Density Estimation (II)

- Yesterday
  - Overview & Issues
  - Histogram
  - Kernel estimators
  - Ideogram
- Today
  - Further development of optimization
  - Estimating variance and bias
  - Adaptive kernels
  - Multivariate kernel estimation

Some References (I)

Richard A. Tapia & James R. Thompson, Nonparametric Density Estimation, Johns Hopkins University Press, Baltimore (1978).

David W. Scott, Multivariate Density Estimation, John Wiley & Sons, Inc., New York (1992).

Adrian W. Bowman and Adelchi Azzalini, Applied Smoothing Techniques for Data Analysis, Clarendon Press, Oxford (1997).

B. W. Silverman, Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability, Chapman and Hall (1986);
http://nedwww.ipac.caltech.edu/level5/March02/Silverman/Silver_contents.html

K. S. Cranmer, "Kernel Estimation in High Energy Physics", Comp. Phys. Comm. **136**, 198 (2001) [hep-ex/0011057v1];
http://arxiv.org/PS_cache/hep-ex/pdf/0011/0011057.pdf

Some References (II)

M. Pivk & F. R. Le Diberder, "sPlot: a statistical tool to unfold data distributions", Nucl. Instr. Meth. A **555**, 356 (2005).

R. Cahn, "How sPlots are Best" (2005),
http://babar-hn.slac.stanford.edu:5090/hn/aux/auxvol01/rncahn/rev_splots_best.pdf

BaBar Statistics Working Group, "Recommendations for Display of Projections in Multi-Dimensional Analyses",
http://www.slac.stanford.edu/BFROOT/www/Physics/Analysis/Statistics/Documents/MDgraphRec.pdf

Additional specific references will noted in the course of the lectures.

Optimization (continued)

The MSE (discussed yesterday) for a density is a measure of uncertainty at a point. It is useful to somehow summarize the uncertainty over all points in a single quantity. We wish to establish a notion for the "distance" from function $\widehat{p}(x)$ to function $p(x)$.

- Familiar subject for physicists; just dealing with normed vector spaces!
  - Choice of norm is a bit arbitrary; obvious extremes are:

$$\|\widehat{p}(x) - p(x)\|_{L_\infty} \equiv \sup_x |\widehat{p}(x) - p(x)|$$

$$\|\widehat{p}(x) - p(x)\|_{L_1} \equiv \int |\widehat{p}(x) - p(x)|\, dx.$$

- We'll use (following the crowd; for convenience, not necessarily because it is "best") the L_2 norm, or more precisely, the "Integrated Squared Error", ISE:

$$\text{ISE} \equiv \int [\widehat{p}(x) - p(x)]^2\, dx.$$

MISE

In fact, the ISE is still a difficult beast, as it depends on the true density, the estimator, and the sampled data. We may remove this latter dependence by evaluating the "Mean Integrated Squared Error" (MISE):

$$\text{MISE} \equiv E\left[\text{ISE}\right] = E\left[\int \left[\widehat{p}(x) - p(x)\right]^2 dx\right]$$

$$= \int E\left[\left(\widehat{p}(x) - p(x)\right)^2\right] dx = \int \text{MSE}\left[\widehat{p}(x)\right]\, dx \equiv \text{IMSE}.$$

(Exercise: prove MISE=IMSE).

Consistency

A desirable property of an estimator is that the error decreases as the number of samples increases. (Familiar notion from parametric statistics.)

Def: A density estimator $\widehat{p}(x)$ is consistent if:

$$\text{MSE}\left[\widehat{p}(x)\right] \equiv E\left[\widehat{p}(x) - p(x)\right]^2 \to 0$$

as $n \to \infty$.

Optimal Histograms? (I)

Noting that the bin contents of a histogram are binomial-distributed, we could show (exercise!) that, for the histogram density estimator $\widehat{p}(x) = h(x)/nw$:

$$\text{Var}\left[\widehat{p}(x)\right] \leq \frac{p(x_j^*)}{nw}$$

$$\left|\text{Bias}\left[\widehat{p}(x)\right]\right| \leq \gamma_j w,$$

where:

- $x \in$ bin j,
- x_j^* is defined (and exists by mean value theorem) by:

$$\int_{\text{bin}j} p(x)\,dx = wp(x_j^*),$$

- γ_j is a positive constant (existing by assumption) such that

$$|p(x) - p(x_j^*)| < \gamma_j |x - x_j^*|, \quad \forall x \in \text{bin } j,$$

- equality is approached as the probability to be in bin j decreases (e.g., by decreasing bin size).

Optimal Histograms? (II)

Thus,

$$\mathrm{MSE}\left[\widehat{p}(x)\right] = E\left[\widehat{p}(x) - p(x)\right]^2 \leq \frac{p(x_j^*)}{nw} + \gamma_j^2 w^2.$$

Thm: The MSE of the histogram estimator $\widehat{p}(x) = h(x)/nw$ is consistent if the bin width $w \to 0$ as $n \to \infty$ such that $nw \to \infty$.

- Note that the $w \to 0$ requirement insures that the bias will approach zero, according to our earlier discussion.
- The $nw \to \infty$ requirement ensures that the variance asymptotically vanishes.

Thm: The MSE(x) bound above is minimized when

$$w = w^*(x) = \left[\frac{p(x_j^*)}{2\gamma_j^2 n}\right]^{1/3}.$$

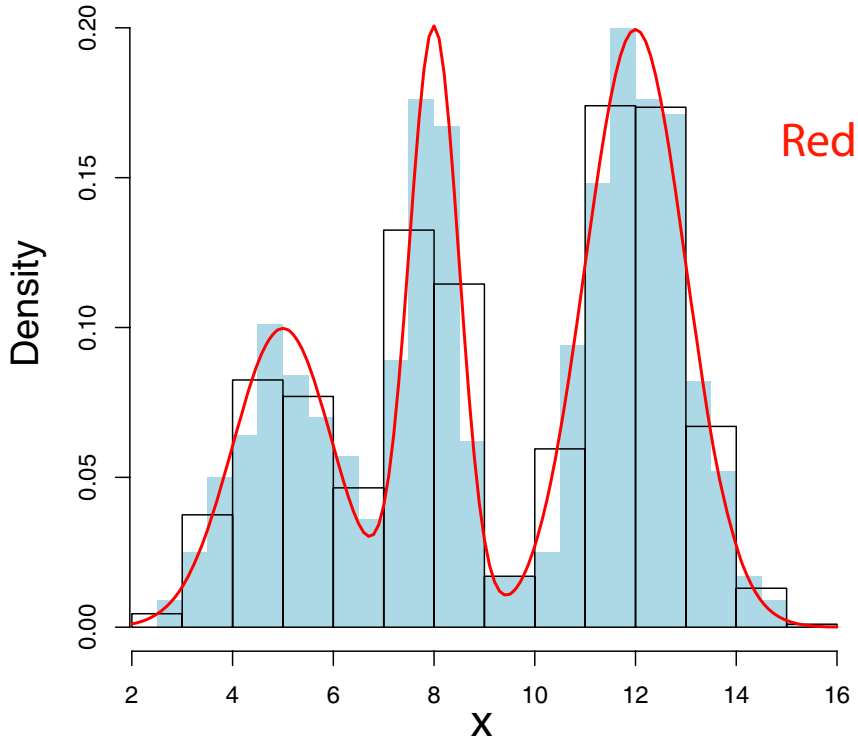
Optimal Histograms? (III)

Notes:

- The optimal bin size decreases as $1/n^{1/3}$.
- The $1/n$ dependence of the variance is our familiar result for Poisson statistics.
- The optimal bin size depends on the value of the density in the bin. Suggests "adaptive binning"... [However, Scott: "...in practice there are no reliable algorithms for constructing adaptive histogram meshes."]
- Alternatively, the MISE error is minimized (Gaussian kernel, asymptotically, for normally distributed data...) when

$$w^* = \left(\frac{4}{3}\right)^{1/5} \sigma n^{-1/5}.$$

Optimal Histograms? (IV)

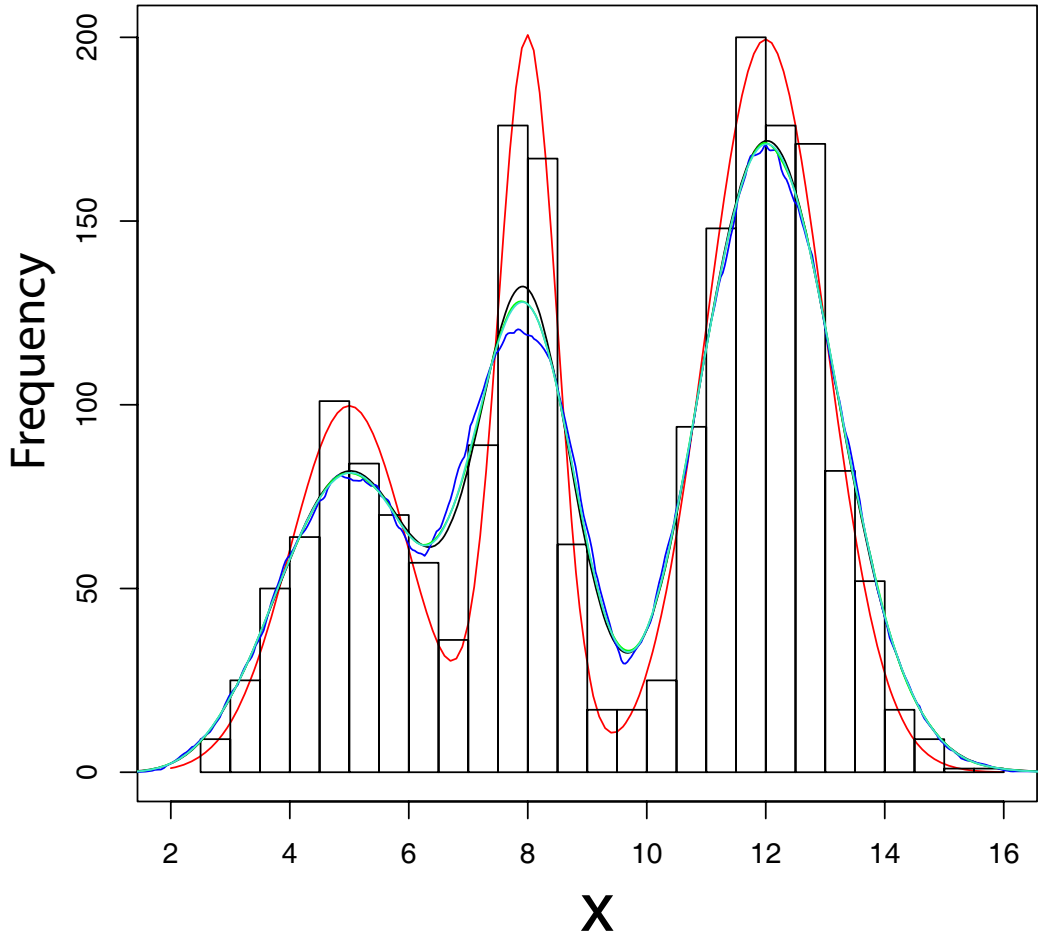


The "standard rules" (Sturges, Scott, Freedman-Diaconis) correspond roughly to the coarser binning above.

Impression: The standard rules for selecting constant bin widths tend to be lower limits on the optimum.

Choice of Kernel

Often, the form of the kernel used doesn't matter very much:



Red: Sampling PDF
Black: Gaussian kernel
Blue: Rectangular kernel
Turquoise: Cosine kernel
Green: Triangular kernel

Application of the Bootstrap to Density Estimation

A means to evaluating how much to trust our density estimate.

Bootstrap Algorithm:

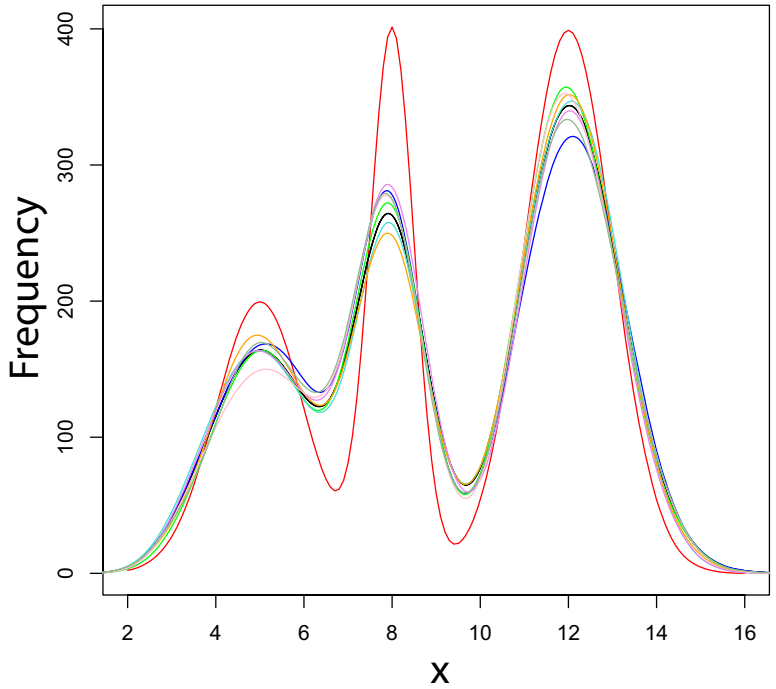
1. Form density estimate $\widehat{p}$ from data $\{x_i; i = 1, \ldots, n\}$.
2. Resample (uniformly) n values from $\{x_i; i = 1, \ldots, n\}$, with replacement, obtaining $\{x_i^*; i = 1, \ldots, n\}$ (bootstrap data).
3. Form density estimate $\widehat{p}^*$ from data $\{x_i^*; i = 1, \ldots, n\}$.
4. Repeat steps 2&3 many (N) times to obtain a family of bootstrap density estimates $\{\widehat{p}_i^*; i = 1, \ldots, N\}$.
5. The distribution of $\widehat{p}_i^*$ about $\widehat{p}$ mimics the distribution of $\widehat{p}$ about p.

Does the Bootstrap Distribution Really Work?

Not quite: Consider, for kernel density estimator (Exercise),

$$E\left[\widehat{p}^*(x)\right] = E\left[k(x - x_i^*; w)\right] = \widehat{p}(x).$$

Thus, the bootstrap distribution about $\widehat{p}$ does not reproduce the bias which may be present in $\widehat{p}$ about p. However, it does properly reproduce the variance of $\widehat{p}$.



Red: Sampling PDF
Black: Gaussian Kernel Estimator
Other colors: Resampled Kernel Estimators

Estimating Bias: The Jackknife (I)

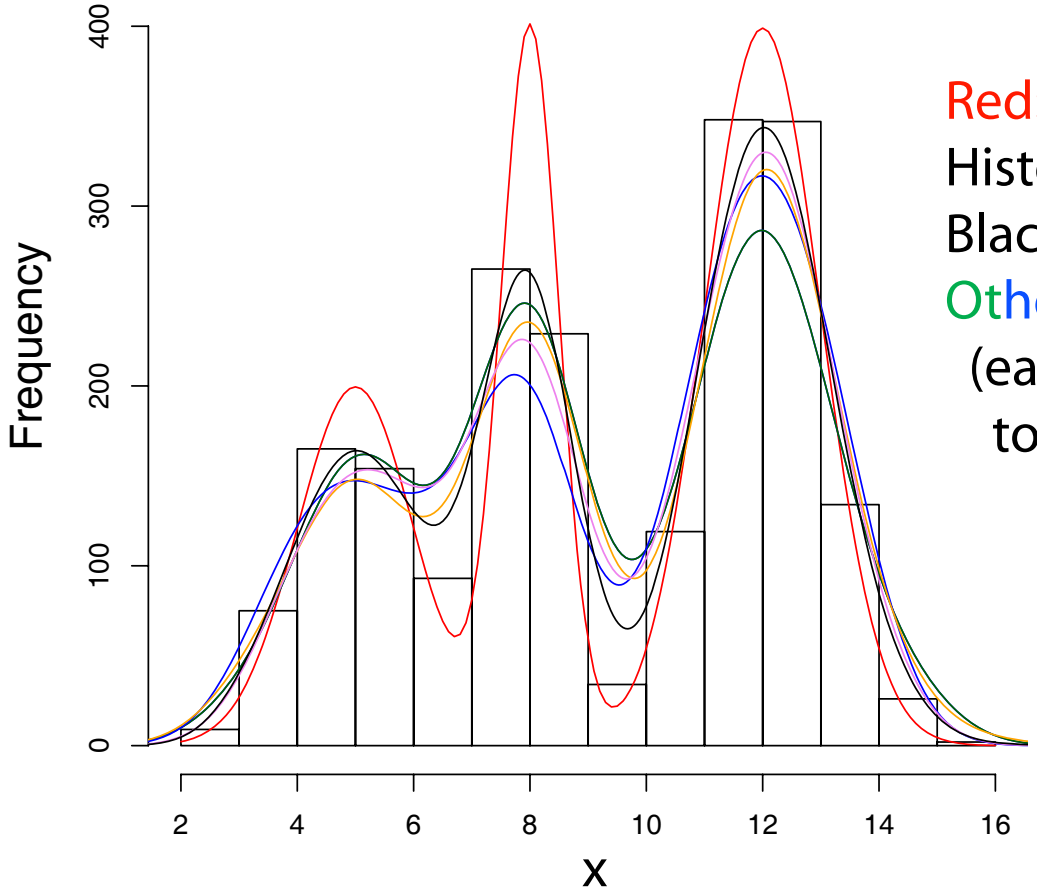
- The idea:
  - Bias depends on sample size.
  - Assume the bias vanishes asymptotically.
  - We can use the data to estimate the dependence of bias on sample size.
- Jackknife algorithm:

(1) Divide data into k random disjoint subsamples.

(2) Evaluate estimator for each subsample.

(3) Compare average of estimates on subsamples with estimator based on full dataset.

Estimating Bias: The Jackknife (II)



Red: Sampling PDF
Histogram: 2000 samples
Black: Gaussian kernel estimator
Other Colors: Kernel jackknife estimators
(each is on 500 samples; to be averaged together for jackknife)

- Jackknife techniques may also be used to reduce bias (see Scott)

Cross-validation (I)

- Similar to the jackknife, but different in intent, is the cross-validation procedure (see also Ilya Narsky lectures).
- In density estimation, cross-validation is used to optimize smoothing bandwidth selection.
  - Improves on "theoretical" optimizations by making use of the actual sampling distribution, via the available samples.

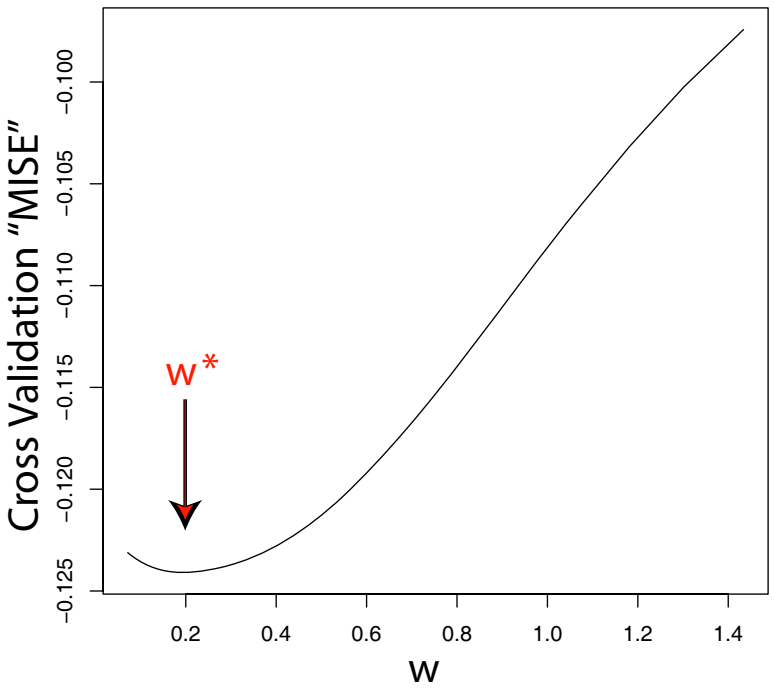
Cross-validation (II)

- Basic method ("leave-one-out cross-validation"):
  - Form n subsets of the dataset, each one leaving out a different datum. Use subscript $-i$ to denote subset omitting datum x_i.
  - For density estimator $\widehat{p}(x; w)$ evaluate the following average over these subsets:

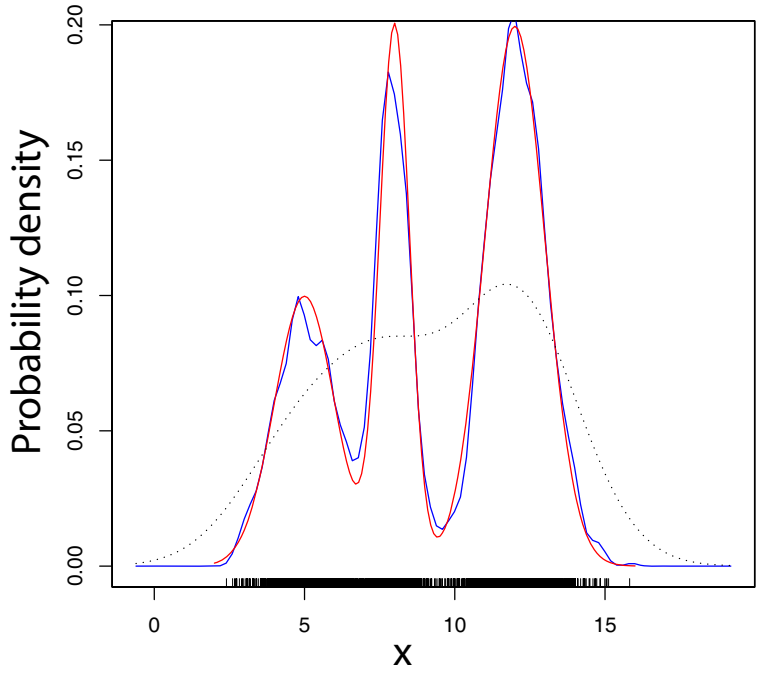
$$\frac{1}{n}\sum_{i=1}^{n}\int \widehat{p}^2_{-i}(x)\,dx - \frac{2}{n}\sum_{i=1}^{n}\widehat{p}_{-i}(x_i).$$

    [Exercise: Show that the expectation of this quantity is the MISE of $\widehat{p}$ for $n-1$ samples, except for a term which is independent of w.]
  - Evaluate the dependence of this quantity on smoothing parameter w, and select the value w^* for which it is minimized.

Cross-validation (III)



Choosing smoothing parameter which minimizes estimated MISE.



Red: Actual PDF

Blue: Kernel estimate, using smoothing from cross-validation optimization

Dots: Kernel estimate with a default value for smoothing

Adaptive Kernel Estimation (I)

- We saw in our discussion of histograms that it is probably more optimal to use variable bin widths. This applies to other kernels as well.
- Indeed, the use of a fixed smoothing parameter, deduced from all of the data introduces a non-local, hence parametric, aspect into the estimation. It is more consistent to look for smoothing which depends on data locally. This is "Adaptive kernel estimation."

Adaptive Kernel Estimation (II)

- We argue that the more data there is in a region, the better that region can be estimated. Thus, in regions of high density, we should use narrower smoothing. In Poisson statistics (e.g., histogram binning), the relative uncertainty scales as

$$\frac{\sqrt{N}}{N} \propto \frac{1}{\sqrt{p(x)}}.$$

Thus, in the region containing x_i, the smoothing parameter should be:

$$w(x_i) = w^*/\sqrt{p(x_i)}.$$

Two issues:

(1) What is w^*?

(2) We don't know $p(x)$!

Adaptive Kernel Estimation (III)

For $p(x)$, we may try substituting our fixed kernel estimator, call it $\widehat{p}_0(x)$.

For w^*, we use dimensional analysis:

$$D\left[w(x_i)\right] = D[x]; \quad D\left[p(x)\right] = D[1/x] \quad \Rightarrow \quad D[w^*] = D\left[\sqrt{x}\right] = D\left[\sqrt{\sigma}\right].$$

Then, e.g., using the "MISE-optimized" choice earlier, we iterate on our fixed kernel estimator to obtain:

$$\widehat{p}_1(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{w_i} K\left(\frac{x - x_i}{w_i}\right),$$

where

$$w_i = w(x_i) = \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\rho\sigma}{\widehat{p}_0(x_i)}} n^{-1/5}.$$

ρ is a factor which may be further optimized, or typically set to one.

The iteration on the fixed-kernel estimator nearly removes the dependence on our intitial choice of w. [See refs for: Discussion of boundaries.]

KEYS Adaptive Kernel Estimation

The KEYS ("Kernel Estimating Your Shapes") PDF is a package for adaptive kernel estimation, see KS Cranmer reference.

http://java.freehep.org/jcvslet/JCVSlet/diff/freehep/freehep/hep/aida/ref/pdf/NonParametricPdf.java/1.1/1.2

Implementation in RooFit:

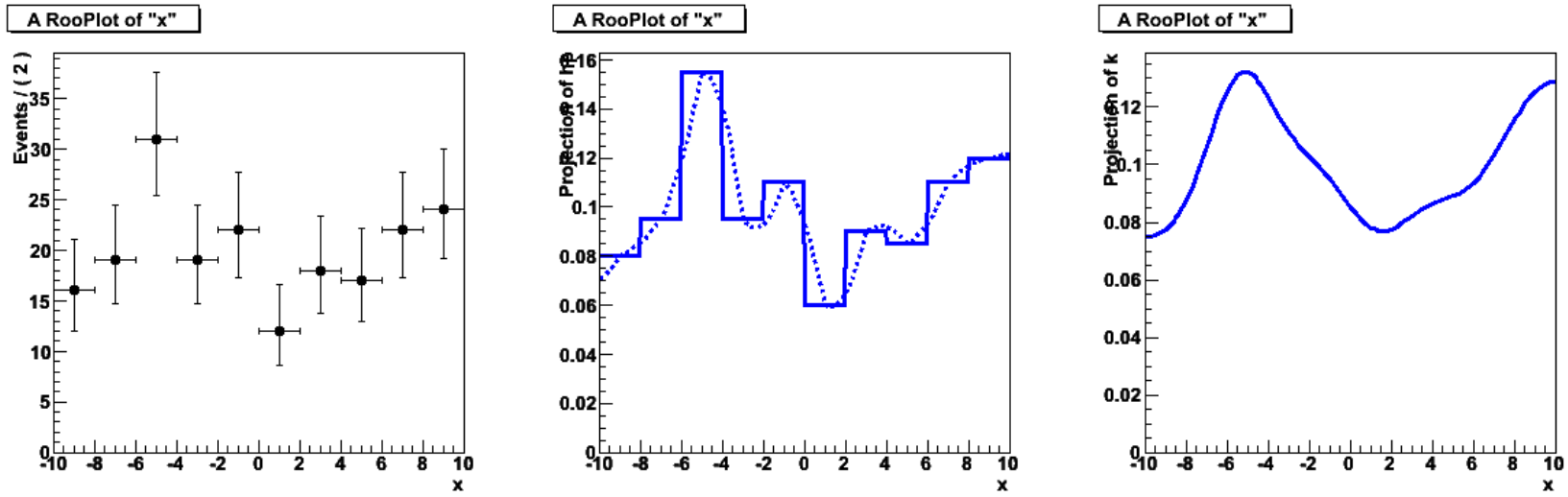


Figure 9 – Non-parametric p.d.f.s: Left: histogram of unbinned input data, Middle: Histogram-based p.d.f (2nd order interpolation), Right: KEYS p.d.f from original unbinned input data.

from: W. Verkerke and D. Kirkby, RooFit Users Manual V2.07:

http://roofit.sourceforge.net/docs/RooFit_Users_Manual_2.07-29.pdf

KEYS (II): Example

Figure showing comparison of KEYS adaptive kernel estimator (dashed curve) with the true PDF (solid curve):

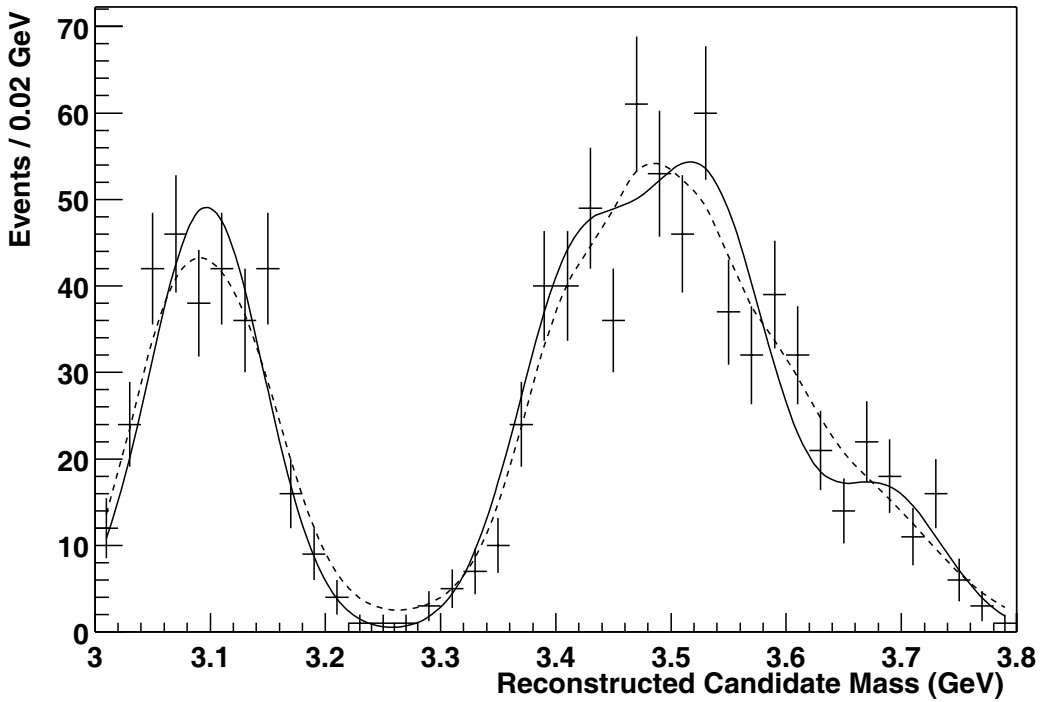


Figure 12: A comparison of the histogram of 1000 events (data points) with the underlying PDF used to generate these events (solid curve) and a non-parameteric KEYS model of these events (dashed curve).

[From BaBar BAD 18v14 (internal document, with permission. . .)]

Multivariate Kernel Estimation (I)

- Besides the curse of dimensionality, the multi-dimensional case introduces the complication of covariance.
- When using a product kernel, the local estimator has diagonal covariance matrix.
- In principle, we could apply a local linear transformation of the data to a coordinate system with diagonal covariance matrices. This amounts to a non-linear transfomation of the data in a global sense, and may not be straightforward. We can at least work in the system for which the overall covariance matrix of the data is diagonal.

Multivariate Kernel Estimation (II)

- If $\{y_i\}$ is the suitably diagonalized data, the product fixed kernel estimator in d dimensions is

$$\widehat{p}_0(y) = \frac{1}{n}\sum_{i=1}^{n}\left[\prod_{j=1}^{d}\frac{1}{w_j}K\left(\frac{y^{(j)} - y_i^{(j)}}{w_j}\right)\right],$$

  where $y^{(j)}$ denotes the j-th component of the vector y.

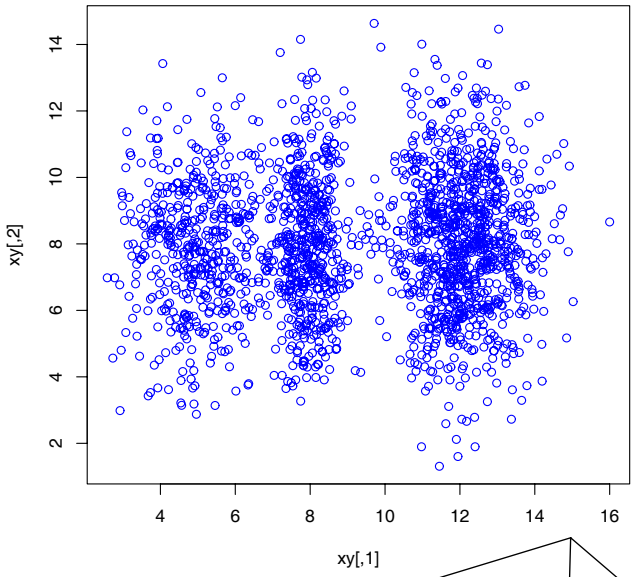
- The asymptotic, normal MISE-optimized smoothing parameters are now:

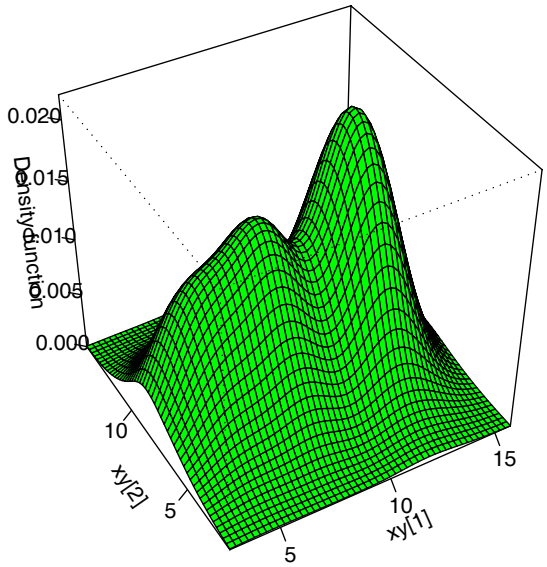
$$w_j = \left(\frac{4}{d+2}\right)^{1/(d+4)}\sigma_j n^{-1/(d+4)}.$$

- The corresponding adaptive kernel estimator follows the discussion as for the univariate case. An issue in the scaling for the adaptive bandwidth arises when the multivariate data is approximately sampled from a lower dimensionality than the dimension d. See references.

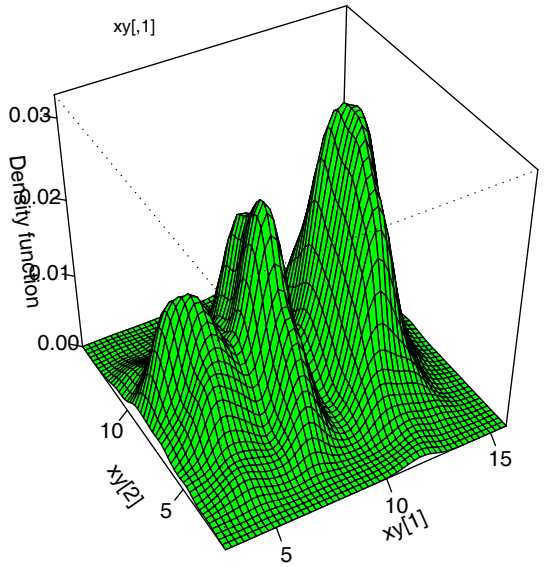
Multivariate Kernel Estimation: Example (I)

Example in which the sampling distribution has diagonal covariance matrix (locally and globally).





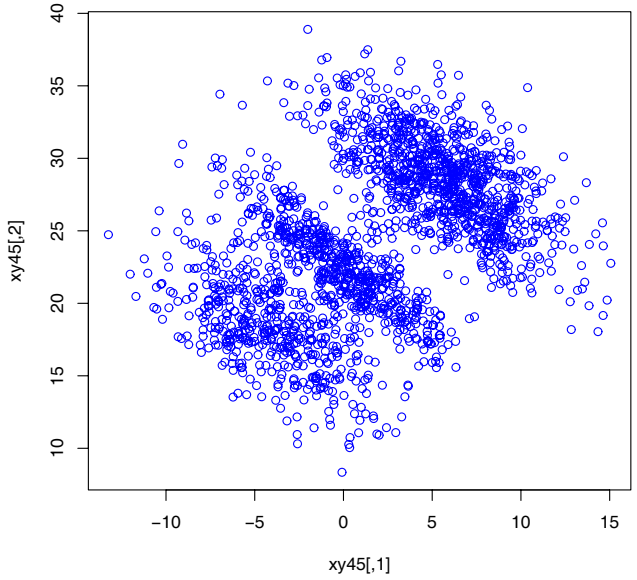
Default smoothing (w)

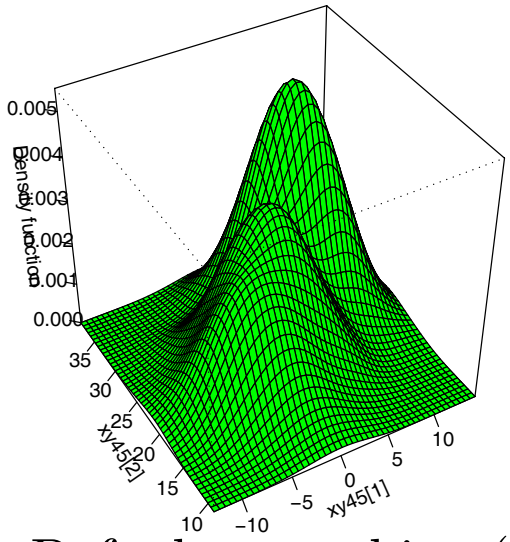


$w/2$ smoothing

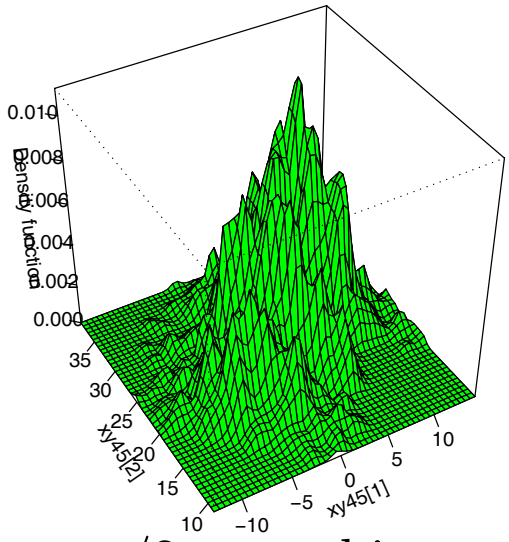
Multivariate Kernel Estimation: Example (II)

Example in which the sampling distribution has non-diagonal covariance matrix.

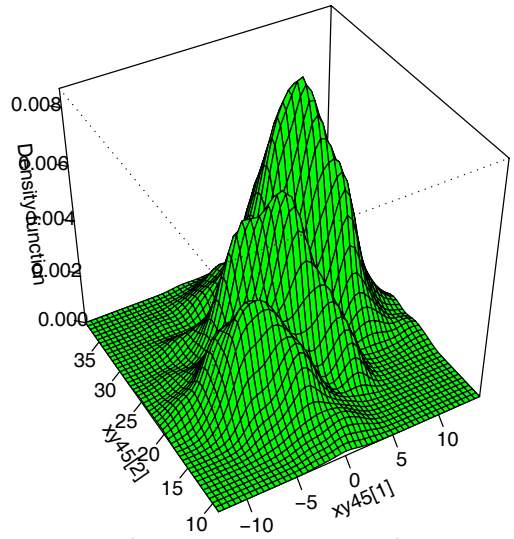




Default smoothing (w)



$w/2$ smoothing



Intermediate smoothing

Summary

We have looked at:

- Optimization
  - Cross-validation
  - Adaptive kernels
- Error estimation
  - Variance (bootstrap)
  - Bias (jackknife)
- A complete analysis puts these ingredients together.
- Multivariate kernel estimation

Next: Series estimation; Monte Carlo weighting; unfolding; non-parametric regression; sPlots.