRI-CT-2001-00128 hosted by IRISS-C/I at CEPS/INSTEAD Differdange (Luxembourg). (This document is CEPS/INSTEAD internal doc. 07-03-0001.) Contact: `philippe.vankerm@ceps.lu`.

and Salgado-Ugarte & Pérez-Hernández (2003) provide discussions in the context of Stata, addressing both fixed and varying bandwidth methods.

An adaptive kernel approach adapts to the sparseness of the data by using a broader kernel over observations located in regions of low density. This is done by varying the bandwidth inversely with the density. As Silverman (1986, p.101) puts it, "An obvious practical problem is deciding in the first place whether or not an observation *is* in a region of low density." Adaptive kernel density estimation deals with this question by using an iterative procedure: An initial (fixed bandwidth) density estimate is computed to get an idea of the density at each of the data points, and this pilot estimate is used to adapt the size of the bandwidth over the data points when computing a new kernel density estimate.

The second feature of `akdensity` is the possibility to request the estimation of pointwise variability bands around the estimated density functions. These bands are constructed as the estimated density at a given grid point x, $\hat{f}(x)$, plus or minus b times the estimated standard error of $\hat{f}(x)$. Note that one should not interpret the bands as providing (pointwise) confidence intervals for $f(x)$ (setting, for example, b at 1.96 to obtain a 95% confidence interval). Kernel density estimates are asymtotically biased, with a bias varying with the bandwidth and the shape of the underlying 'true' density function. For a given bandwidth, the bias does not tend to 0 as the sample size increases. Use of the words 'variability bands', rather than 'confidence bands', is meant to emphasise that the bands quantify the variability of the density estimate but do not take the bias of the estimate into account, and thence do not provide a means of examining particular hypotheses about the density function (Bowman & Azzalini 1997, pp.29–30).

3 Methods and formulae

The method implemented in `akdensity` is the now standard *adaptive two-stage estimator* proposed in Abramson (1982). It is based on the construction of a *local bandwidth factor*, λ_i, at each sample point. The local bandwidth factors have unit (geometric) mean and multiply a global fixed bandwidth, h. Thence h controls the overall degree of smoothing while the λ_i stretch or shrink the sample points bandwidths to adapt to the density of the data.

The adaptive kernel density estimate is given by

$$\hat{f}(x) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} \frac{w_i}{h_i} K\left(\frac{x - x_i}{h_i}\right) \qquad (1)$$

where the x_i's are the data points (associated with weights w_i), K is a kernel function, and $h_i = h \times \lambda_i$. (Compare with [R] **kdensity**.)

The local bandwidth factors are proportional to the square root of the underlying density functions at the sample points:

$$\lambda_i = \lambda(x_i) = \left(G/\tilde{f}(x_i)\right)^{0.5} \qquad (2)$$

where G is the geometric mean over all i of the pilot density estimate $\tilde{f}(x)$. The pilot density estimate is a standard fixed bandwidth kernel density estimate obtained with h as bandwidth.[1]

The variability bands are based on the following expression for the variance of $\hat{f}(x)$ given in Burkhauser et al. (1999):

$$V\left(\hat{f}(x)\right) = \left(\sum_{i=1}^{n} \frac{w_i^2}{n^2}\right) \frac{f(x)}{h\lambda(x)} \int (K(s))^2\, ds \qquad (3)$$

The b parameter that controls the number of standard errors to add around $\hat{f}(x)$ to construct the variability bands is specified by the user.

4 Implementation notes

`akdensity` is packaged in two modules. The engine of the package is `akdensity0`. It allows kernel density estimation with either fixed or observation-specific bandwidths (i.e. the bandwidth parameter can be either a scalar or a variable name), and optionally generates local bandwidth factors after estimation of the density function. It produces no graphical output. `akdensity` is a user-friendly wrapper that mimicks the syntax of the official `kdensity` and generates the 2-stage adaptive kernel density estimates by making repeated calls to `akdensity0`. The first call uses a fixed bandwidth and generates the local bandwidth factors, the second call uses the varying bandwidths obtained from the local bandwidth factors.

Equations (1) and (2) show that local bandwidth factors must be computed for each sample point. This requires an estimate of the pilot density function at each sample point. Computing a kernel density estimate for each sample point can be prohibitively slow for large datasets. To speed up calculations, `akdensity0` estimates the pilot density function for a grid of points (specified by the user), and uses linear interpolation to approximate the

[1] In the unweighted case, with a Gaussian kernel function, the methods are exactly as in Salgado-Ugarte et al. (1993) and Salgado-Ugarte & Pérez-Hernández (2003): estimates obtained with both `akdensity` and the existing `adgakern` or `varwiker` are identical, although `akdensity` offers some extra flexibility in practice.

density at sample points located between two grid points. It is thus useful to use a grid that spans outside of the data range. This procedure leads to considerable speed gains with large datasets.

`akdensity` is more limited than `kdensity` in one respect: The choice of the kernel function. Only Epanechnikov and Gaussian kernel functions have been implemented. Note, however, that these are popular choices and it is widely accepted that the choice of kernel is not a crucial issue.

5 Syntax

The syntax for `akdensity` follows `kdensity`'s:

`akdensity` *varname* [*weight*] [`if` *exp*] [`in` *range*] [`, nograph noadaptive generate(`*newvar_x newvar_density*`) n(`#`) width(`#`)` [`epan` | `gauss`] `normal student(`#`) at(`*var_x*`) stdbands(`#`) symbol(...) connect(...) title(`*string*`)` *graph_options*]

The only new options are `stdbands` and `noadaptive`. All the other options are described in [R] **kdensity**.

`stdbands(`#`)` requests the estimation of variability bands, and specifies the number of standard errors above and below the estimates to be used (a positive number). If the `generate` option is specified, the estimated bands are stored in two new variables *newvar_density_up* and *newvar_density_lo*.

`noadaptive` can be specified to obtain the standard fixed bandwidth kernel density estimate. The resulting density is exactly as produced by `kdensity`. This may be used to obtain the variability bands around the fixed kernel density estimates.

`akdensity` is compatible with both Stata 7 and Stata 8. It uses the newly implemented graphics engine if called by Stata 8, and runs otherwise the former engine for Stata 7. As a consequence, the allowed graphics options differ according to the release of Stata being used.[2]

The syntax for the engine command, `akdensity0`, is similar:

`akdensity0` *varname* [*weight*] [`if` *exp*] [`in` *range*] `, width(`# | *varname*`) at(`*var_x*`) generate(`*newvar_density*`)` [`stdbands(`#`) lambda(`*string*`)` [`epan` | `gauss`] `double`]

`at`, `width` and `generate` are not optional. Most options are as in `kdensity` or `akdensity`. Note, however, that the `width` option can here be either a scalar or a variable name containing observation-specific bandwidths. Also, `generate` must specify a *single* new variable name to store the estimated value of the density function at the grid points. The options specific to `akdensity0` are the following:

`lambda(`*string*`)` requests the estimation of local bandwidth factors based on the estimated density function, and specifies a new variable name where these values are to be stored.

`double` requests the use of double precision in the estimation of the density functions and standard error bands.

[2]Remember that, if need be, the Stata 7 engine can be called from within Stata 8 by using the `version 7:` prefix command, i.e. `version 7: akdensity (...)`.

References

Abramson, I. S. (1982), 'On bandwidth variation in kernel estimates–a square root law', *Annals of Statistics* **10**(4), 1217–1223.

Bowman, A. W. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, Vol. 18 of *Oxford Statistical Science Series*, Oxford University Press, Oxford, UK.

Burkhauser, R. V., Crews, A. D., Daly, M. C. & Jenkins, S. P. (1999), 'Testing the significance of income distribution changes over the 1980s business cycle: A cross-national comparison', *Journal of Applied Econometrics* **14**(3), 253–272.

Pagan, A. & Ullah, A. (1999), *Nonparametric Econometrics*, Themes in Modern Econometrics, Cambridge University Press, New York, USA.

Salgado-Ugarte, I. H. & Pérez-Hernández, M. A. (2003), 'Exploring the use of variable bandwidth kernel density estimators', *Stata Journal* **3**(2).

Salgado-Ugarte, I. H., Shimuzu, M. & Taniuchi, T. (1993), 'snp6: exploring the shape of univariate data using kernel density estimators', *Stata Technical Bulletin* **16**, 8–19.

Salgado-Ugarte, I. H., Shimuzu, M. & Taniuchi, T. (1995), 'snp6.2: practical rules for bandwidth selection in univariate density estimation', *Stata Technical Bulletin* **27**, 5–19.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman and Hall, London, UK.