Statistics IV

Frank Porter

September 21, 2013

Plan for the Statistics Lectures

- Lecture I (Wednesday, September 18, 11:45-12:30)
  1. Important probability concepts
  2. Point estimation
- Lecture II (Thursday, September 19, 10:45-12:30)
  1. Frequency and Bayes interpretations
  2. Interval estimation
  3. Systematic uncertainties
- Lecture III (Friday, September 20, 10:45-12:30)
  1. Hypothesis tests
  2. Resampling methods
- Lecture IV (Saturday, September 21, 10:45-12:30)
  1. Density estimation

Density estimation

1. Empirical density estimator
2. Histograms
3. Kernel estimation
4. Ideogram
5. Parametric vs non-paerametric
6. Optimization
7. Estimating errors
8. Curse of dimensionality
9. Naive Bayes
10. Orthogonal series
11. Monte Carlo models
12. Unfolding
13. sPlots

Density estimation

- Density estimation deals with the problem of estimating a pdf based on some data sampled from the pdf
- It may use assumed forms of the distribution, parameterized in some way (parametric statistics)
- Or it may avoid making assumptions about the form of the pdf (non-parametric statistics)
- We have discussed parametric statistics, now we are concerned more with the non-parametric case
  distinct concepts

We'll assume we have a dataset of iid observations (possibly vectors), $x_1, \ldots, x_N$ from pdf $f(X)$. The problem is to estimate f
Our estimator will be denoted $\hat{f}$. $\hat{f}(X)$ is a RV

Motivation

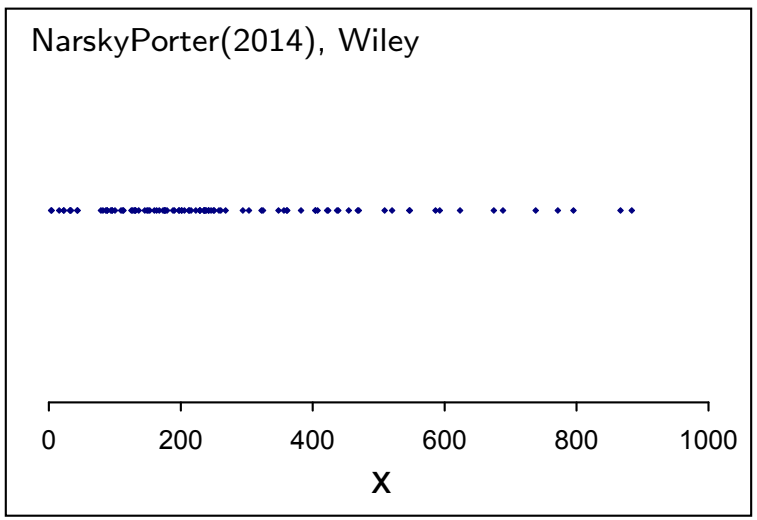
- Why non-parametric estimates?
- Maybe don't have a parametric model
- Maybe can't do an analytic calculation, and must simulate our model
- May be easier/better than parametric modeling for efficiency corrections and background subtraction
- Visualization
- Smoothing
- Unfolding
- Comparing samples (eg, when simple moments may not capture enough complexity)

Empirical density estimator

A basic density estimate is the Empirical Probability Density Function (epdf). Place a delta function at each data point:

$$\hat{f}(X) = \frac{1}{N} \sum_{n=1}^{N} \delta(X - x_n)$$

X could be multi-dimensional here; the scatter plot presents a representation of a 2-dimensional epdf



The points are at ∞, with the "area" under a point equal to $1/N$. Here, $N = 100$, and the sampling distribution is a $\Delta(1232)$ Breit-Wigner (with pion and nucleon masses subtracted) on a 2^{nd} degree polynomial background. The background probability is 50%.

Have already seen the epdf as the sampling density for the bootstrap procedure

The Histogram

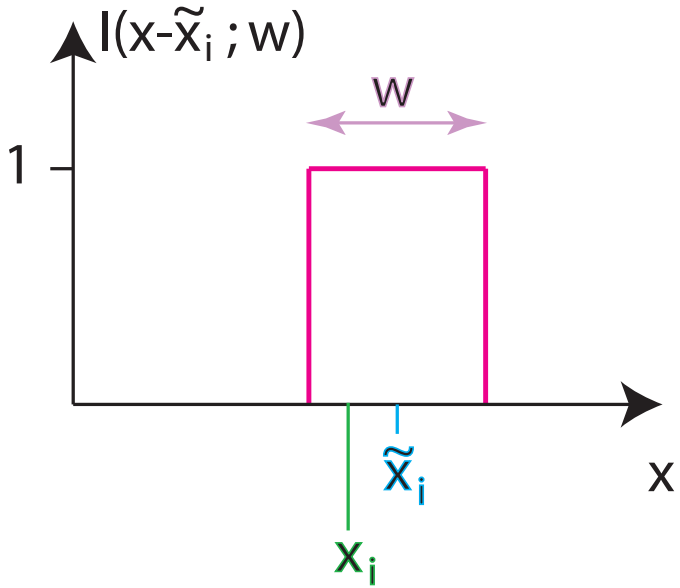
A familiar density estimator is based on the histogram:

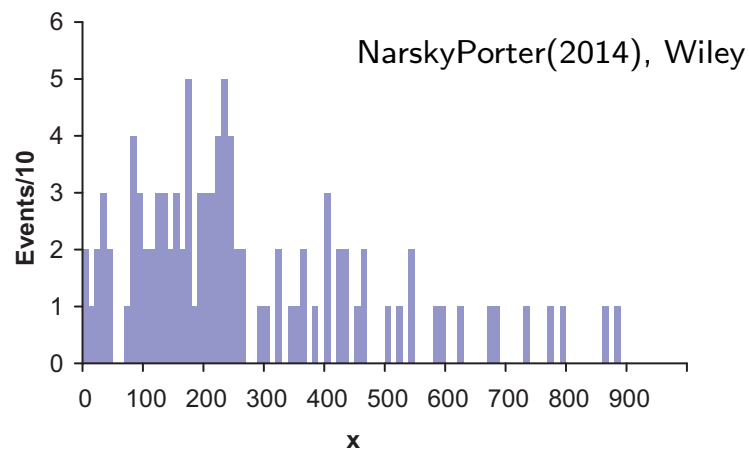
$$h(x) = \sum_{n=1}^{N} I(x - \tilde{x}_n; w),$$

where $\tilde{x}_n$ is the center of the bin in which observation x_n lies, w is the bin width, and

$$I(x; w) = \begin{cases} 1 & x \in [-w/2, w/2) \\ 0 & \text{otherwise} \end{cases}$$

This function is called the indicator function





Histogram with $w = 10$, for the sample above

The Histogram

- This is written for uniform bin widths, but may be generalized to differing widths with appropriate relative normalization factors
- Given a histogram, the estimator for the probability density function is:

$$\hat{f}(x) = \frac{1}{Nw} h(x)$$

Criticisms of Histogram as Density Estimator

- Discontinuous even if pdf is continuous
- Dependence on bin size and bin origin
- Information from location of datum within a bin is ignored

Kernel Estimation

Take the histogram, but replace indicator function I with something else:

$$\hat{f}(x) = \frac{1}{N} \sum_{n=1}^{N} k(x - x_n; w)$$

where $k(x, w)$ is the kernel function, normalized to unity:

$$\int_{-\infty}^{\infty} k(x; w)\, dx = 1$$

Usually interested in kernels of the form

$$k(x - x_i; w) = \frac{1}{w} K\left(\frac{x - x_i}{w}\right),$$

The kernel estimator for the pdf is then:

$$\hat{p}(x) = \frac{1}{nw} \sum_{i=1}^{n} K\left(\frac{x - x_i}{w}\right)$$

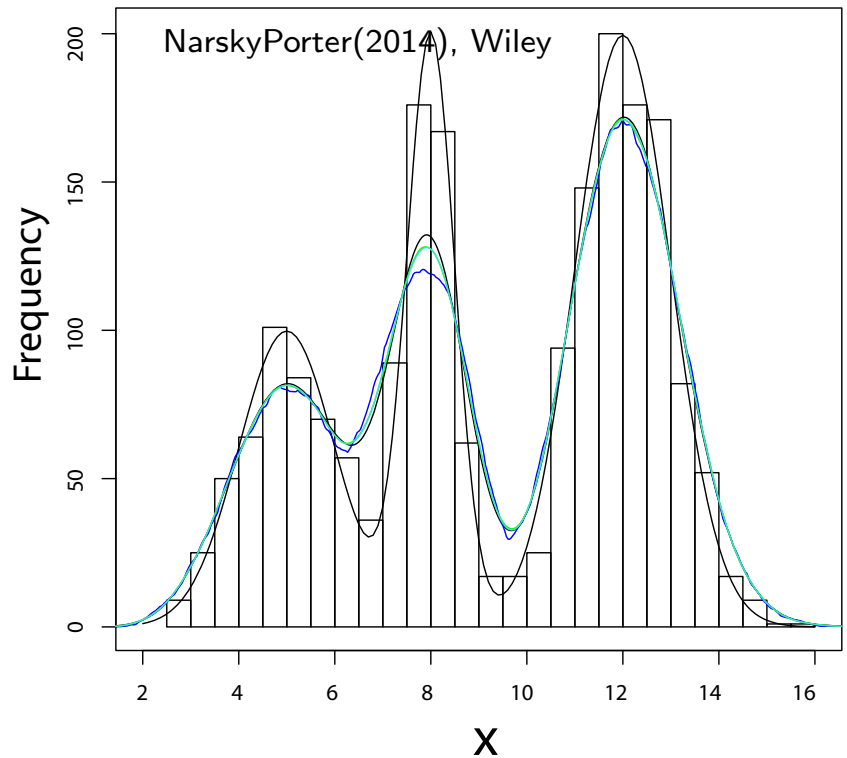
The role of parameter w as a smoothing parameter is evident

Kernel Estimation

- Often, the particular form of the kernel used doesn't matter very much. Illustrated below for several kernels (with commensurate smoothing parameters)
- Gaussian kernel is probably the most popular, and is smooth

Comparison of density estimates using different kernels

- Black (highest): sampling distribution
- Black (lower): estimate with Gaussian kernel
- Green: indistinguishable triangular and cosine kernel estimates
- Blue: rectangular kernel estimate



Multi-Variate Kernel Esitmation

Explicit multi-variate case, $d = 2$ dimensions:

$$\hat{f}(x,y) = \frac{1}{N w_x w_y} \sum_{n=1}^{N} K\left(\frac{x - x_n}{w_x}\right) K\left(\frac{y - y_n}{w_y}\right)$$

This is a product kernel form, with the same kernel in each dimension, except for possibly different smoothing parameters. It does not have correlations

The kernels we have introduced are classified more explicitly as fixed kernels: The smoothing parameter is independent of x

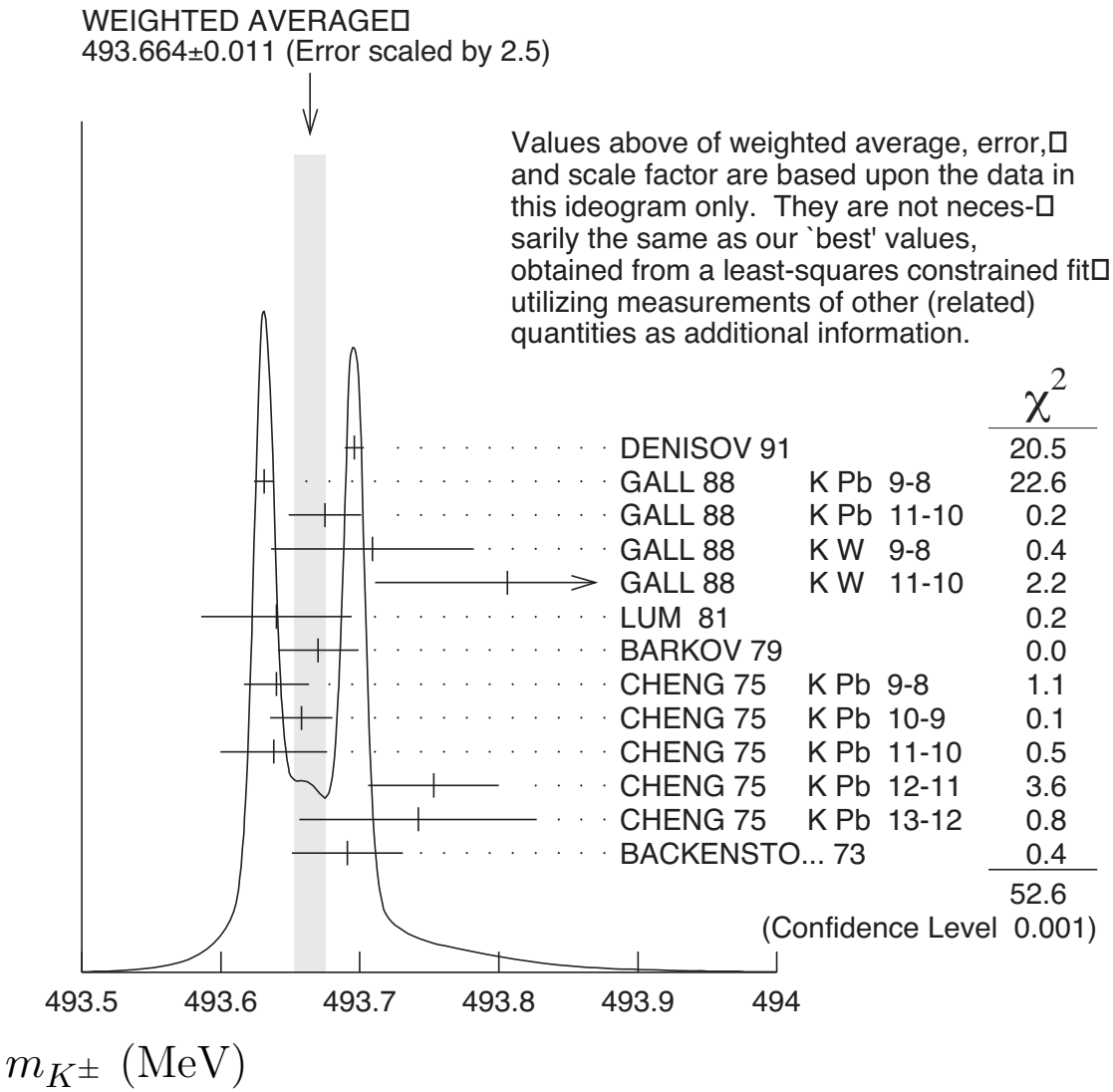
Ideogram

A simple variant on the kernel idea is to permit the kernel to depend on additional knowledge in the data.

- Physicists call this an ideogram
- Most common is the Gaussian ideogram, in which each data point is entered as a Gaussian of area one and standard deviation appropriate to that datum
- This addresses a way that the iid assumption might be broken

[Aside: Be careful to get your likelihood function right if you are incorporating variable resolution information in your fits; see, e.g., Punzi: `http://www.slac.stanford.edu/econf/C030908/papers/WELT002.pdf`]

Sample Ideogram



(from RPP 2006)

Parametric vs non-Parametric Density Estimation (I)

- Distinction is fuzzy
- A histogram is non-parametric, in the sense that no assumption about the form of the sampling distribution is made
- Often an implicit assumption that distribution is "smooth" on scale smaller than bin size. Eg, we know something about the resolution of our apparatus
- But the estimator of the parent distribution made with a histogram is parametric – the parameters are populations (or frequencies) in each bin. The estimators for those parameters are the observed histogram populations. Even more parameters than a typical parametric fit!

Parametric vs non-Parametric Density Estimation (II)

- Essence of difference may be captured in notions of "local" and "non-local":
  - If a datum at x_n influences the density estimator at some other point x this is non-local
  - A non-parametric estimator is one in which the influence of a point at x_n on the estimate at any x with distance$(x_n, x) > \epsilon$ vanishes, asymptotically
- Notice that for a kernel estimator, the bigger the smoothing parameter w, the more non-local the estimator,

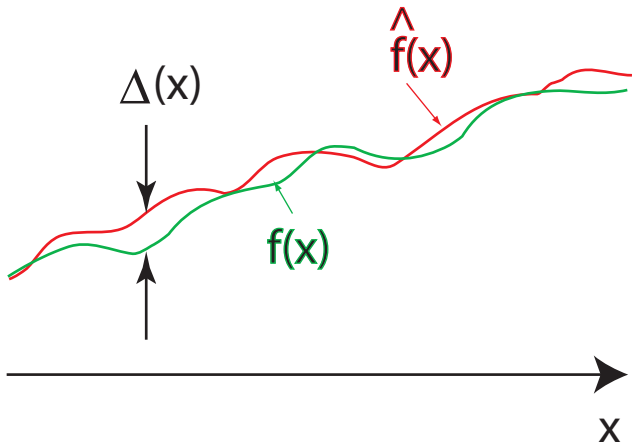
$$\hat{f}(x) = \frac{1}{Nw} \sum_{n=1}^{N} K\left(\frac{x - x_n}{w}\right)$$

- The "optimal" choice of smoothing parameter depends on N

Optimization

We would like to make an optimal density estimate from our data.

- What does that mean?
- Need a criterion for "optimal"
- Choice of criterion is subjective; it depends on what you want to achieve.
- We may compare the estimator for a quantity (here, value of the density at x) with the true value: $\Delta(x) = \hat{f}(x) - f(x)$



Mean Squared Error (I)

A common choice in parametric estimation is to minimize the sum of the squares. We may take this idea over here, and form the Mean Squared Error (MSE):

$$\mathrm{MSE}[\widehat{f}(x)] \equiv E\left\{\left[\hat{f}(x) - f(x)\right]^2\right\} = \mathrm{var}[\widehat{f}(x)] + \mathrm{bias}^2[\widehat{f}(x)]$$

where

$$\mathrm{var}[\widehat{f}(x)] \equiv E\left[\left(\widehat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

$$\mathrm{bias}[\widehat{f}(x)] \equiv E[\widehat{f}(x)] - f(x)$$

Mean Squared Error (II)

Since this isn't quite our familiar parameter estimation, let's take a little time to make sure it is understood:

Suppose $\hat{p}(x)$ is an estimator for the pdf $f(x)$, based on data $\{x_n; n = 1, \ldots, N\}$, iid from $f(x)$. Then

$$E[\hat{f}(x)] = \int \cdots \int \hat{f}(x; \{x_n\}) \, \text{Prob}\,(\{x_n\}) d^n(\{x_n\})$$

$$= \int \cdots \int \hat{f}(x; \{x_n\}) \prod_{n=1}^{N} [f(x_n) dx_i]$$

*Exercise: Proof of formula for the MSE

$$
\begin{aligned}
\mathrm{MSE}[\hat{f}(x)] &= E\left[(\hat{f}(x) - f(x))^2\right] \\
&= \int \cdots \int \left[\hat{f}(x;\{x_i\}) - f(x)\right]^2 \prod_{n=1}^{N} [f(x_n)dx_n] \\
&= \int \cdots \int \left[\hat{f}(x;\{x_i\}) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x)\right]^2 \prod_{n=1}^{N} [f(x_n)dx_n] \\
&= \int \cdots \int \Big\{ \left[\hat{f}(x;\{x_i\}) - E(\hat{f}(x))\right]^2 + \left[E(\hat{f}(x)) - f(x)\right]^2 \\
&\qquad -2\left[\hat{f}(x;\{x_i\}) - E(\hat{f}(x))\right]\left[E(\hat{f}(x)) - f(x)\right] \Big\} \prod_{n=1}^{N} [f(x_n)dx_n] \\
&= \mathrm{var}[\hat{f}(x)] + \mathrm{bias}^2[\hat{f}(x)] + 0
\end{aligned}
$$

[In typical treatments of parametric statistics, we assume unbiased estimators. That isn't a good assumption here.]

The Problem With Smoothing (I)

Theorem: [Rosenblatt (1956)] A uniform minimum variance unbiased estimator for $f(x)$ does not exist.

- Unbiased:

$$E[\hat{f}(x)] = f(x)$$

- Uniform minimum variance:

$$\text{var}\left[\hat{f}(x)|f(x)\right] \leq \text{var}\left[\hat{g}(x)|f(x)\right], \ \forall x,$$

for all $f(x)$, where $\hat{g}(x)$ is any other estimator of $f(x)$

The Problem With Smoothing (II)

For example, suppose we have a kernel estimator:

$$\hat{f}(x) = \frac{1}{N} \sum_{n=1}^{N} k(x - x_n; w)$$

Its expectation is:

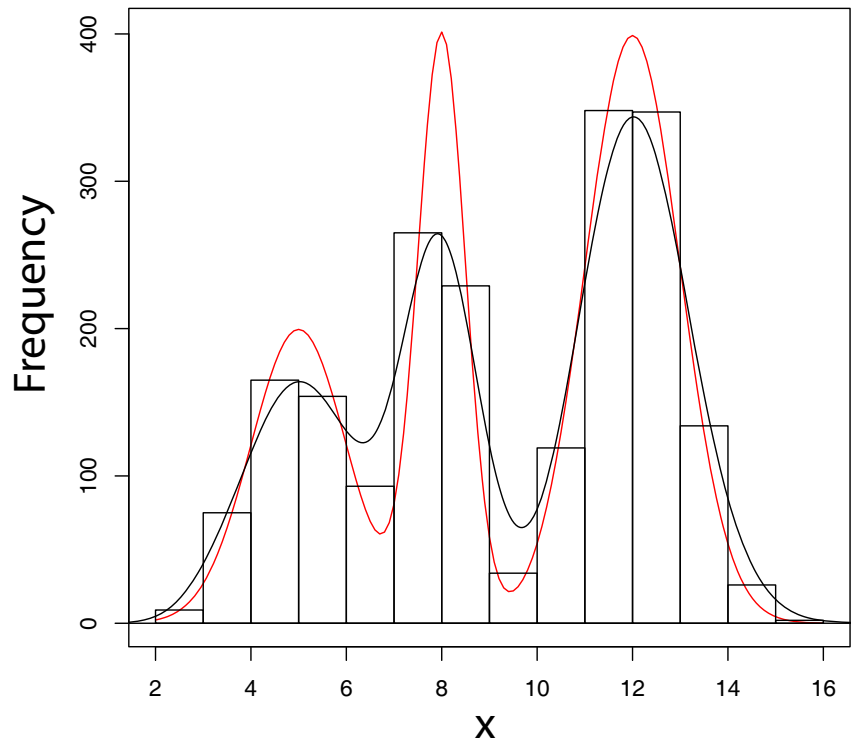
$$E[\hat{f}(x)] = \frac{1}{N} \sum_{n=1}^{N} \int k(x - x_n; w) f(x_n) dx_n$$
$$= \int k(x - y) f(y) dy$$

Unless $k(x - y) = \delta(x - y)$, $\hat{f}(x)$ will be biased for some $f(x)$.
But $\delta(x - y)$ has infinite variance

The Problem with Smoothing (III)

So the nice properties we strive for in parameter estimation (and sometimes achieve) are beyond reach.
Intuition: smoothing lowers peaks and fills in valleys.

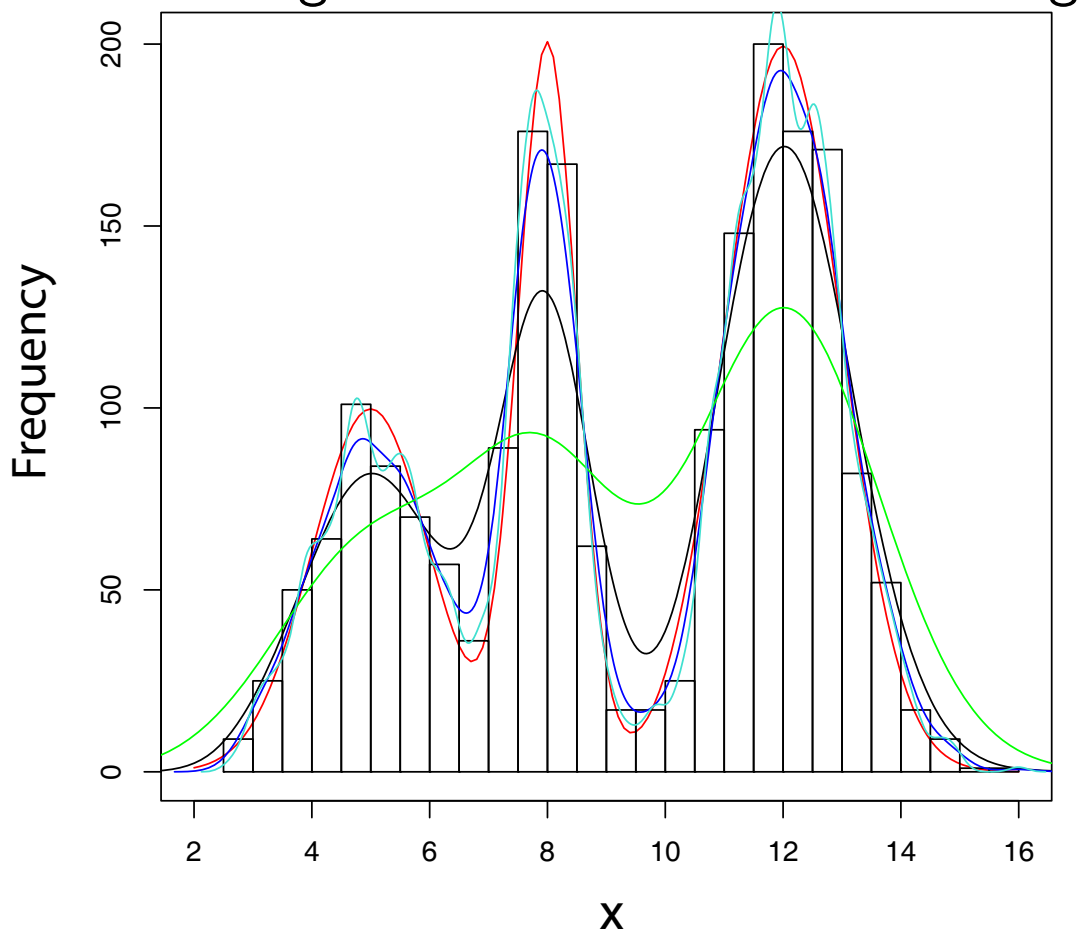
- Red curve: pdf
- Histogram for a sampling from pdf
- Black curve: Gaussian kernel estimator for pdf using the same sampling



NarskyPorter(2014), Wiley

Dependence on Smoothing Parameter

Plot showing effect of choice of smoothing parameter:



Red: Sampling PDF
Black: Default smoothing (w)
Blue: w/2 smoothing
Turquoise: w/4 smoothing
Green: 2w smoothing

NarskyPorter(2014), Wiley

More on Optimization

- The MSE for a density is a measure of uncertainty at a point
- Useful to summarize the uncertainty over all points, a notion for the "distance" from function $\hat{f}(x)$ to function $f(x)$
- A convenient choice is the Integrated Squared Error (ISE):

$$\text{ISE} \equiv \int \left[\hat{f}(x) - f(x)\right]^2 dx$$

- The ISE depends on the true density, the estimator, and the sampled data
- Remove this latter dependence by evaluating the Mean Integrated Squared Error (MISE), or equivalently, the integrated mean square error (IMSE):

$$\begin{aligned}\text{MISE} &\equiv E\left[\text{ISE}\right] = E\left[\int \left[\hat{f}(x) - f(x)\right]^2 dx\right] \\ &= \int E\left[\left(\hat{f}(x) - f(x)\right)^2\right] dx = \int \text{MSE}\left[\hat{f}(x)\right] dx \equiv \text{IMSE}\end{aligned}$$

Optimization – Consistency

A desirable property of an estimator is that the error decreases as the number of samples increases. This is a familiar notion from parametric statistics

Definition

A density estimator $\hat{f}(x)$ is consistent iff:

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

as $N \to \infty$

*Choosing Histogram Binning

Considerations such as minimizing the MSE may be used to choose an "optimal" bin width for a histogram

Theorem: The MSE of the histogram estimator is consistent if the bin width $w \to 0$ as $N \to \infty$ such that $Nw \to \infty$

- 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.
- Arguments exist that optimal bin size should decrease as $1/N^{1/3}$
- A popular choice is Sturges' rule, which says that the number of bins should be

$$k = 1 + \log_2 N$$

  It is the default choice when making a histogram in R.
- Another popular choice is Scott's rule for the bin width is:

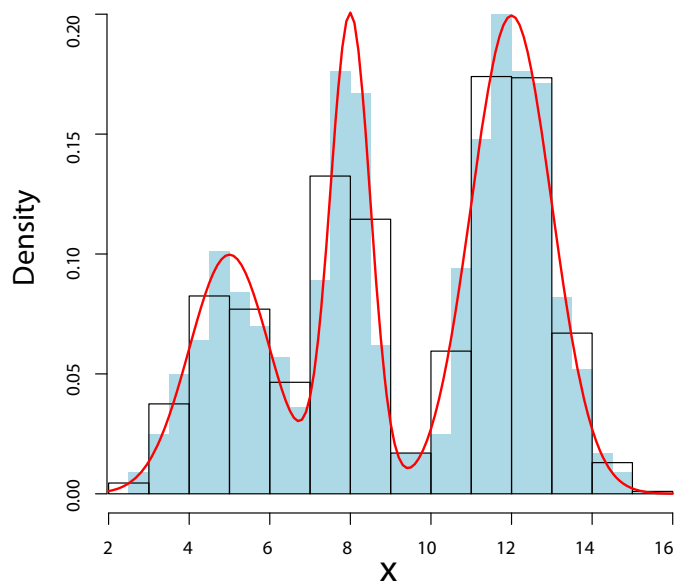
$$w = 3.5 s N^{-1/3},$$

  where s is the sample standard deviation

Choosing Histogram Binning

- These rules often leave the impression that the binning could usefully be finer. If data are unimodal, then the rules may reasonably apply. If not then a more adaptive approach is required to obtain optimal results. If additional information is available, eg, the experimental resolution, this can help to inform the bin width choice

- Curve is the sampling pdf
- The "standard rules" (Sturges, Scott, Freedman-Diaconis) correspond roughly to the coarser binning above
- The shaded histogram seems like a better choice, illustrating that blind application of the rules to complicated data may not yield desired result

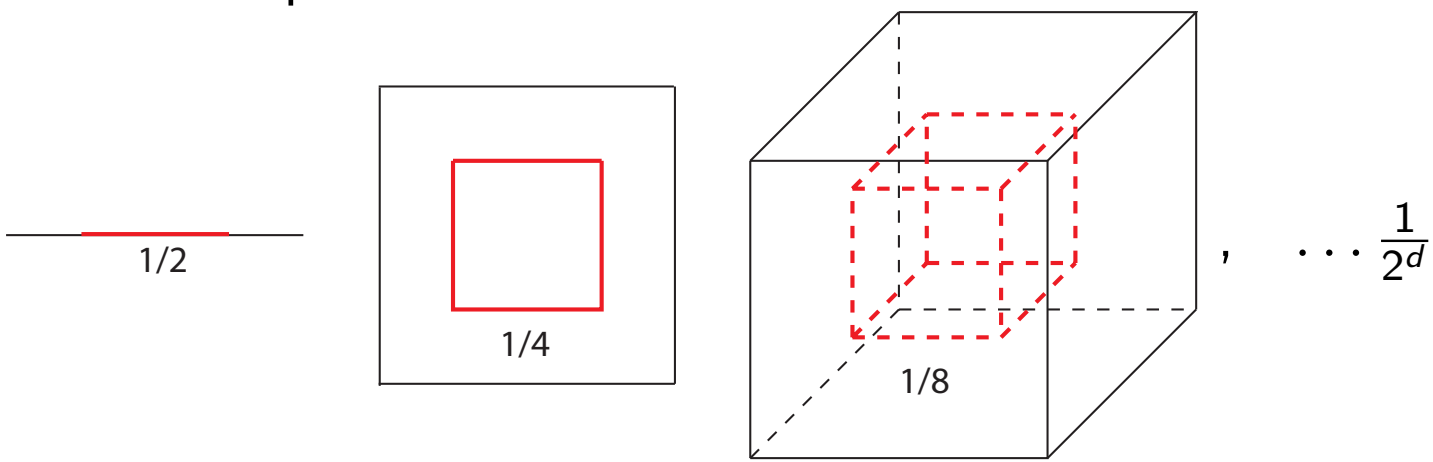


NarskyPorter(2014), Wiley

The Curse of Dimensionality

The The Curse of Dimensionality is a significant affliction in density estimation

- Difficult to display and visualize as the number of dimensions increases.
- "All" the volume (of a bounded region) goes to the boundary (exponentially!) as the dimensions increases. I.e., data becomes "sparse"



- Tendency for exponentially growing computation requirement with dimensions
- Even worse than parametric statistics

Summary

We have introduced:

- Basic notions in (non-parametric) density estimation
- Some simple variations on the theme
- A foundation towards optimization
- An idea of where and how things will fail

Next: Further sophistication on these ideas; and introduction of other variations in approach and application

Estimating errors

The bootstrap provides a way to evaluate how good is our density estimate

The bootstrap algorithm in this context is as follows:

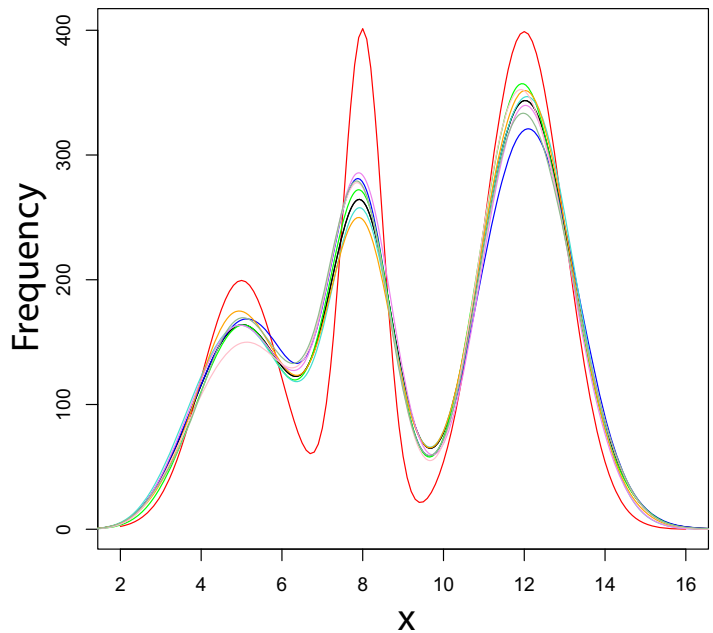
1. Form density estimate $\hat{f}$ from data $\{x_n; n = 1, \ldots, N\}$
2. Resample (uniformly) N values from $\{x_n; n = 1, \ldots, N\}$, with replacement, obtaining $\{x_n^*; n = 1, \ldots, N\}$ (bootstrap replica)
3. Form density estimate $\hat{f}^*$ from replica $\{x_n^*; n = 1, \ldots, N\}$
4.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

Estimating errors – variance

- Consider, for a kernel density estimator, the expectation of the boostrap dataset (exercise):

$$E\left[\hat{f}^*(x)\right] = E\left[K(x - x_n^*; w)\right] = \hat{f}(x)$$

- The bootstrap distribution about $\hat{f}$ does not reproduce the bias which may be present in $\hat{f}$ about f. It does reproduce the variance of $\hat{f}$, hence the bootstrap is a useful tool for estimating the variance of a density estimator



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

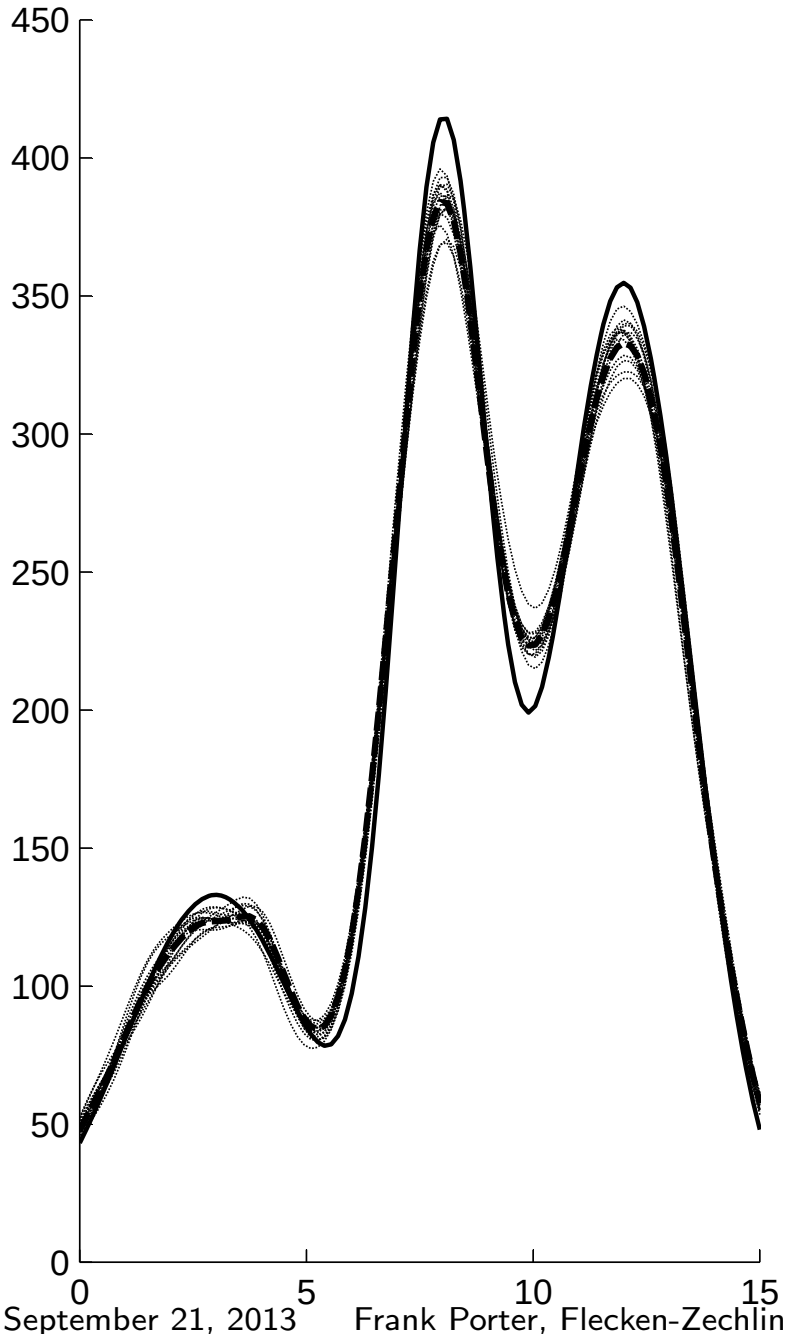
Estimating errors – bias

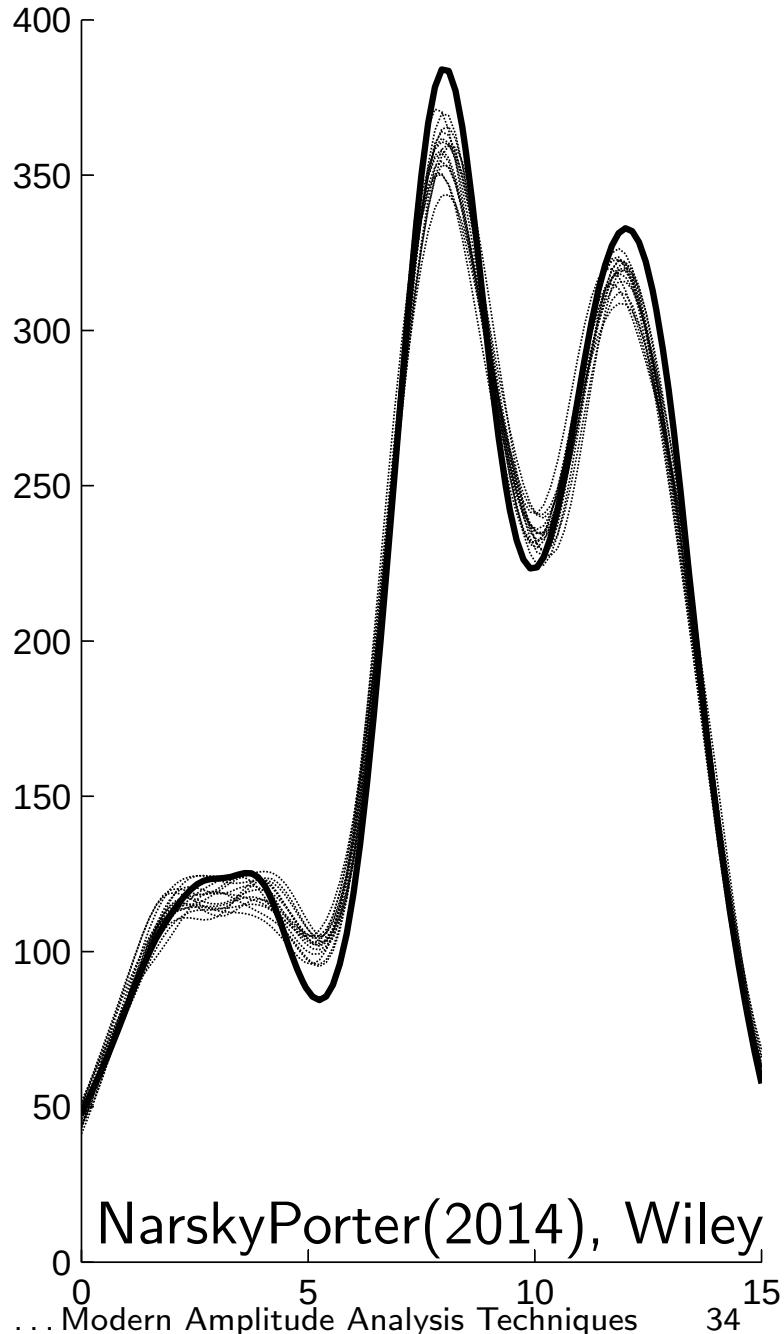
- We use a variation of the bootstrap, called the smoothed bootstrap, to obtain an estimate for the bias
- In this case, the replicas are sampled from the (kernel) estimate for the density, instead of from the empirical density
- Denote the kernel estimate by $\hat{f}_w(x)$, where w indicates the dependence on the smoothing parameter
- Suppose we draw a smoothed bootstrap replica x^* from this distribution
- We can make a kernel density estimate from this replica, $\hat{f}^*_w(x)$
- Now the bias of $\hat{f}^*_w(x)$ compared with $\hat{f}_w(x)$ will mimic the bias of $\hat{f}_w(x)$ compared with $f(x)$. Thus, using the smoothed bootstrap we may estimate the full MSE

Estimating errors – bias

- On the next slide: Use of the bootstrap to determine the variance and bias of a Gaussian kernel density estimator. The sample is size $N = 1000$
  - Left: Solid curve shows the sampling distribution; heavy dashed curve shows a kernel estimator; lighter dotted curves show 15 bootstrap replica kernel estimators
  - Right: Solid curve is the kernel density estimator from the dashed curve on the left plot; lighter dotted curves show 15 smoothed bootstrap replica kernel estimators
- The bias of the smoothed bootstrap replicas about the kernel density estimator mimics the bias of the kernel estimator compared with the true distribution
- Can use this to correct for bias
- Could further construct confidence intervals (e.g., using the percentile method) including the effect of bias
- The smoothed bootstrap introduces some extra variance, which may be corrected for

Estimating errors – bias

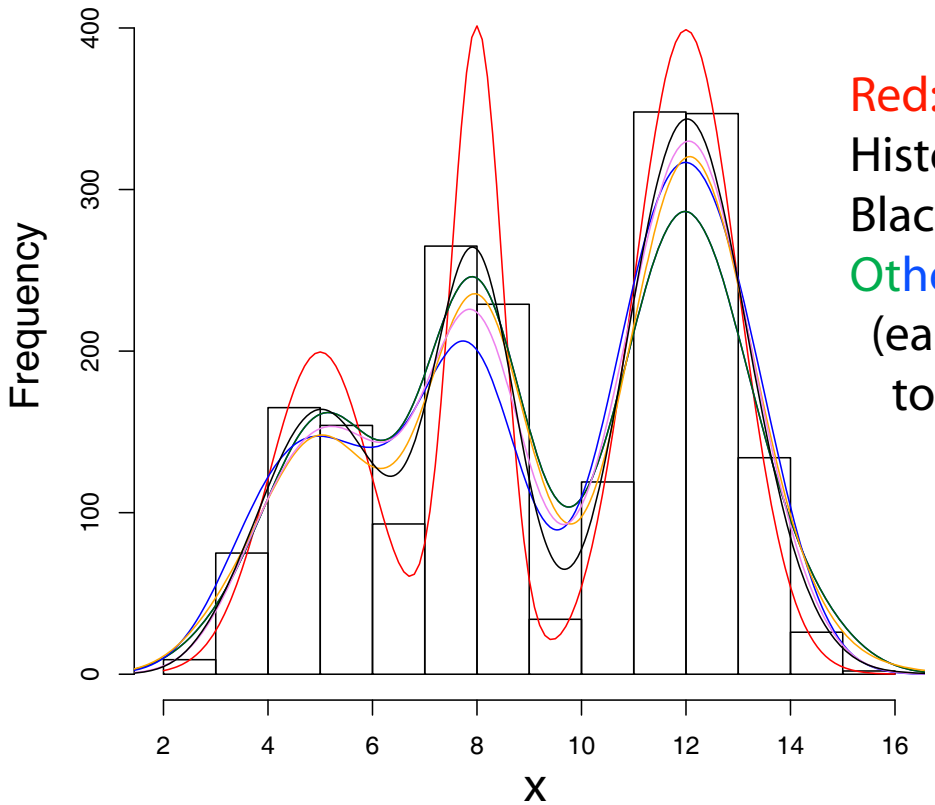




Estimating errors – bias

- Besides the bootstrap, the jackknife and cross-validation may also be used to estimate errors (and hence smoothing optimization) in kernel estimation
- Eg, we may use the jackknife to estimate bias.
- The idea is that bias depends on sample size. If we can assume that the bias vanishes asymptotically, we may use the data to estimate the dependence of the bias on sample size.
- We use a delete-d version here, with $k = N/d$. A simple algorithm is:
  1. Divide the data into k random disjoint subsamples.
  2. Evaluate the estimator for each subsample.
  3. Compare the average of the estimates on the subsamples with the estimator based on the full dataset.

Estimating errors – bias



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)

NarskyPorter(2014), Wiley

Jackknife estimation of bias

Adaptive Kernel Estimation

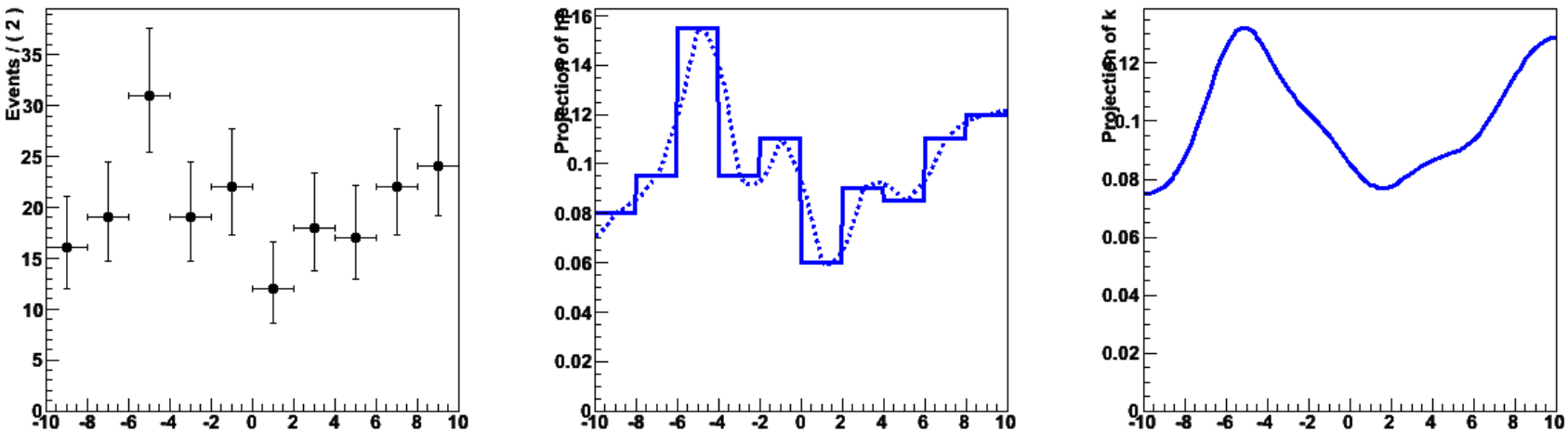
- We saw that it is probably more optimal to use variable bin widths in histograms
- This applies also to other kernels
- Indeed, the use of a fixed smoothing parameter, deduced from all of the data introduces a non-local, hence parametric, aspect into the estimation
- More consistent to look for smoothing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

Adaptive Kernel Estimation

- We could derive algorithms based on optimizing MISE to carry this out.
- Generally begin with initial fixed kernel estimate to get a starting density estimate, then iterate as needed
- E.g., for Poisson statistics, adaptive smoothing parameter $w(x) \sim \sqrt{\sigma/f(x)}N^{-1/5}$ where the power of N is the value that minimizes MISE for a Gaussian kernel and normal sampling (Taylor, *Biometrika* **76** (1989) 705)
- There are packages for adaptive kernel estimation, eg, KEYS (Kernel Estimating Your Shapes) (Cranmer, Comp.Phys.Comm., **136** (2000) 198)

Adaptive Kernel Estimation

Example of the KEYS adaptive kernel estimation



- Left: Input data
- Middle: Histogram-based estimate with second order interpolation
- Right: KEYS adaptive kernel estimate

(RooFit manual, Verkerke and Kirkby)

Multivariate Kernel Estimation

- Multi-dimensional case introduces issue of covariance
- With a product kernel, the local estimator has diagonal covariance matrix
- Could apply a local linear transformation of the data to a coordinate system with diagonal covariance matrices. Amounts to a non-linear transfomation of the data in a global sense; may not be easy
- But we can try the system for which the overall covariance matrix of the data is diagonal

Multivariate Kernel Estimation

- If $\{y_n\}_{n=1}^N$ is the diagonalized data, the product fixed kernel estimator in D dimensions is:

$$\hat{f}_0(y) = \frac{1}{N}\sum_{n=1}^{N}\left[\prod_{d=1}^{D}\frac{1}{w_d}K\left(\frac{y^{(d)} - y_n^{(d)}}{w_d}\right)\right],$$

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

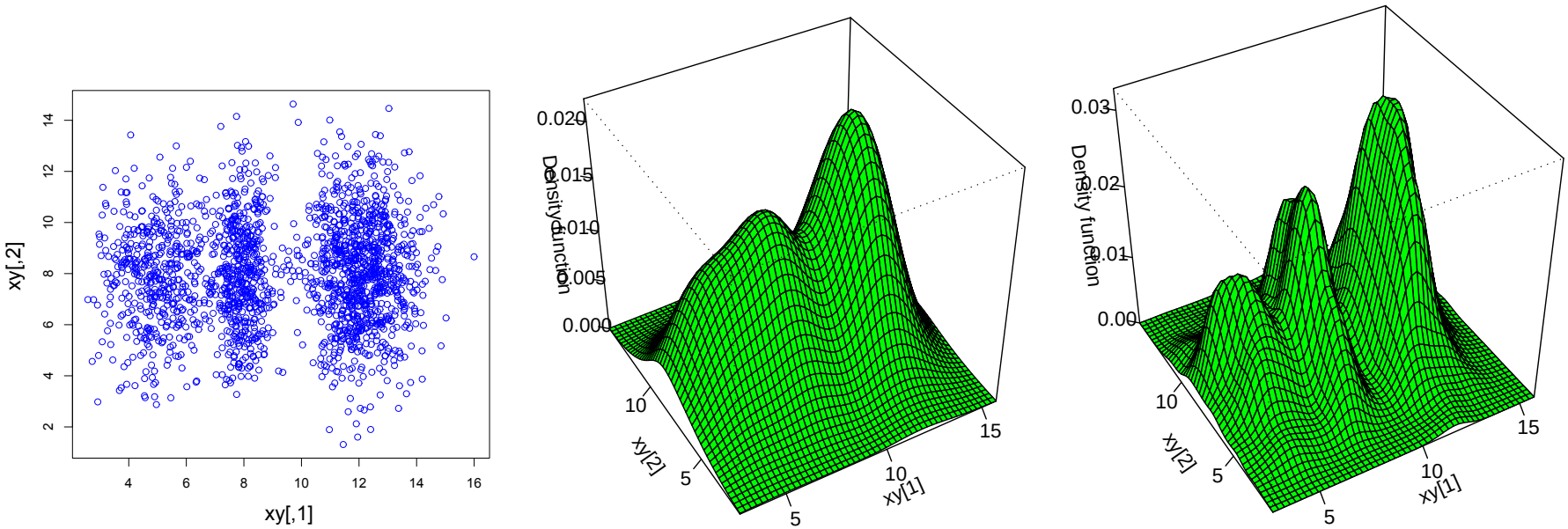
- The asymptotic, normal MISE-optimized smoothing parameters are now:

$$w_d = \left(\frac{4}{D+2}\right)^{1/(D+4)} \sigma_d N^{-1/(D+4)}$$

- Corresponding adaptive kernel estimator follows the discussion as for the univariate case

Multivariate Kernel Estimation

Example where the sampling distribution has diagonal covariance matrix (locally and globally) NarskyPorter(2014), Wiley

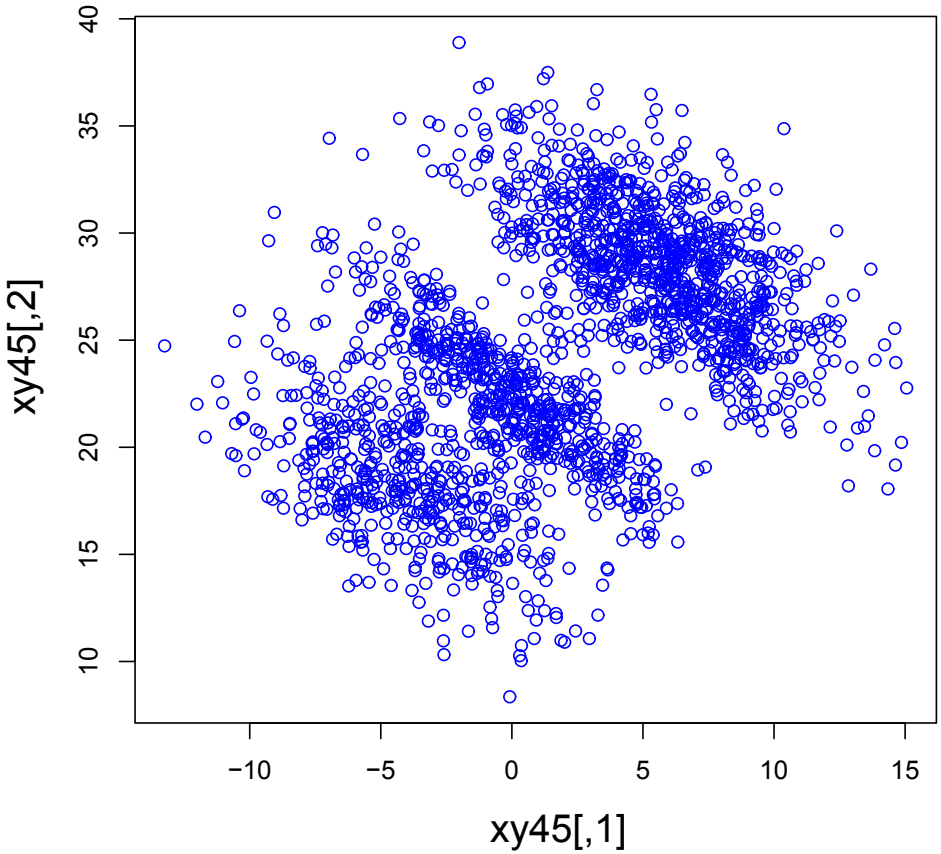


- Left: 2-D distribution with diagonal covariance matrix
- Middle: application of kernel estimation to this distribution
- Right: Same as middle, except using one-half the default smoothing parameter

Multivariate Kernel Estimation

Example with non-diagonal covariance matrix (above example rotated 45°)

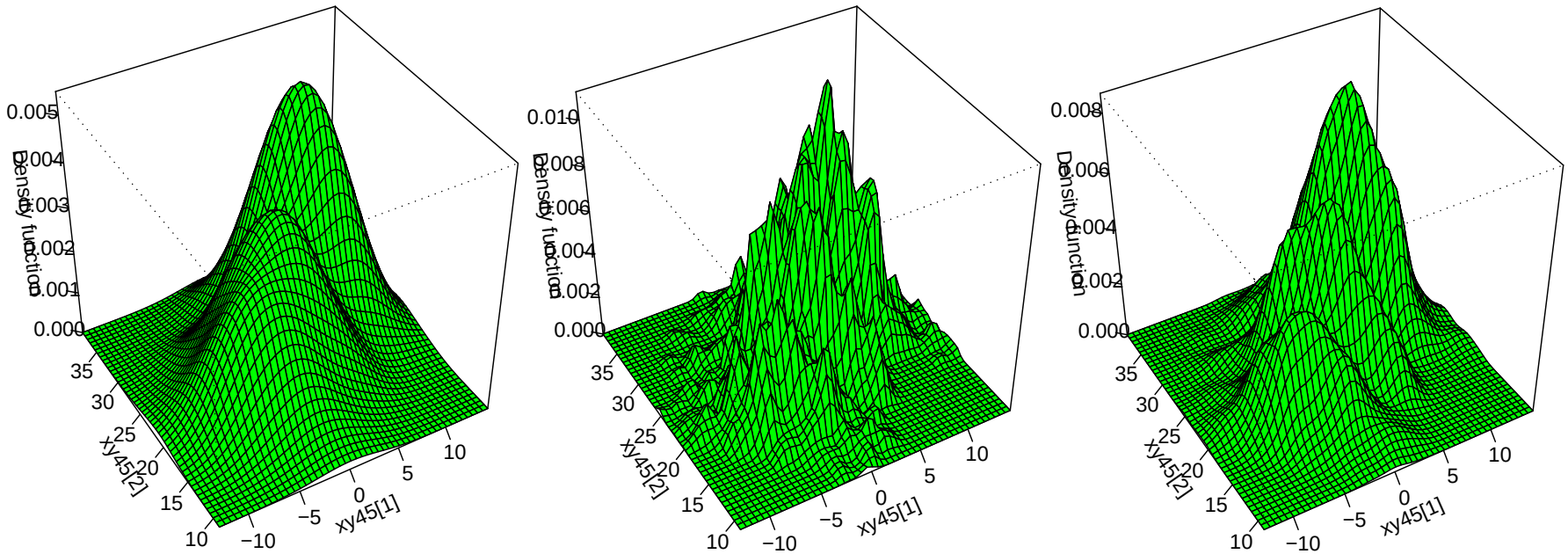
NarskyPorter(2014), Wiley



We see on next slide that this is more difficult to handle with our product kernel technology

Multivariate Kernel Estimation

Kernel estimation applied to the 2-D data on the previous slide



NarskyPorter(2014), Wiley

- Left: Default smoothing parameter
- Middle: Using one-half of the default smoothing parameter
- Right: Intermediate smoothing

Estimation Using Orthogonal Series

Alternative approach: expand the pdf in a series of orthonormal functions:

$$f(x) = \sum_{k=0}^{\infty} a_k \psi_k(x),$$

where

$$a_k = \int \psi_k(x) f(x) \rho(x)\, dx = E\left[\psi_k(x)\rho(x)\right],$$

$$\int \psi_k(x)\psi_\ell(x)\rho(x)\, dx = \delta_{k\ell},$$

$\rho(x) > 0$ is a weight function

Estimation Using Orthogonal Series

- Expansion coefficients are expectation values of functions; Natural to substitute sample averages as estimators (substitution method!). This corresponds to using the empirical probability distribution:

$$\begin{aligned}\hat{a}_k &= \int \psi_k(x)\hat{f}(x)\rho(x)\,dx \\ &= \int \psi_k(x)\frac{1}{N}\sum_{n=1}^{N}\delta(x-x_n)\rho(x)\,dx \\ &= \frac{1}{N}\sum_{n=1}^{N}\psi_k(x_n)\rho(x_n)\end{aligned}$$

- Thus:

$$\hat{f}(x) = \sum_{k=1}^{m}\hat{a}_k\psi_k(x)$$

Number of terms m is chosen by some optimization criterion

Estimation Using Orthogonal Series

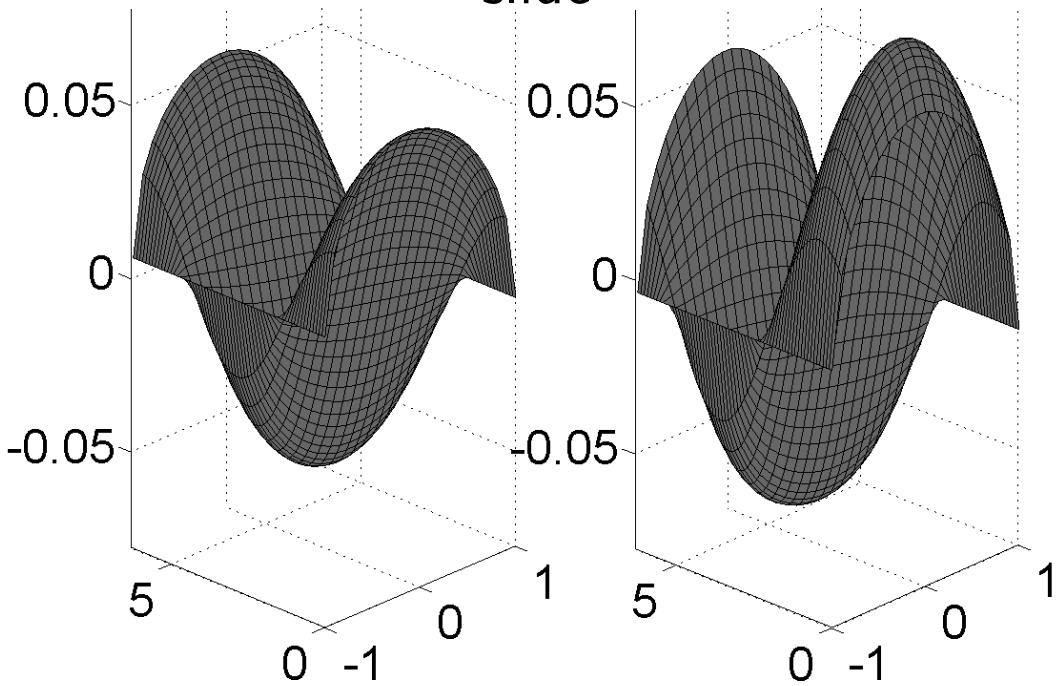
- Analogy between m and smoothing parameter w in kernel estimators; and between K and $\{\psi_k\}$
- Often applied with angular distributions with Legendre polynomials or spherical harmonics
- We may try an example in a two-dimensional sampling space. A dataset of size $N = 1000$ is generated according to density:

$$f(\cos\theta, \phi) = \frac{1}{4\pi}\left(1 + \frac{1}{2}\cos\theta + \frac{1}{2}\sin\theta\cos\phi\right).$$

- Use a series of real linear cominations of $Y_{\ell m}$ spherical harmonics to fit this data, according to the above substitution method
- Since we know the true distribution, we may compute the error in our density estimate
- We show result on next slide

Estimation Using Orthogonal Series

Errors from fits to sample dataset generated according to previous slide



NarskyPorter(2014), Wiley

- Left: Errors from fit with exactly the same angular functions in the series as used to generate the data
- Right: Error for a series that has extra terms $\sin\theta \sin\phi$ and $\cos^2\theta$ (in the form of suitable additional orthonormal $Y_{\ell m}$'s)

Estimation Using Orthogonal Series

- The errors become more serious when we add the terms
- We fit the empirical distribution more accurately with the additional terms, but we do a poorer job of estimating the actual distribution
- Extreme is the limit of keeping infinite terms in our expansion; we recover the epdf

Using Monte Carlo Models

- Often build a model using MC computations of different processes, adding together to get the complete model
- May involve weighting of events, if more data is simulated for some processes
- Overall simulated epdf is (x_n is $x_n^{(d)}, d = 1 \ldots D$):

$$\hat{f}(x) = \sum_{n=1}^{N} \rho_n \delta(x - x_n),$$

where $\sum \rho_n = 1$ (or N for an event sample of some total size).

- The weights must be included in computing the sample covariance matrix:

$$\Sigma_{k\ell} = \sum_{n=1}^{N} \rho_n \frac{(x_n^{(k)} - \hat{\mu}_k)(x_n^{(\ell)} - \hat{\mu}_\ell)}{\sum_j \rho_j},$$

where $\hat{\mu}_d = \sum_n \rho_n x_n^{(d)} / \sum_j \rho_j$ is the sample mean in dimension d

Using Monte Carlo Models

- The simulated data is discrete; may use kernel smoothing to obtain a continuous model
- Assuming we have transformed to a diagonal system using the sample covariance matrix, the product kernel density based on our simulation is then:

$$\hat{f}_0(x) = \frac{1}{\sum_j \rho_j} \sum_{n=1}^{N} \rho_n \prod_{d=1}^{D} \frac{1}{w_d} K\left(\frac{x^{(d)} - x_n^{(d)}}{w_d}\right)$$

- May be iterated to obtain an adaptive kernel estimator

Unfolding

- May not be satisfied with estimating the density from which our sample X was drawn
- Interesting physics may be obscured (smeared) by transformation with uninteresting functions, eg, efficiency dependencies, resolution, classification errors, etc.
- Similar to reconstructing an image from blurred photo
- We refer to the problem as one of unfolding the interesting distribution from the sampled (i.e., smeared) distribution
- The term deconvolution is also used. Strictly, "unfolding" is more general, not restricted to convolution smearing
- Theoretical reference Meister, *Deconvolution problems in Nonparametric Statistics* (2009) Springer
- Image reconstruction in astrophysics: Puetter et al., Ann.Rev.Astron.Astrophys. **43** (2005) 139
- There are pitfalls to unfolding, so if all you really want is a comparison of data with theory, it is better to smear the theory than to unfold the data

Unfolding

- Suppose there is a "fundamental" distribution, $f(x)$, and a transformation K mapping this distribution to the "experimental" distribution, $e(x)$:

$$e(x) = \int K(x,y) f(y) dy$$

- We sample from $e(x)$, but want to learn about $f(x)$
- If $f(x)$ were known up to parameters, we would address this problem in the context of parameteric statistics
- Instead, consider the context of non-parametric density estimation
- Assume that transformation $K(x,y)$ is known. It is called the point spread function because it gives the density at x from a point density "source" at y. [Often, K is estimated from auxillary data, and we must consider the uncertainties ("systematic errors") introduced]

Unfolding

- When we sample from e, we obtain an estimator $\hat{e}$ for e. Eg, $\hat{e}$ could be the empirical density estimator
- In principle, the estimation of f from this is easy:

$$\hat{f}(y) = \int K^{-1}(y,x)\hat{e}(x)dx,$$

where

$$\int K^{-1}(x,y)K(y,x')\,dy = \delta(x-x')$$

- If $\hat{e}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x-x_n)$ is the empirical distribution, then:

$$\hat{f}(y) = \frac{1}{N}\sum_{n=1}^{N} K^{-1}(y,x_n)$$

*Unfolding

- If we don't know how to invert K, we may try an iterative (e.g., Neumann series) solution
- Eg, consider the problem of unfolding radiative corrections in a cross section measurement.
- The observed cross section, $\sigma_E(s)$, is related to the "interesting" cross section $\sigma(s)$ according to:

$$\sigma_E(s) = \sigma(s) + \delta\sigma(s),$$

where

$$\delta\sigma(s) = \int K(s,s')\sigma(s')\,ds'$$

- Form iterative estimate for $\sigma(s)$ according to:

$$\begin{aligned} \hat{\sigma}_0(s) &= \sigma_E(s) \\ \hat{\sigma}_i(s) &= \sigma_E(s) - \int K(s,s')\hat{\sigma}_{i-1}(s')\,ds', \quad i = 1,2,\ldots \end{aligned}$$

Unfolding

- Important case when measurement results in addition of a stochastic error to a statistical sampling from the fundamental distribution
- Then concerned with the addition of two independent RVs
- With fundamental distribution f and resolution density g, the distribution of the sum of the two contributions is a convolution:

$$e(x) = \int g(x-y)f(y)dy$$

- This suggests working in Fourier transform space, since then the convolution simplifies to a product:

$$e_{\rm FT}(t) = g_{\rm FT}(t)f_{\rm FT}(t)$$

where, eg,

$$f_{\rm FT}(t) = \int e^{it\cdot x} f(x)dx$$

is the Fourier transfom of $f(x)$

Unfolding

- To implement such an approach, we could use data to form an empirical estimate $\hat{e}_{FT}$ of e_{FT}:

$$\hat{e}_{FT}(t) = \int e^{it\cdot x}\hat{e}(x)dx = \frac{1}{N}\sum_{n=1}^{N} e^{it\cdot x_n}$$

- We then divide out g_{FT} to obtain estimator $\hat{f}_{FT}$. Then we take the inverse Fourier transfom to get $\hat{f}$
- A difficulty emerges in this last step – we may need to do something (regularization) to ensure the inverse transform exists and is not erratic (i.e., sensitive to small changes in the data). We look at an example of this sort of problem below

Unfolding

- We may equally well apply unfolding to a histogram with Poisson-distributed bin contents $x = x_1, \ldots, x_b$. Suppose the fundamental distribution is

$$f(x) = \prod_{i=1}^{b} \frac{\mu_i^{x_i} e^{-\mu_i}}{x_i!}, \quad x_i = 0, 1, \ldots$$

and the experimental distribution is

$$e(x) = \prod_{i=1}^{b} \frac{\nu_i^{x_i} e^{-\nu_i}}{x_i!}, \quad x_i = 0, 1, \ldots,$$

where the ν_i are related to μ by some transformation.

- Given a sampling x, the empirical (ML) distribution is

$$\hat{e}(y) = \prod_{i=1}^{b} \frac{x_i^{y_i} e^{-x_i}}{y_i!}, \quad y_i = 0, 1, \ldots$$

Unfolding

- Now suppose that the transformation is of the form:

$$\nu = A\mu + B$$

- Since A is a matrix, this includes the possibility that events get assigned to the wrong bin, as well as possible inefficiencies
- The B term allows for possible "background" contributions
- Our estimate for ν is x. Thus, assuming A is invertible, we estimate the unfolded distribution according to:

$$\hat{\mu} = A^{-1}(\hat{\nu} - B)$$

[In practice we solve with an appropriate numerical algorithm]

- This provides an unbiased estimator for μ, in fact the MLE, which is minimum variance (among unbiased estimators) since our distribution is in the exponential family

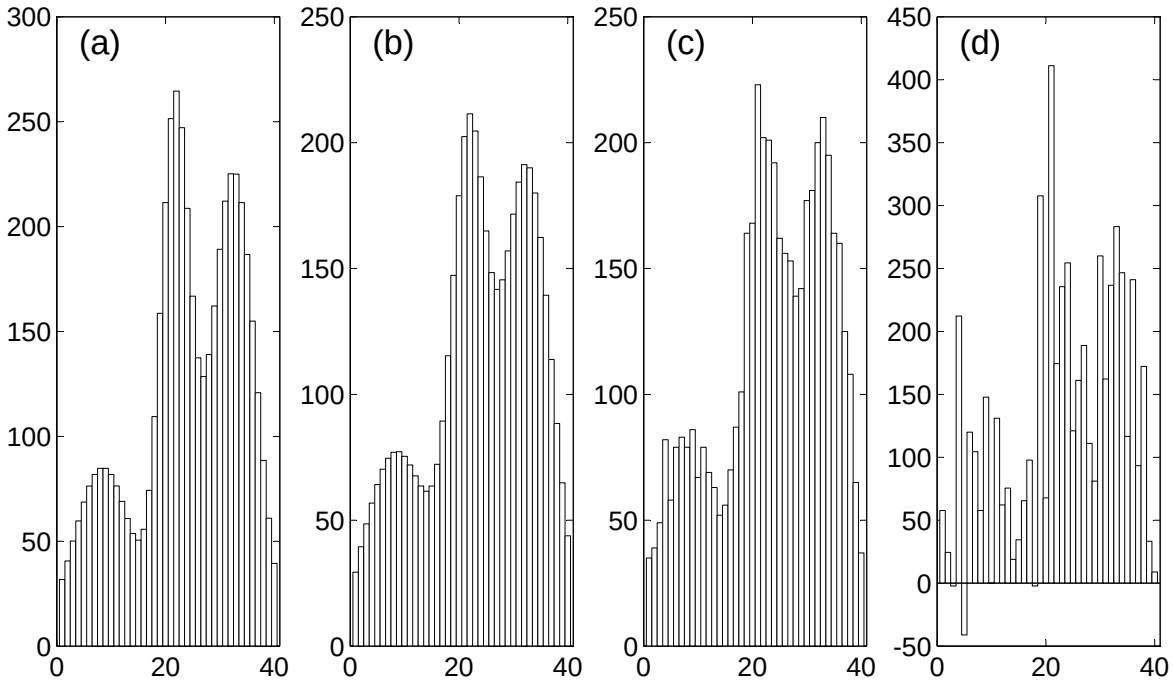
Unfolding – example

- Example: fundamental distribution with three Gaussian peaks, and a total expected sample size of 5000
- The matrix A is taken to be:

$$A_{ij} = 0.3/2^{|i-j|},$$

  mimicing a situation with an overall inefficiency plus migration among bins, where the migration probability decreases as bins are further apart
- The "background" vector B is taken to be a vector with the number five in each position

Unfolding – example



NarskyPorter(2014), Wiley

- Illustration of unfolding a histogram:
  - (a) The fundamental expected bin contents
  - (b) The expected bin contents after passing through the detector and analysis
  - (c) A sampling from (b), according to Poisson statistics. This is our MLE for (b)
  - (d) The unfolded distribution using the sampling in (c). This is our estimate for the fundamental distribution

Unfolding – example

- Plot (d) replicates the basic features of (a), but the fluctuations may be disconcerting. This is a minimum variance (among unbiased estimators) unbiased estimator for the expectations in (a)
- In spite of its correctness, people may conclude that the unfolding is unsatisfactory because of the large fluctuations. What is the problem?
- Paying large price in variance for unbiasedness
  - Transformation to the experimental distribution smoothes the data, as may be seen comparing (b) with (a). This tends to introduce bias and reduce variance
  - We take data from the smoothed distribution, and apply the "unsmoothing" transformation
  - The smallish fluctuations in the dataset are amplified in the process. We eliminate the bias, but increase the variance

Unfolding – Regularization

- Often desirable to obtain a result with less erratic behavior, essentially taking into account our expectation that the fundamental distribution is smooth
- Accomplish this by accepting some bias, applying some smoothing or interpolation in our unfolding process
- This is regularization

*Unfolding – Remark, if needed

- In our histogram example, we might have imagined that our experiment corresponds to a sampling x from the fundamental distribution f, which is transformed to the data we actually observe with $y = Ax + B$. If this is correct, we can simply compute $x = A^{-1}(y - B)$ and exactly recover the sampling x from the fundamental distribution and there is no additional variance introduced
- However, this is not correct. Our sampling is from the transformed distribution e, with whatever fluctuations that allows
- Imagine a transformation that takes an expected signal of one million events down to an expectation of just ten events. We must deal with the statistics of ten instead of a million. Effectively, the transformation has introduced noise into the measurement

Unfolding: Regularization

Many possible approaches to regularization in unfolding. All are ways to trade bias for variance to achieve some optimal balance, perhaps as measured by MISE. Smoothing techniques such as kernels or orthogonal series that we have discussed already are candidates

Eg, consider using kernel density estimation in the context of a convolution kernel and the Fourier transform approach. Instead of the empirical density estimate, we use kernel estimator

$$\hat{e}'(x) = \frac{1}{Nw} \sum_{n=1}^{N} K\left(\frac{x - x_n}{w}\right)$$

Unfolding: Regularization

- Fourier transform is

$$\hat{e}'_{\mathrm{FT}}(t) = \frac{1}{Nw}\sum_{n=1}^{N}\int e^{itx}K\left(\frac{x-x_n}{w}\right)$$
$$= \hat{e}_{\mathrm{FT}}(wt)K_{\mathrm{FT}}(wt),$$

  where e_{FT} is the empirical Fourier transform estimator

- Thus, our smoothed estimater for $f(x)$ becomes:

$$\hat{f}'(x) = \frac{1}{2\pi N}\sum_{n=1}^{N}\int e^{-it(x-x_n)}\frac{K_{\mathrm{FT}}(wt)}{g_{\mathrm{FT}}(t)}dt$$

  Called the standard deconvolution kernel density estimator

- K is chosen to ensure that this transform exists, at least if $g_{\mathrm{FT}}(t) \neq 0$, eg, $K(x) = \sin x/\pi x$, with transform $K_{\mathrm{FT}}(t) = 1$ if $t \in [-1, 1]$ and zero otherwise

Unfolding: Regularization

- Alternative popular approach is to add a penalty function to the optimization problem
- Eg, in a least-squares minimization, where we find $\mu = \hat{\mu}$ that minimizes $(A\mu + B - x)^2$, may instead minimize:

$$(A\mu + B - x)^2 + \lambda^2 (L\mu)^2,$$

  where L is some linear operator that measures lack of smoothness in μ and λ is a regularization or smoothing parameter
- Larger values of λ result in smoother density estimates. This techinique is usually referred to as Tikhonov regularization
- As derivatives measure (lack of) smoothness, L may suitably be chosen as a differential operator

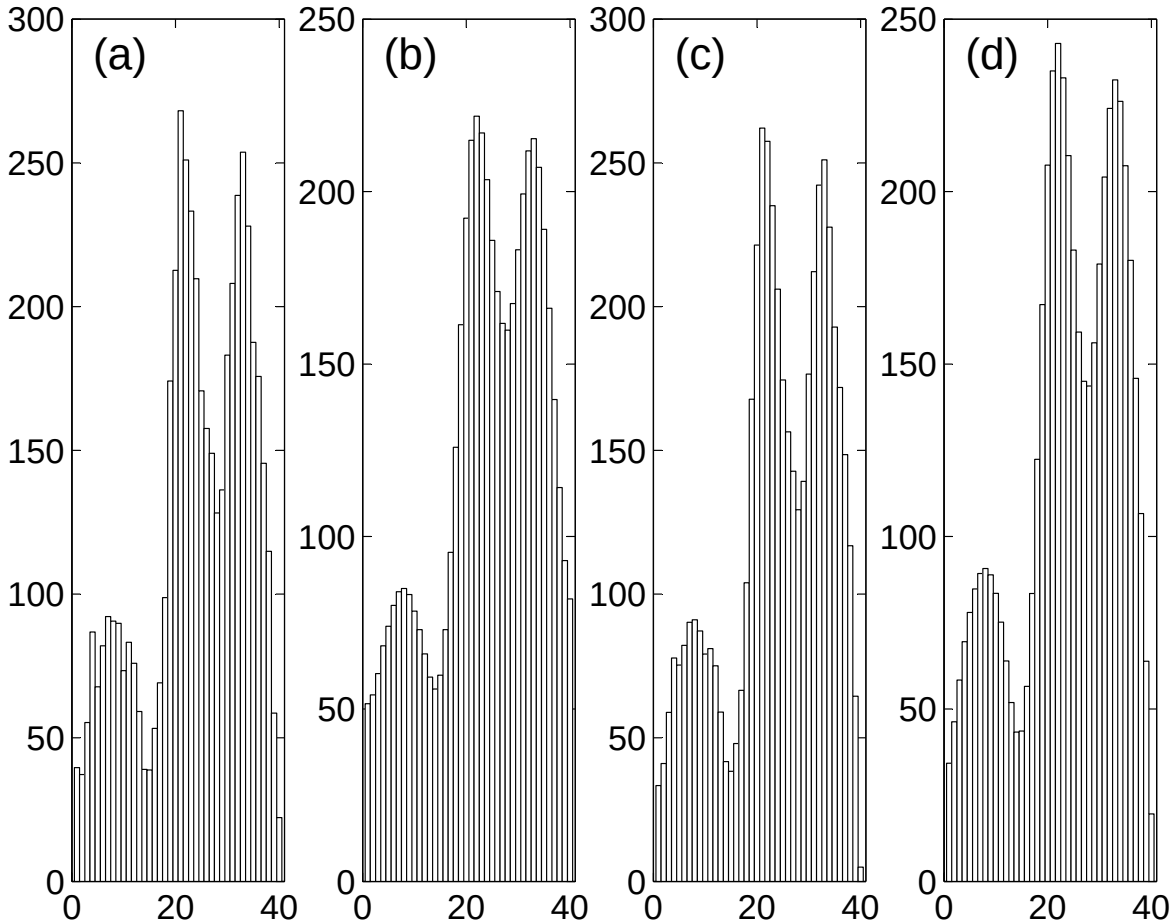
Unfolding: Regularization – Example

- For example, we return to our histogram example
- Since this involves discrete binning, our derivative operator becomes a discrete approximation.The simplest choices, for first- and second-derivatives, are [D_1 is $(N-1)\times N$ and D_2 is $(N-2)\times N$]:

$$D_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 1 & -1 \end{pmatrix}$$

$$D_2 = \begin{pmatrix} -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \end{pmatrix}$$

Unfolding: Regularization – Example



NarskyPorter(2014), Wiley

- Results of Tikhonov regularization for sample histogram
  - (a) and (b) use first derivative smoothing
  - (c) and (d) use the second derivative
  - (a) and (c) use smoothing parameter $\lambda = 0.2$
  - (b) and (d) use $\lambda = 1$

Unfolding: Regularization – Example

- Regularization substantially reduces the variance while adding bias
- At least in this case, the choice of regulator is not very important, but there is a clear, and plausible, dependence on the parameter λ
- The paramter λ may be chosen, for example, to minimize MISE
- It didn't happen in this case, but Tikhonov regularization allows estimators with negative values (if we had picked smaller values for λ, we would have seen this). Not bad or wrong, but may present difficulty if intended use is estimator as a density in further sampling, for example in MC simulations

Bayesian Unfolding

- The unfolding problem may be addressed from Bayesian perspective
- May have prior belief $\pi(f)$ concerning f. As prior belief relates f at different points, the prior contributes a smoothing effect in forming the posterior distribution, given by $P(f|x) \propto P(x|f)\pi(f)$
- Let's try this out for the case of a histogram. We wish to estimate the expected bin contents μ. Pick as the best estimates those for which the posterior distribution is maximal

$$P(\hat{\mu}|x) = \max_{\mu} P(x|\mu)\pi(\mu)$$

- Notice that, without the prior function π, this just gives us the MLE that we obtained earlier. The prior provides regulation

Bayesian Unfolding

- If we have a principled choice for π, we can use it
- But what if we wish to start with a notion of complete ignorance?
- One view of complete ignorance is that a sampling is just as likely to land in one bin as any other. That is, the probability to land in bin i is

$$p_i = \frac{\mu_i}{\mu_T} = p_j = \frac{\mu_j}{\mu_T} = p = \frac{1}{b}$$

That is, our prior distribution for μ is

$$\pi(\mu) = \frac{\mu_T!}{\prod_{i=1}^{b} \mu_i!} \frac{1}{b^{\mu_T}},$$

where $\mu_T = \sum_{i=1}^{b} \mu_i$

Bayesian Unfolding

- Instructive to introduce the notion of entropy, defined by:

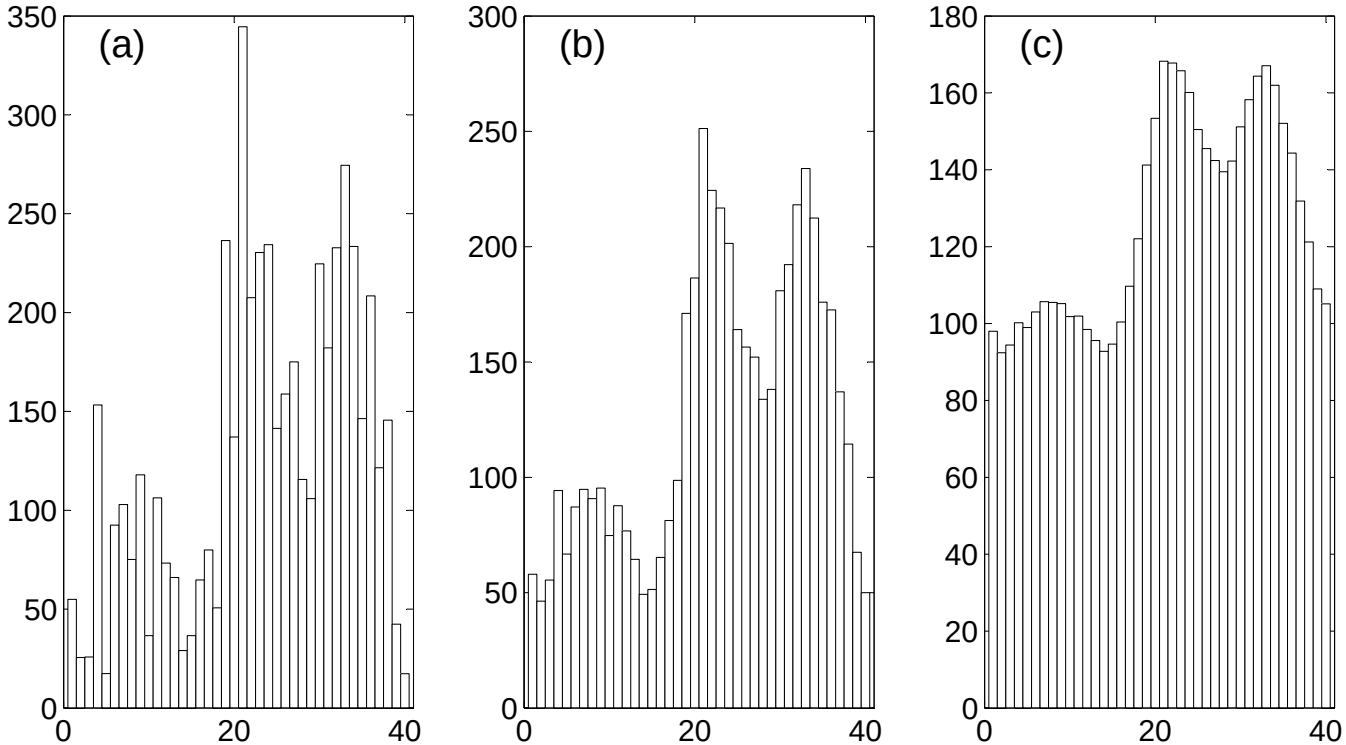
$$H = -\sum_{i=1}^{b} p_i \log p_i$$

This is maximized when $p_i = p, i = 1, \ldots, b$. We recover our notion of complete ignorance as the distribution with maximum entropy

- Apply this technique to our three-peak spectrum. Specifically, maximize (dropping terms independent of μ):

$$\log L(\mu; x) + \log \pi(\mu) = \sum_{i=1}^{b} (x_i \log \nu_i - \nu_i) + \log \Gamma(\mu_T + 1) - \sum_{i=1}^{b} \log \Gamma(\mu_i + 1) - \mu_T \log b$$

Bayesian Unfolding



NarskyPorter(2014), Wiley

- Unfolding a histogram with entropy regularization ($\lambda = 1$ corresponds to formula on previous slide)
  - (a) $\lambda = 0.01$
  - (b) $\lambda = 0.1$
  - (c) $\lambda = 1$

Bayesian Unfolding

- See that this smoothing works, but is smoother than we probably wanted
- Unless one wants to insist that this is the "correct prior", things are getting pulled too far towards the ignorance prior (or towards maximum entropy)
- We can easily salvage the situation if we are willing to give up on a strict Bayesian interpretation
- We accomplish this by multiplying the prior (or the entropy, if we use that for our regulating function) by a regularization parameter, λ. That is, we find μ maximizing:

$$\log L(\mu; x) + \lambda \log \pi(\mu)$$

  The results for $\lambda = 0.01$ and $\lambda = 0.1$ are shown as (a) and (b) on the previous slide
- As before, the value of λ may be optimized to minimize the MISE, in the context of some reasonable approximate model

Unfolding – Conclusions

- To check and tune the unfolding procedure we may use the estimated unfolded distribution as the "actual" distribution in a simulated experiment. The unfolding procedure can be used on the simulated data and compared with the distribution used in the simulation.
- Possible to imagine refinements, such as adaptive regularization. Eg, in image reconstruction there may be true sharp edges that shouldn't be smoothed as much as softer areas
- Unfolding has pitfalls, and caution should be exercised lest one creates a significant-looking peak where none really exists. It should also be kept in mind that the result of unfolding includes correlations. In an unfolded histogram the bin contents are not independent RVs

sPlots

A technique known as the sPlot may be employed to:

- Display estimated contributions of various components in a distribution
- Perform background subtraction
- Reconstruct (eg) Dalitz-plot distributions for signal, and hence to correct the signal yield for a selection efficiency varying across the Dalitz plot, providing the signal branching fraction without assumptions on the resonance structure of the signal

The sPlot is a multivariate technique that uses the distribution on a subset of variables to predict the distribution in another subset. It is based on a (parametric) model in the predictor variables, with different categories or classes (e.g., "signal" and "background").

sPlots

- Assume there are $K + R$ parameters in the overall fit to the data
  1. The expected number of events (observations), $n = \{n_k, k = 1, \ldots, K\}$ in each class
  2. distribution parameters, $\theta = \{\theta_i, i = 1, \ldots, R\}$
- Use a total of N events to estimate these parameters via a ML fit to the (iid) sample $x = x_1, \ldots, x_N$, where x_e is a vector of measurements for event number e
- The likelihood function is:

$$L(n, \theta; x) = \prod_{e=1}^{N} \sum_{k=1}^{K} \frac{n_k}{N} f_k(x_e; \theta),$$

where f_k is the normalized probability density function for category k, and the constraint $\sum_{k=1}^{K} n_k = N$ [can also do case without this constraint]

sPlots

- Goal is to find event weights $w_k(x'_e)$, depending only on $x'_e \subseteq x_e$ (and implicitly on the unknown parameters), such that the asymptotic distribution in $Y_e \notin X'_e$ of the weighted events is the sampling distribution in Y_e, for any chosen class k
- Set relation on x_e, etc., refers to elements of the set of vector components. The possibility that $x'_e = x_e$ is included because y_e could refer to quantities that are not contained in x_e
- It is assumed that Y_e and X'_e are statistically independent within each class. The $w_k(x'_e)$ are the weights we obtain in an actual experiment of size N. They are RVs
- The empirical frequency distribution for y, in class k, is estimated using the weights according to:

$$\hat{g}_k(y) = \sum_{e=1}^{N} w_k(x'_e)\delta(y - y_e)$$

sPlots

- Optimizing the variance of $\hat{g}$ over all y and using the substitution method to estimate unkown quantities yieds the result (Pivk and Le Diberder, NIM A **555** (2005) 356):

$$w(x'_e) \equiv \frac{\hat{V} f(x'_e;\theta)}{\sum_{k=1}^{K} n_k f_k(x'_e;\theta)}$$

- Here, $K \times K$ matrix V is estimated with

$$\hat{V^{-1}} \equiv \sum_{e=1}^{N} \frac{f(x'_e;\theta) f^T(x'_e;\theta)}{\left[\sum_{k=1}^{K} n_k f_k(x'_e;\theta)\right]^2}$$

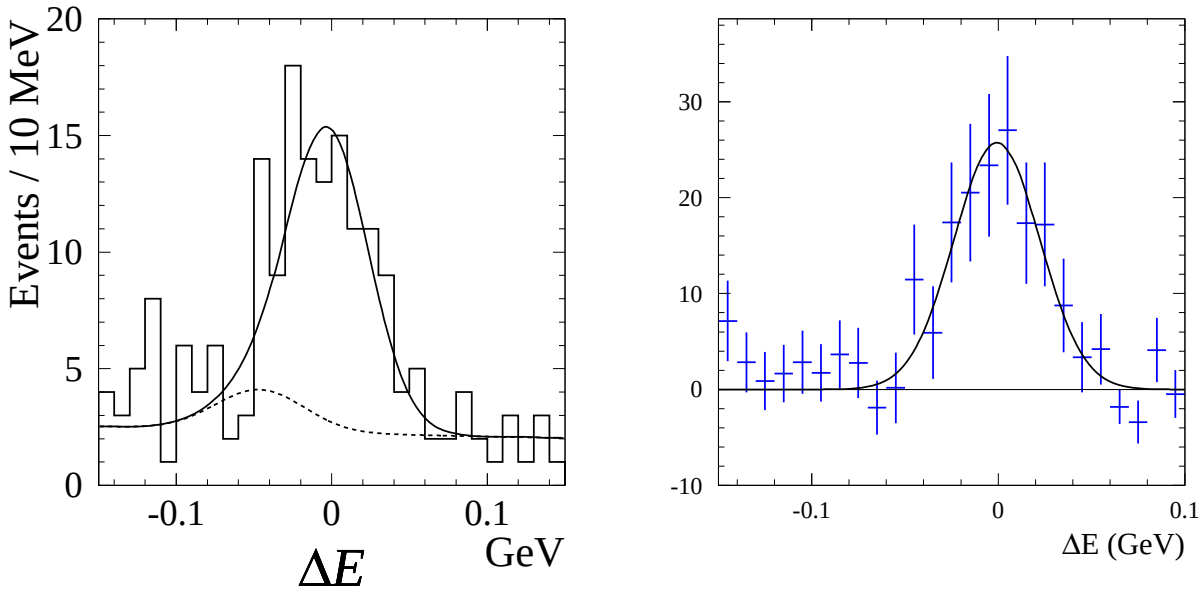
  or from the covarince matrix of a fit excluding y
- The weights may be negative as well as positive

sPlots

- Once the weights are determined, the sPlot is constructed by adding each event e with $y = y_e$ to the y-histogram (or scatter plot, etc, if y is multivariate), with weight $w_j(x'_e)$
- Resulting histogram is an estimator for the true distribution in y for class j
- Typically the sPlot error in a bin is estimated simply according to the sum of the squares of the weights. This sometimes leads to visually misleading impressions, due to fluctuations on small statistics
  - If the plot is being made for a distribution for which there is a prediction, then that distribution might be used to estimate expected uncertainties, and these plotted
  - Or, a (smoothed) estimate from the empirical distribution may be used to estimate the expected errors

sPlot – Example I

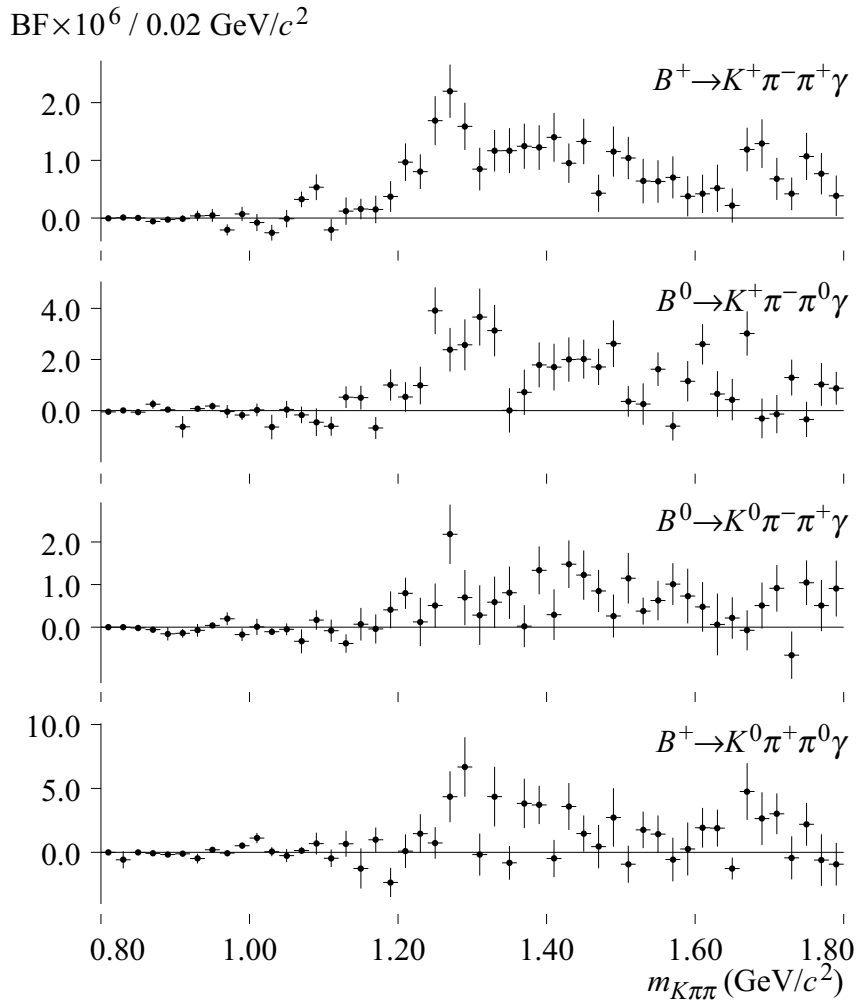
- Illustration of the sPlot technique



  - Left: A non-sPlot, which uses a subset of the data, selected on signal likelihood, in an attempt to display signal behavior (Aubert et al., PRL **89** (2002) 281802)
  - Right: An sPlot for the signal category. Curve is expected signal. Note excess of events at low ΔE. This turned out to be an unexpected portion of the signal distribution, found using the sPlot (Pivk, arXiv:physics/0602023v1(2006))

sPlot – Example II

The sPlot technique used for background subtraction in mass spectra:
The $K\pi\pi$ mass spectra in $B \to \gamma K\pi\pi$ after background subtraction using sPlots



Aubert et al., PRL **98** (2007) 211804