The Stata Journal

The *Stata Technical Journal* (ISSN 1536-867X) is a publication of Stata Press, and Stata is a registered trademark of Stata Corporation.

The Stata Journal (2003)
3, Number 2, pp. 148–156

Adaptive kernel density estimation

Philippe Van Kerm
CEPS/INSTEAD, G.-D. Luxembourg

Abstract. This insert describes the module `akdensity`. `akdensity` extends the official `kdensity` that estimates density functions by the kernel method. The extensions are of two types: `akdensity` allows the use of an "adaptive kernel" approach with varying, rather than fixed, bandwidths; and `akdensity` estimates pointwise variability bands around the estimated density functions.

Keywords: st0037, adaptive kernel density, local bandwidths, variability bands

1 Overview

Stata offers one official command for nonparametric 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 and 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. First, `akdensity` allows the use of a varying, rather than fixed, bandwidth as in Salgado-Ugarte et al. (1993) and Salgado-Ugarte and 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. Second, `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.

1.1 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 and Azzalini (1997), or Pagan and Ullah (1999). Salgado-Ugarte et al. (1993), Salgado-Ugarte et al. (1995), and Salgado-Ugarte and 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, 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, $\widehat{f}(x)$, plus or minus b times the estimated standard error of $\widehat{f}(x)$. 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 emphasize that the bands quantify the variability of the density estimate but do not take the bias of the estimate into account, and thus do not provide a means of examining particular hypotheses about the density function (Bowman and Azzalini 1997, 29–30).

1.2 Methods and formulas

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. Thus, 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

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

where the x_is 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,

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

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 $\widehat{f}(x)$ given in Burkhauser et al. (1999):

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

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

2 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 mimics the syntax of the official `kdensity` and generates the two-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 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.

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

3 Syntax

The syntax for `akdensity` follows `kdensity`:

`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 otherwise runs 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`]

`width()`, `at()`, 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 (...)`.

4 Example

To illustrate the features of `akdensity` and compare the fixed and adaptive kernel density estimates, I use the coral-trout-length data illustrating both the STB article by Salgado-Ugarte et al. (1993) and the official documentation of `kdensity`. The data, available from the Stata Press web site, consist of 316 length observations of coral trout (in mm.). The graphics presented have been obtained with the Stata 7 graphics engine.

Let me first draw the default estimates obtained with both `kdensity` and `akdensity`. Both use the default number of 50 equally spaced grid points and use the "rule-of-the-thumb" global bandwidth described in [R] **kdensity**.

```
. use http:\\www.stata-press.com\data\r7\trocolen.dta, clear
. kdensity length, nogr gen(x fixed)
. akdensity length, nogr gen(adaptive) at(x)
Two-stage adaptive kernel density estimation
Step 1: Pilot density and local bandwidth factors estimation
Step 2: Adaptive kernel density estimation
. label var fixed "Fixed kernel"
. label var adaptive "Adaptive kernel"
. graph fixed adaptive x, xlab ylab c(ll) s(dS)
```

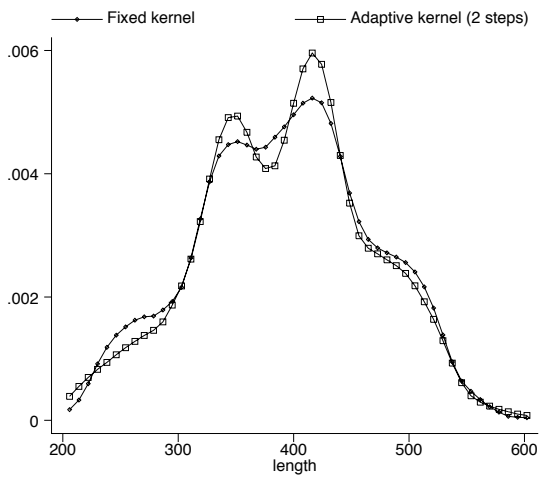

Figure 1: Fixed vs. adaptive kernel density estimates (bandwidth=20).

The main effect of using adaptive kernel estimation in this example is to separate out more clearly the two modes of the distribution (see Figure 1). The fixed bandwidth tends to oversmooth the middle of the distribution. On the contrary, the adaptive kernel estimate is smoother in the tails (especially in the lower tail).

One advantage of the methods implemented here is that the overall smoothing is still controlled by the choice of the global bandwidth. This allows fine tuning by changing the global bandwidth parameter. Figure 2 depicts the previous estimators obtained by setting the global bandwidth to 15 (the automatically computed global bandwidth as in Figure 1 is about 20). See Salgado-Ugarte and Pérez-Hernández (2003) for a detailed discussion of bandwidth selection procedures in this context.

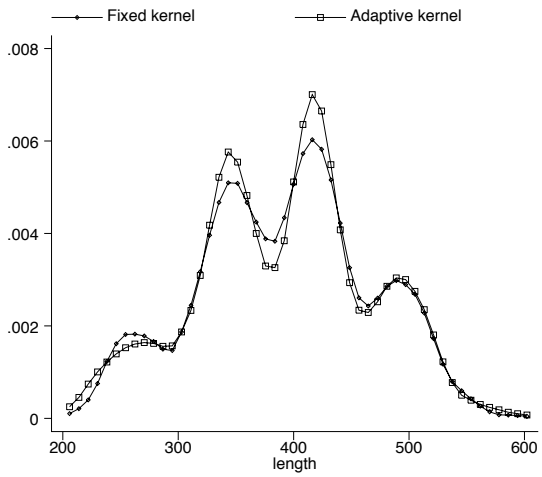

Figure 2: Fixed vs. adaptive kernel density estimates (bandwidth=15).

By drawing the standard error bands, it is possible to see the impact of varying the bandwidth on the variability of the estimates.[3] As mentioned earlier, adaptive kernel methods tend to reduce the variability of the estimates in areas of low density and increase it in areas with many observations. This accompanies a bias reduction (resp. increase) in areas where the data are numerous (resp. sparse). Variability bands for both the fixed and adaptive kernel density estimates in the current example are depicted in Figure 3. These pictures are obtained as follows (the commands also illustrate the use of the `n(#)` option to change the number of evaluation points, and of the `at(`*varname*`)` option for user-supplied grids):

```
. akdensity length, nogr stdbands(2) gen(x2 fixed3) noadapt n(200)
Standard kernel density estimation
. akdensity length, nogr stdbands(2) gen(adaptive3) at(x2)
Two-stage adaptive kernel density estimation
Step 1: Pilot density and local bandwidth factors estimation
Step 2: Adaptive kernel density estimation
. label var fixed3 "Fixed kernel"
. label var adaptive3 "Adaptive kernel"
. label var fixed3_up "with var. bands"
```

[3]Remember that the bands are centered on $\widehat{f}(x)$ and ignore the bias of the estimates.

```
. label var adaptive3_up "with var. bands"
. gr fixed3 adaptive3 fixed3_up adaptive3_up fixed3_lo adaptive3_lo x2,
> c(l[-]ll[-]ll[-]l)  s(iiiiii) pen(232323) xlab ylab
```

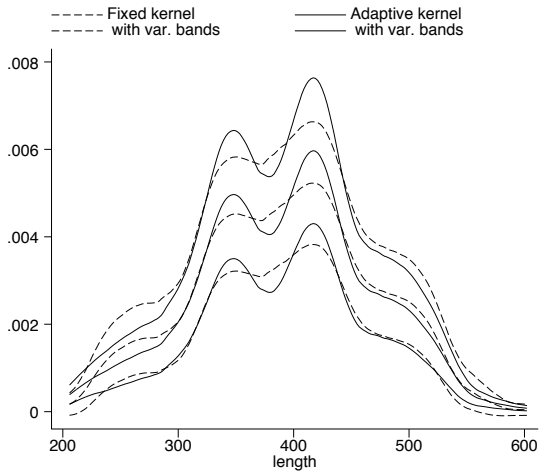

Figure 3: Fixed vs. adaptive kernel density estimates with variability bands.

Finally, to illustrate how the engine `akdensity0` can be used on its own, let me draw a three-step adaptive estimator. In this case, the kernel density estimate obtained in the second step is used as a new pilot density, and the estimation is repeated. The three-step estimator is obtained as follows:

```
. akdensity0 length, at(x) gen(pilot) width(20) lambda(lambda)
. gen hi = 20*lambda
. akdensity0 length, at(x) gen(pilot2) width(hi) lambda(lambda2)
. replace hi = 20*lambda2
(316 real changes made)
. akdensity0 length, at(x) gen(adaptive4) width(hi)
. label var adaptive "Adaptive kernel (2 steps)"
. label var adaptive4 "Adaptive kernel (3 steps)"
. gr fixed adaptive adaptive4 x, c(ll[-]l[_]) s(odS) xlab ylab
```

(*Continued on next page*)

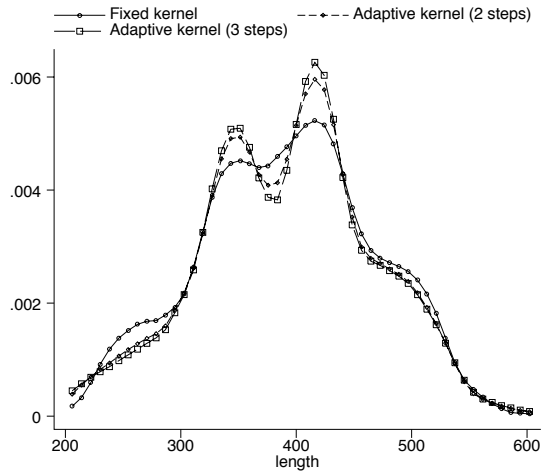

Figure 4: Fixed vs. two-step and three-step adaptive kernel density estimates.

The first call to `akdensity` generates the fixed bandwidth pilot estimate and a first set of local bandwidth factors. In the second call, these local factors multiply the global bandwidth to obtain an adaptive kernel density estimate. A second set of local bandwidth factors is constructed based on this new density estimate. Finally, a third call is made using this second set of local factors to obtain the final estimate. Note, however, from Figure 4, that this three-step estimate does not differ widely from the two-step estimates.

5 Acknowledgments

The `akdensity` module 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 document.)

6 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. and A. Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*, vol. 18. Oxford, UK: Oxford University Press.

Burkhauser, R. V., A. D. Crews, M. C. Daly, and S. P. Jenkins. 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. and A. Ullah. 1999. *Nonparametric Econometrics*. New York: Cambridge University Press.

Salgado-Ugarte, I. H. and M. A. Pérez-Hernández. 2003. Exploring the use of variable bandwidth kernel density estimators. *Stata Journal* 3(2): 133–147.

Salgado-Ugarte, I. H., M. Shimizu, and T. Taniuchi. 1993. snp6: Exploring the shape of univariate data using kernel density estimators. *Stata Technical Bulletin* 16: 8–19. In *Stata Technical Bulletin Reprints*, vol. 3, 155–173. College Station, TX: Stata Press.

—. 1995. snp6.2: Practical rules for bandwidth selection in univariate density estimation. *Stata Technical Bulletin* 27: 5–19. In *Stata Technical Bulletin Reprints*, vol. 5, 172–190. College Station, TX: Stata Press.

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

About the Author

Philippe Van Kerm is research associate at CEPS/INSTEAD (G.-D. Luxembourg). His research focuses on applied microeconometrics, with particular reference to income distribution issues.