Exploring High Dimensional Objects

Danny Kaye

Nottingham Trent University

Exploring High Dimensional Data

Exploratory data analysis is essential to serious analysis of data.

The most desirable and useful exploratory tool is often visualising data.

When the data are from familiar, low dimensional processes this is often fairly easy (e.g., we know lots of ways to display approximately normal data or deal with mild positive skew etc.).

When the data are from an unfamiliar process or of more than three dimensions traditional methods break down.

Histogram

Traditionally, for one dimensional data, the histogram provides simple density estimation, however, the information obtained is very dependant on the number and size of bins chosen and, even, the starting point of the first bin.

Histogram

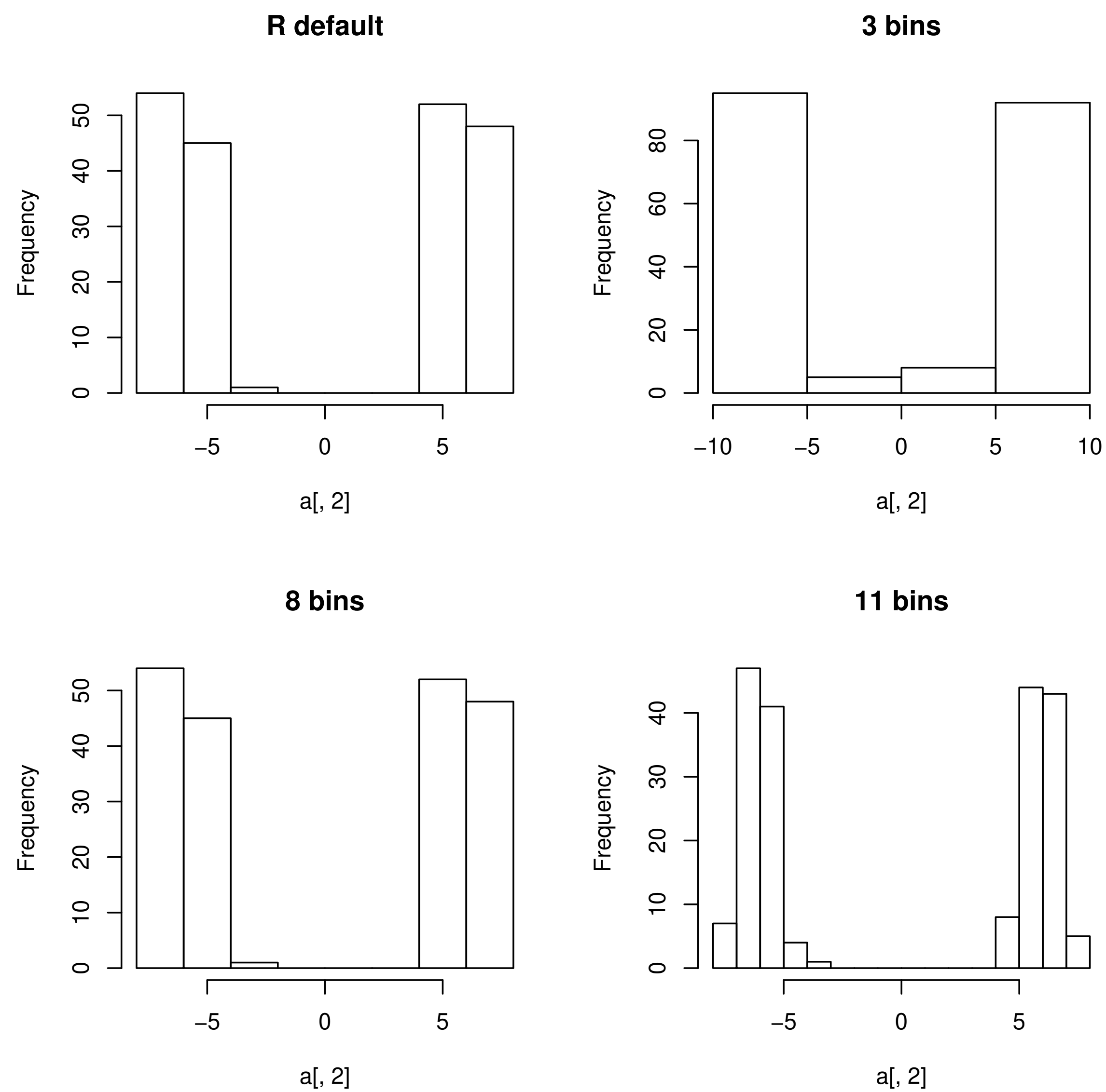


Figure 1: Four histograms of the same data.

Density Estimation

A better alternative to a histogram is a density estimator, this attempts to recover the underlying density that gave rise to a sample:

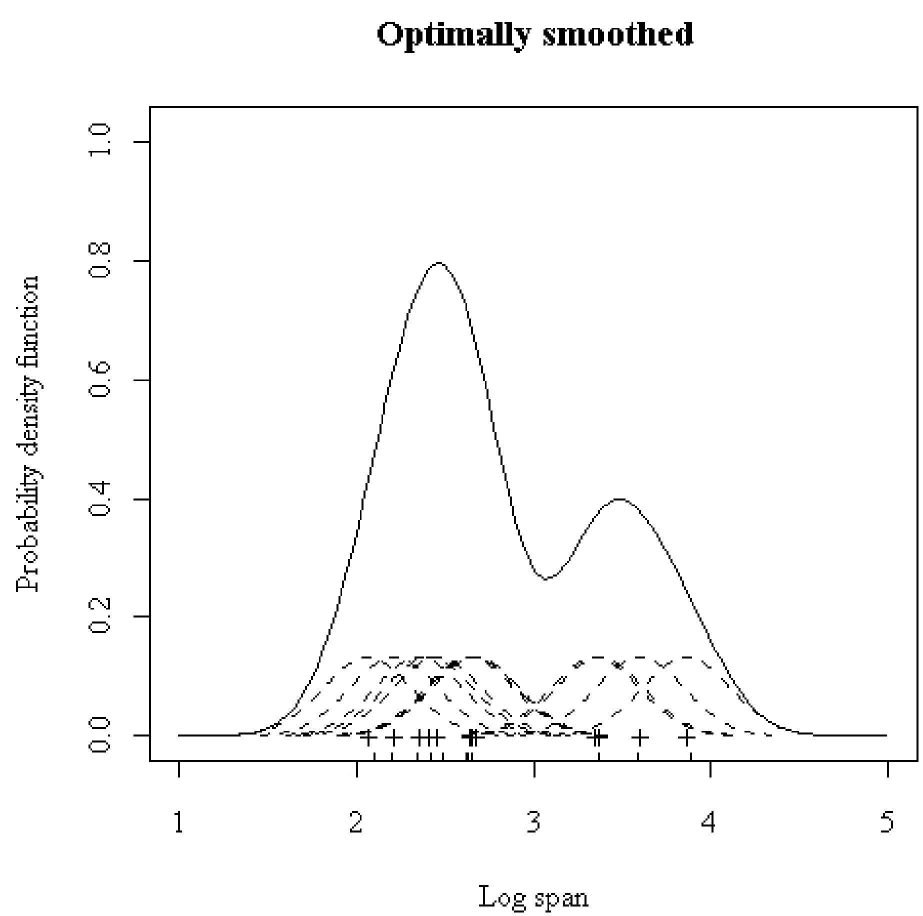


Figure 2: Kernel Density Estimation.

Kernel Density Estimation (kde)

Given data x from a density $f_X(\cdot)$ we can define an estimate of the underlying density as:

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

(see for example Silverman (1986, p15)). This is commonly known as the *Kernel density estimator* where h is the *window width*, *smoothing parameter* or *bandwidth*.

However, this has a problem in that the bandwidth is unknown and has to be estimated.

Bayes Kernel Density Estimation (bkde)

Make h a parameter of the problem, avoiding both specifying the bandwidth and the assumption that all projections from a density have the same smoothness.

Specify a KDE in terms of some function $K(\cdot|\mu, \theta)$, μ a location parameter and θ a scale parameter.

The estimate of $f_X(\cdot)$, based on the sample $\boldsymbol{x}$, at the point, $X = t$ is:

$$\hat{f}_X(t|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} K(t|x_i, \boldsymbol{\theta})$$

where n is the number of data items.

Likelihood

The need to choose h, (more generally, $\boldsymbol{\theta}$), is a problem of inference about $\boldsymbol{\theta}$. We have sample data $\boldsymbol{x}$ and a model from which to build a likelihood.

Belief about the smoothness can be expressed in a prior $p(\boldsymbol{\theta})$. By taking $\hat{f}_X(x)$ to be a model for the data we can construct a likelihood function.

Defining $\boldsymbol{x}_{(i)} = \{x_1, x_2, \ldots, x_i\}$ and assuming no particular ordering to the sample data, we may write the likelihood as:

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{x}_{(i-1)},\boldsymbol{\theta})$$

where $p(x_i|\boldsymbol{x}_{(i-1)},\boldsymbol{\theta})$ is the probability density of x_i, given by our chosen model.

Likelihood

This is of the style commonly used for a time series $\{x_i\}$, but is more generally true. If, for example, the density were $f_X(x|\boldsymbol{\theta})$ we would have the usual formulation for a random sample:

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta})$$

If we now take the KDE as our model for observation x_i having seen $\boldsymbol{x}_{(i-1)}$ we can rewrite our likelihood as:

$$\ell(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} \hat{f}_X(x_i|\boldsymbol{x}_{(i-1)},\boldsymbol{\theta})$$

Prior

Choice of prior $p(\boldsymbol{\theta})$ may, typically, reflect belief about the smoothness of the true density that $\hat{f}_X()$ is intended to estimate.

For example for the basic KDE we take $\boldsymbol{\theta} = \{h\}$, in this case a normal prior for $\boldsymbol{\theta}$ with a high value of μ implies a belief in a rather smooth, near uniform density while a low value implies a non-smooth density.

The variance is chosen to reflect the strength of belief, a small value indicating a strongly held belief and a large value the opposite.

Posterior

With small data sets the choice of prior has a large influence on the smoothness of the final estimate. We may expect that with large data sets the likelihood will dominate in the posterior and we will choose roughly non-informative priors as a convenient initial choice.
Once we have the likelihood function and the prior we apply Bayes theorem to obtain the posterior density:

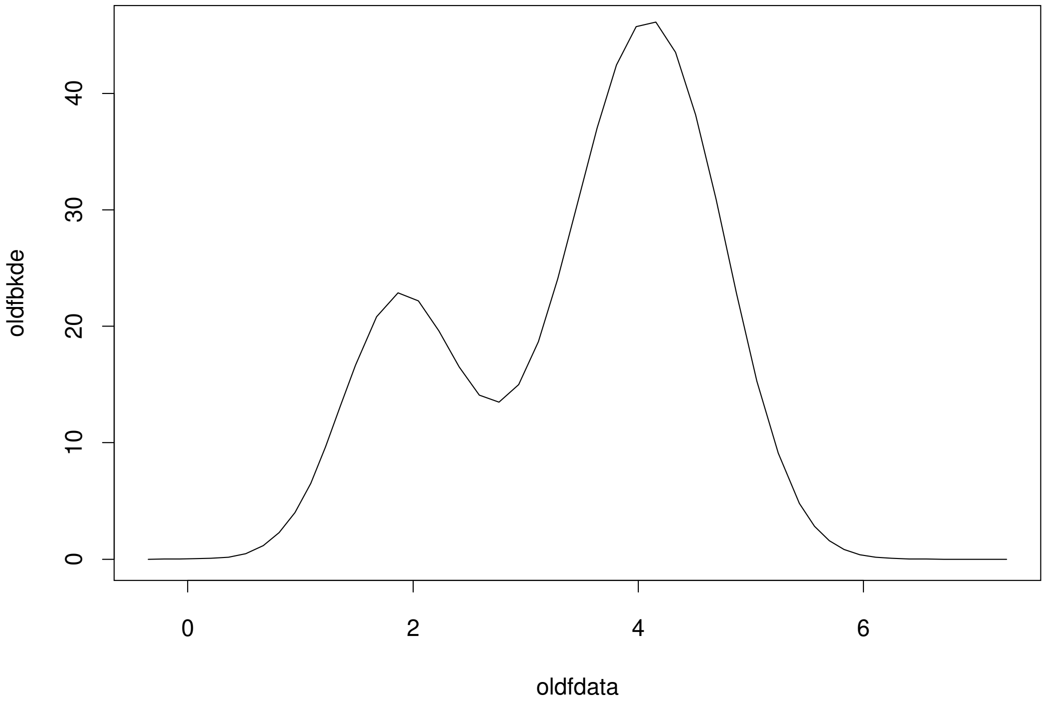
$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{\ell(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{x}_{(n_0)})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}$$

in which:

$$p(\boldsymbol{x}) = \int_{\Theta} \ell(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{x}_{(n_0)})p(\boldsymbol{\theta})d\theta$$

Example

Using the posterior θ in a predictive density gives an easy to plot output; the well known Old Faithful data, Weisberg (1980), gives the following:



Adaptive Bayesian Kernel Estimation.

As can be seen from the figure the density has two areas of high density and three of relatively low density.

Intuitively a large bandwidth is required in an area of low density and a small bandwidth in an area of high density, smoothing out the tails and revealing more detail in the high density areas. This leads to the notion of an adaptive estimate where the bandwidth is inversely proportional to the local density.

We introduce an estimate that works in a similar way. An initial KDE is used to modify the bandwidth used in the final estimate giving:

$$\hat{f}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(h\lambda_i)^d} K\left\{\frac{t - x_i}{h\lambda_i}\right\}$$

The steps given in Silverman (1986) for an adaptive kde are

1. Find a *pilot estimate* $\tilde{f}(\boldsymbol{t})$ that satisfies $\tilde{f}(x_i) > 0$ for all i
2. Define *local bandwidth factor* λ_i by:

$$\lambda_i = \left\{ \frac{\tilde{f}(x_i)}{g} \right\}^{-\alpha}$$

   where g is the geometric mean of the $\tilde{f}(x_i)$: and α is the *sensitivity parameter,* a number satisfying $0 \leq \alpha \leq 1$.

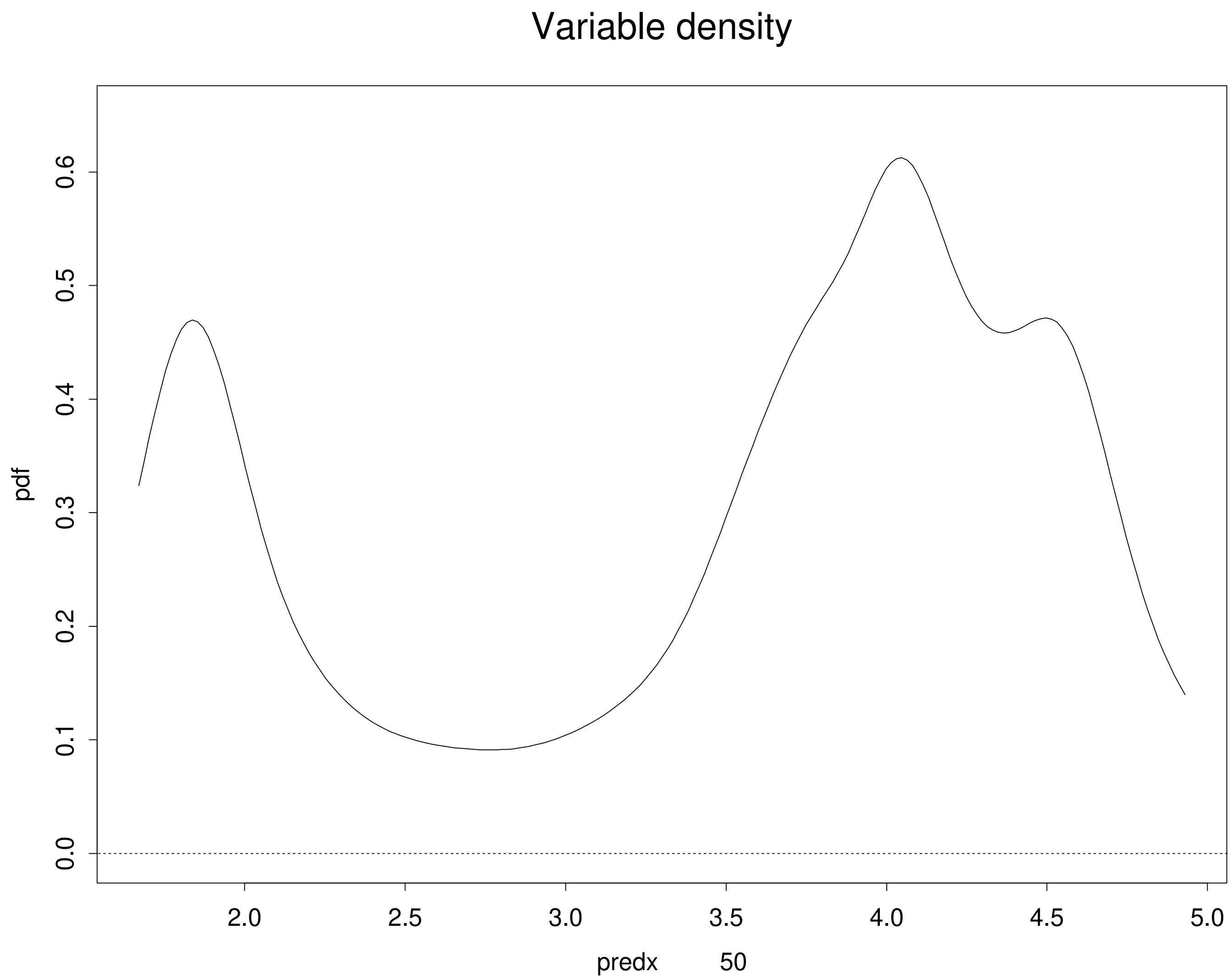
3. Define the *adaptive kernel estimate* $\hat{f}$ by:

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{(h\lambda_i)^d} K \left\{ \frac{t - x_i}{h\lambda_i} \right\}$$

where K is the kernel function and h is the bandwidth. As in the ordinary kernel method, K is a symmetric function integrating to unity.

Example

Examining the Old Faithful data again gives the following:



Visualisation

Real data is seldom univariate and so, ideally, we need a way of visualising data in a human friendly way. If we take a three dimensional data set we can project that data onto a one or two dimensional target and so get an image of what it looks like from one direction. If we rotate the target and repeat we can get a series of views known as a spin plot.

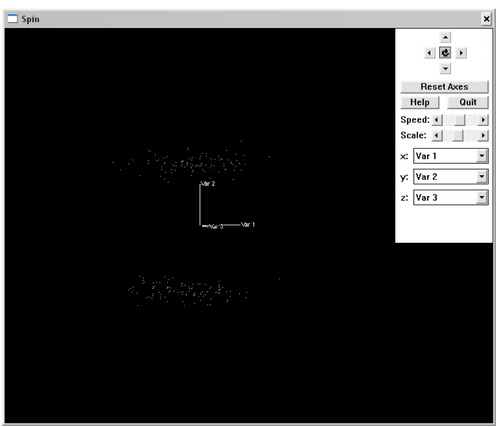


Figure 3: Views of Spin(a)

The Grand Tour and Projection Pursuit

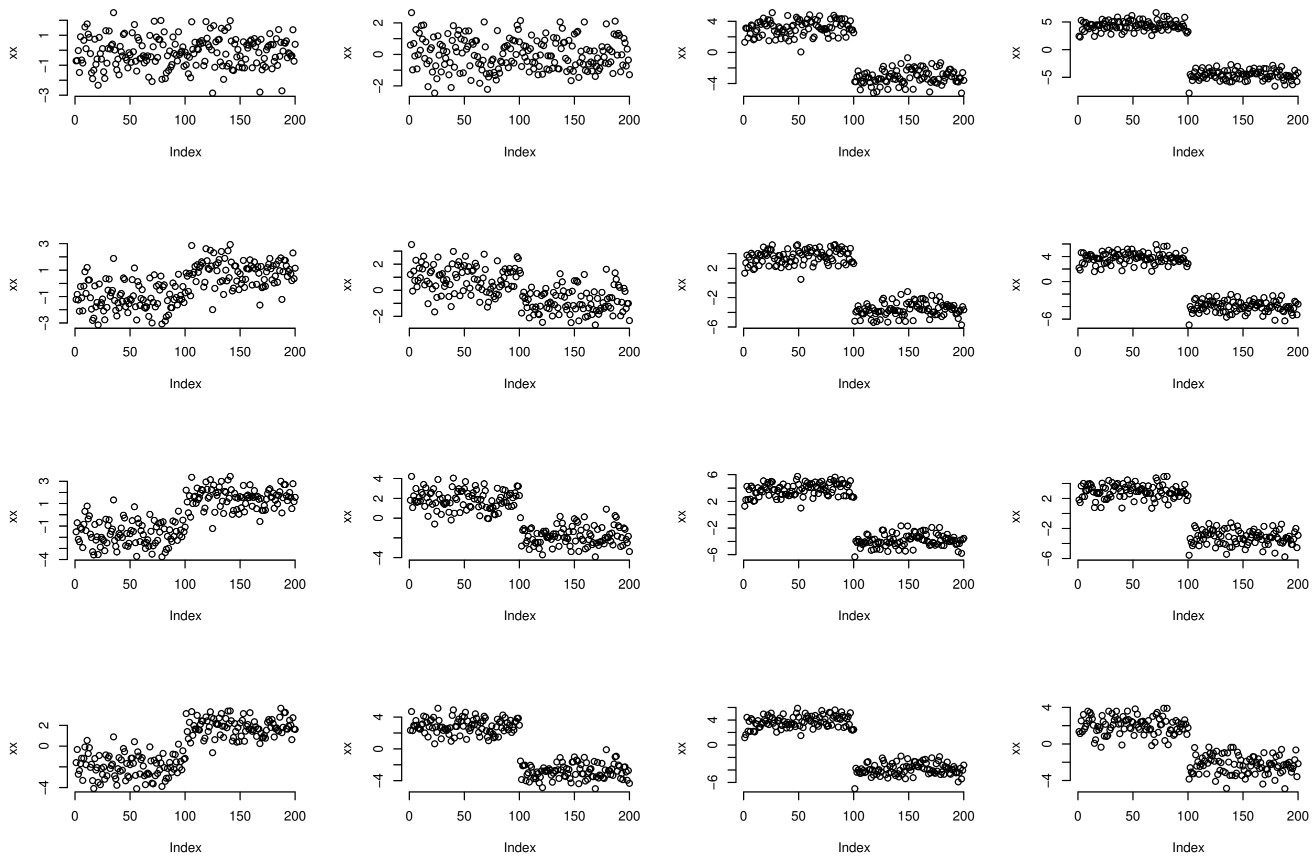
Extending this to more than three dimensions turns out to be easy and leads to two techniques.

The Grand Tour is

> *"a method for viewing multivariate statistical data via orthogonal projections onto a sequence of two dimensional subspaces."*
>
> Asimov (1985)

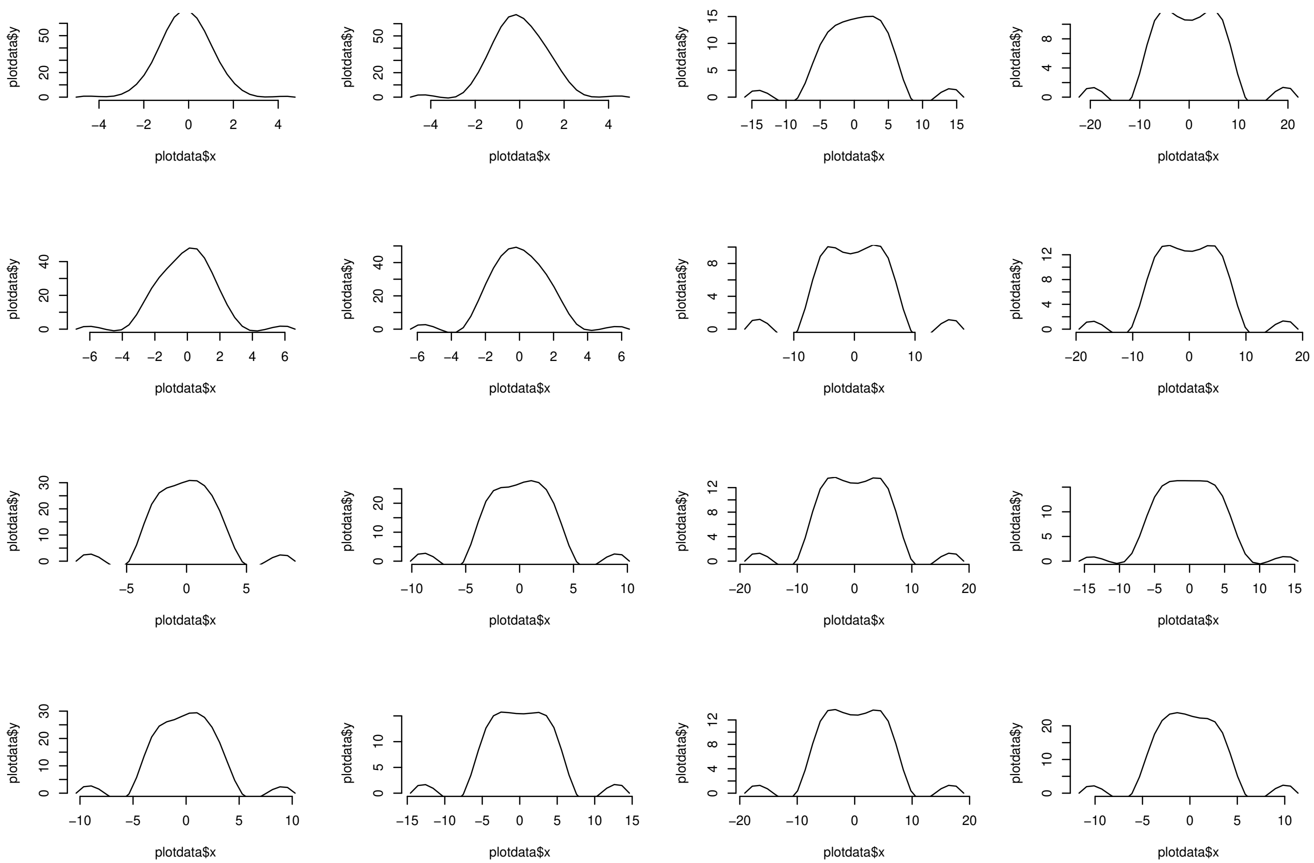
where the sequence is dense in the space. Projection pursuit, Jones (1983), is also a sequence of projections but, rather than attempting to create a dense sequence, it is steered by some definition of interesting when applied to the sequence of subspaces.

Grand Tour



Grand Tour example with BKDE

Finally we add density estimation to the grand tour and obtain a series of estimated densities:



References

[1] D Asimov. The grand tour - a tool for viewing multidimensional data. *SIAM journal of scientific, statistical computing*, 6(1):128–143, January 1985.

[2] M. C. Jones. *The projection pursuit algorithm for exploratory data analysis*. PhD thesis, University of Bath, 1983.

[3] B. W. Silverman. *Density estimation*. Chapman and Hall, London, 1986.

[4] S. Weisberg. *Applied linear regression*. Wiley, Sanford, 1980.