Statistics 240 Lecture Notes

P.B. Stark www.stat.berkeley.edu/∼stark/index.html

DRAFT–Revised 1 May 2008

1 Part 10: Density Estimation. ROUGH DRAFT

References:

Daubechies, I. 1992. *Ten lectures on wavelets*, SIAM, Philadelphia, PA.

Donoho, D.L., I.M. Johnstone, G. Kerkyacharian, and D. Picard, 1993. Density estimation by wavelet thresholding. http://www-stat.stanford.edu/ donoho/Reports/1993/dens.pdf

Evans, S.N. and P.B. Stark, 2002. Inverse problems as statistics, *Inverse Problems*, *18*, R55–R97.

Hengartner, N.W. and P.B. Stark, 1995. Finite-sample confidence envelopes for shape-restricted densities. *Ann. Stat.*, *23*, pp. 525–550.

Silverman, B.W., 1990. *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

1.1 Background

Suppose we observe $\{X_j\}_{j=1}^n$ i.i.d. F, where F has density f with respect to Lebesgue measure on the real line. What can we learn about f from these data?

Estimating f can play a role in exploratory data analysis (EDA) as a graphical summary of the data set. In some contexts, more rigorous estimates and inferences about f and properties of f such as its value at a point $f(x_0)$, its derivative at a point $f'(x_0)$, a Sobolev norm of f such as $\|f\|_S^2 = \int (f^2 + f'^2 + f''^2)dx$, and the number and locations of modes of f, also are interesting. We

will look at some approaches to estimating f, to finding confidence regious for f, and to testing hypotheses about f. We will not dwell on optimality considerations.

1.1.1 The Histogram and the Naive Estimator

This section follows *Silverman* (1990).

Let $\hat{F}_n$ denote the empirical cdf of the data $\{X_j\}_{j=1}^n$:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \le x}. \tag{1}$$

Although $\hat{F}_n$ is often a good estimator of F, $d\hat{F}_n/dx$ is usually not a good estimator of $f = dF/dx$. The derivative of the empirical cdf is a sum of point masses at the observations. It usually is not an enlightening representation of the data.

Suppose we have a collection of *class intervals* or *bins* $\{\mathcal{I}_k = (a_k, a_{k+1}]\}_{k=1}^K$ such that every X_j is in some $\mathcal{I}_k$. (Choosing the intervals to be open on the left and closed on the right is arbitrary; the essential point is that they be disjoint and that their union include all the data.) Let

$$w_k = \text{diam}(\mathcal{I}_k) = a_{k+1} - a_k. \tag{2}$$

The *histogram* of the data using these bins is

$$h(x) = \frac{1}{n}\sum_{k=1}^{K} \frac{1}{w_k} 1_{x \in \mathcal{I}_k} \# \left\{ \{X_j\}_{j=1}^n \cap \mathcal{I}_k \right\}. \tag{3}$$

The histogram is an estimate of f. Its general appearance, including the number and locations of its modes and its smoothness, depends strongly on the locations and widths of the bins. It is blocky and discontinuous. If the bin widths and locations are chosen well, its performance—in the sense of convergence to f in a norm as the sample size n grows—can be reasonable.

Another estimate of f derives from the definition of f as the derivative of F:

$$f(x) = \lim_{\epsilon \to 0} \frac{F(x+\epsilon) - F(x)}{\epsilon} = \lim_{h \to 0} \frac{1}{2h} \Pr\{x - h < X \le x + h\} \tag{4}$$

One could imagine estimating f by picking a small value of h and taking

$$\begin{aligned} \hat{f}_h(x) &\equiv \frac{1}{2h}\left(\hat{F}_n(x+h) - \hat{F}_n(x-h)\right) \\ &= \frac{1}{2nh}\sum_{j=1}^{n} 1_{x-h < X_j \le x+h} \\ &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} K\left(\frac{x - X_j}{h}\right), \end{aligned} \tag{5}$$

where $K(x) = \frac{1}{2} \times 1_{-1<x\leq 1}$. This is the *naive density estimate*. It amounts to estimating $f(x)$ by a superposition (sum) of boxcar functions centered at the observations, each with width $2h$ and area $1/n$. This sum is also blocky and discontinuous, but it avoids one of the arbitrary choices in constructing a histogram: the choice of locations of the bins. As $h \to 0$, the naive estimate converges weakly to the sum of point masses at the data; for $h > 0$, the naive estimator smooths the data. The tuning parameter h is analogous to the bin width in a histogram. Larger values of h give smoother density estimates. Whether "smoother" means "better" depends on the true density f; generally, there is a tradeoff between bias and variance: increasing the smoothness increases the bias but decreases the variance.

It follows from the fact that $\int_{-\infty}^{\infty} K(x)dx = 1$ that

$$\begin{aligned} \int_{-\infty}^{\infty} \hat{f}(x)dx &= \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} \int_{-\infty}^{\infty} K\left(\frac{x - X_j}{h}\right) \\ &= \frac{1}{n}\sum_{j=1}^{n} 1 = 1. \end{aligned} \tag{6}$$

It follows from the fact that $K(x) \geq 0$ that $\hat{f} \geq 0$ for all x. Thus $\hat{f}$ is a probability density function.

1.2 Kernel estimates

The two properties of the boxcar just mentioned—integrating to one and nonnegativity—hold whenever $K(x)$ is itself a probability density function, not just when K is a unit-area boxcar function. Using a smoother *kernel* function K, such as a Gaussian density, leads to a smoother estimate $\hat{f}_K$. Estimates that are linear combinations of such kernel functions centered at the data are called *kernel density estimates*. We denote the kernel density estimate with bandwidth (smoothing parameter) h by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right). \tag{7}$$

The dependence of the estimate on the kernel is not evident in the notation—the kernel is understood from context. Kernels are always chosen to integrate to one, but there can be asymptotic advantages to kernels that are negative in places. The density estimates derived using such kernels can fail to be probability densities, because they can be negative for some values of x. Typically, K is chosen to be a symmetric probability density function.

There is a large body of literature on choosing K and h well, where "well" means that the estimate converges asymptotically as rapidly as possible in some suitable norm on probability density

functions. The most common measure of performance is the mean integrated squared error (MISE):

$$\begin{aligned}\text{MISE}(\hat{f}) &\equiv \mathbb{E}\int(\hat{f}(x)-f(x))^2dx \\ &= \int \mathbb{E}(\hat{f}(x)-f(x))^2dx \\ &= \int(\mathbb{E}\hat{f}(x)-f(x))^2dx + \int \mathbf{Var}(\hat{f})dx. \end{aligned} \tag{8}$$

The MISE is sum of the integral of the squared pointwise bias of the estimate and the pointwise variance of the estimate. For kernel estimates,

$$\mathbb{E}\hat{f}(x) = \int \frac{1}{h}K\left(\frac{x-y}{h}\right)f(y)dy, \tag{9}$$

and

$$\mathbf{Var}\hat{f}(x) = \int \frac{1}{h^2}K\left(\frac{x-y}{h}\right)^2 f(y)dy - \left[\frac{1}{h}\int K\left(\frac{x-y}{h}\right)f(y)dy\right]^2. \tag{10}$$

The expected value of $\hat{f}$ is a smoothed version of f, the result of convolving f with the scaled kernel. If f is itself very smooth, smoothing it by convolution with the scaled kernel does not change its value much, and the bias of the kernel estimate is small. But in places where f varies rapidly compared with the width of the scaled kernel, the local bias of the kernel estimate will be large. Note that the bias depends on the kernel function and the scale (bandwidth) h, not on the sample size.

The two previous expressions for bias and variance rarely lead to tractable computations, but good approximations are available subject to some assumptions about K and f. Suppose K integrates to 1, is symmetric about zero so that $\int xK(x)dx = 0$, and has nonzero finite second central moment $\int x^2K(x)dx = k_2 \neq 0$, and that f has as many continuous derivatives as needed. Then, to second order in h,

$$\mathbf{Bias}_h(x) \approx \frac{1}{2}h^2f''(x)k_2. \tag{11}$$

(See *Silverman* (1990), pp. 38ff.) Thus

$$\int \mathbf{Bias}_h^2(x)dx \approx \frac{1}{4}h^4k_2^2\int(f'')^2(x)dx, \tag{12}$$

confirming more quantitatively that (asymptotically) the integrated bias depends on the smoothness of f. A similar Taylor series approximation shows that to first order in h^{-1},

$$\mathbf{Var}\hat{f}(x) \approx n^{-1}h^{-1}f(x)\int K^2(u)du, \tag{13}$$

so

$$\int \mathbf{Var}\hat{f}(x)dx \approx n^{-1}h^{-1}\int K^2(u)du. \tag{14}$$

Thus shows that to reduce the integrated bias, one wants a narrow kernel, which must have large values to satisfy $\int K = 1$, while to reduce the integrated variance, one wants the kernel to have small values, which requires it to be broad to satisfy $\int K = 1$.

By calculus one can show that the approximate MISE is minimized by choosing the bandwidth to be

$$h^* = n^{-1/5} k_2^{-2/5} \left[\int K^2(u) du \right]^{1/5} \left[\int (f'')^2(x) dx \right]^{-1/5}, \tag{15}$$

which depends on the unknown density f. Note that the (approximately) optimal bandwidth for MISE decreases with n as $n^{-1/5}$. For the (approximately) optimal bandwidth h^*,

$$\text{MISE} \approx n^{-4/5} \times 1.25 C(K) \left[\int (f'')^2(x) dx \right]^{1/5}, \tag{16}$$

where

$$C(K) = k_2^{2/5} \left[\int K^2(u) du \right]^{4/5} \tag{17}$$

depends only on the kernel. The kernel that is approximately optimal for MISE thus has the smallest possible value of $C(K)$ subject to the restrictions on the moments of K. If we restrict attention to kernels that are probability density functions, the optimal kernel is the *Epanechnikov kernel* $K_e(u)$

$$K_e(u) = \frac{3}{4\sqrt{5}} (1 - u^2/5)_+. \tag{18}$$

This is the positive part of a parabola.

One can define the relative efficiency of other kernels compared with the Epanechnikov kernel as the ratio of their values of $C(K)^{5/4}$. Other common kernels include Tukey's Biweight (suitably normalized, this is $\frac{15}{16}(1 - u^2)_+^2$), a triangular kernel, the rectangular kernel of the naive estimate, and the Gaussian density. Table 3.1 on p. 43 of *Silverman* (1990) shows that there is not much variation in the efficiency: the rectangular kernel is worst, with an efficiency of about 93%; the efficiency of the Gaussian is about 95%; the efficiency of the triangular kernel is about 99%; and the efficiency of the Biweight is over 99%. Thus the choice of kernel can reflect other concerns, such as desired properties of $\hat{f}$ (continuity, computational complexity, and so on).

Choosing $h = h(n)$ is much more of a concern for the asymptotic behavior of the density estimate. To a large extent, choosing h is a black art, but there are some automatic strategies that behave well subject to some assumptions. One of the most popular is least-squares cross-validation, which is a resampling method related to the jackknife. Here is a sketch of the method, following *Silverman* (1990), pp. 48ff.

The integrated squared error of a density estimate $\hat{f}$ is

$$\int (\hat{f} - f)^2 dx = \int \hat{f}^2 dx - 2\int \hat{f} f dx + \int f^2 dx. \tag{19}$$

The last term does not involve the density estimate, so it is not in our control. Thus it is enough to try to minimize

$$R(\hat{f}) \equiv \int \hat{f}^2 dx - 2\int \hat{f} f dx. \tag{20}$$

Cross validation estimates $R(\hat{f}_h)$ from the data, and chooses h to minimize the estimate. The first term in $R(\hat{f})$ can be calculated explicitly from $\hat{f}$. Estimating the second term is the crux of the method. By analogy to the jackknife, define the *leave one out* kernel density estimate

$$\hat{f}_{h,(i)}(x) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x - X_j}{h}\right). \tag{21}$$

Let

$$M_0(h) \equiv \int \hat{f}_h^2 - \frac{2}{n} \sum_i \hat{f}_{h,(i)}(X_i). \tag{22}$$

Let's compute the expected value of $M_0(h)$. First note that

$$\begin{aligned} \mathbb{E}\frac{1}{n}\sum_i \hat{f}_{h,(i)}(X_i) &= \mathbb{E}\hat{f}_{h,(1)}(X_1) \\ &= \mathbb{E}\int \hat{f}_{h,(1)}(x) f(x) dx \\ &= \mathbb{E}\int \hat{f}_h(x) f(x) dx. \end{aligned} \tag{23}$$

The last step uses the fact that the expected value of the kernel density estimate depends on K and h but not on the sample size. Thus

$$\mathbb{E}R(\hat{f}_h) = \mathbb{E}M_0(h), \tag{24}$$

and $M_0(h)$ is an unbiased estimator of the ISE of $\hat{f}_h$, less the term $\int f^2$, which does not depend on $\hat{f}_h$. Provided $M_0(h)$ is close to $\mathbb{E}M_0(h)$, choosing h to minimize $M_0(h)$ should select a good value of h for minimizing the MISE of the estimate. The form of $M_0(h)$ is not computationally efficient; simplifications are possible, especially if $K(\cdot)$ is symmetric. Moreover, if we use n in place of $n-1$ in the denominators, we get a similar score function $M_1(h)$ that is easier to compute:

$$M_1(h) = \frac{1}{n^2 h} \sum_i \sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh} K(0), \tag{25}$$

where $K^*(u) = (K \star K)(u) - 2K(u)$ (here $\star$ denotes convolution). The score function $M_1(h)$ can be computed very efficiently by Fourier methods; see § 3.5 of *Silverman* (1990).

A theorem due to Charles Stone (1984) justifies asymptotically choosing h by cross validation using the score function M_1. Stone's theorem says that, subject to minor restrictions on K, the ratio of the integrated squared error choosing h by minimizing M_1 to the integrated squared error for the best choice of h given the sample $\{X_j\}$ converges to 1 with probability 1 as $n \to \infty$.

Cross validation tends to fail when the data have been discretized (binned), because the behavior of $M_1(h)$ at small h is sensitive to rounding and discretization. It can be rescued sometimes by restricting the optimization to a range of values of h that excludes very small values.

The MISE (or an estimate of it) is but one of many possible score functions that could be used in a cross validation scheme. For example, one could use the log likelihood instead, which leads to maximizing the score function

$$\mathrm{CV}(h) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_{h,(i)}(X_i). \tag{26}$$

It turns out that under strong restrictions on f and K, $-\mathrm{CV}(h)$ is (within a constant) an unbiased estimator of the Kullback-Leibler distance between $\hat{f}_h$ and f:

$$I(f, \hat{f}_h) \equiv \int f(x) \log \frac{f(x)}{\hat{f}_h(x)} dx. \tag{27}$$

The score function $\mathrm{CV}(h)$ is not resistant.

1.3 Kernel estimates of multivariate densities

This material is drawn from Chapter 4 of *Silverman* (1990).

Let $\{X_j\}_{j=1}^n$ each take values in $\mathbb{R}^d$, $d \geq 1$. Let $K : \mathbb{R}^d \to \mathbb{R}$ satisfy

$$\int_{\mathbb{R}^d} K(x)dx = 1. \tag{28}$$

Typically, the kernel K is a radially symmetric probability distribution such as the standard multivariate normal, or the multivariate Epanechnikov kernel

$$K_e(x) \equiv \frac{d+2}{2c_d}(1 - \|x\|^2)_+, \tag{29}$$

where c_d is the volume of the unit sphere in $\mathbb{R}^d$:

$$c_d = \int_{\mathbb{R}^d} 1_{\|x\|<1} dx. \tag{30}$$

The kernels

$$K_2(x) \equiv \frac{3}{\pi}(1 - \|x\|^2)_+^2 \tag{31}$$

and

$$K_3(x) \equiv \frac{4}{\pi}(1 - \|x\|^2)_+^3 \tag{32}$$

have more derivatives than the Epanechnikov kernels, and thus produce smoother density estimates; also, they are easier to compute than the multivariate normal density.

Given a multivariate kernel function, the multivariate kernel density estimate is

$$\hat{f}_h(x) \equiv \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right), \tag{33}$$

which is directly analogous to the univariate kernel density estimate. The kernel density estimate is a sum of "bumps" centered at the observations, each with mass $1/n$ and a common width that depends on a tuning parameter, the bandwidth h. The bandwidth h is "isotropic" in that all coordinates are scaled in the same way. If the coordinates are incommensurable (*e.g.*, if the variances of different coordinates are radically different), it can help to transform the coordinate system before using the estimator, for example, by transforming so that the covariance matrix of the observations is the identity matrix. The estimate can then be transformed by the inverse change of variables to get the density estimate in the original coordinate system. This corresponds to the estimate

$$\hat{f}_{h,S}(x) = \frac{1}{\sqrt{|S|}nh^d} \sum_{j=1}^{n} k\left(\frac{\|x - X_j\|_{S^{-1}}^2}{h^2}\right), \tag{34}$$

where $\|x\|_{S^{-1}}^2 \equiv x^T S^{-1} x$, and $k(\|x\|^2) = K(x)$.

Most of the treatment of univariate kernel density estimates carries over, *mutatis mutandis*, to the multivariate case. For example, there is an optimal window width for minimizing the (approximate) MISE; it depends on the smoothness of the underlying density (through $\int(\nabla^2 f)^2$) and on the norm of the kernel and on the second moment of the kernel. Stone's theorem shows that choosing the bandwidth by cross validation using the score function

$$M_1(h) = \frac{1}{n^2 h^d} \sum_i \sum_j K^*\left(\frac{X_i - X_j}{h}\right) + \frac{2}{nh^d} K(0) \tag{35}$$

is asymptotically optimal for MISE.

1.3.1 The *Curse of Dimensionality*

The difficulty of density estimation grows very rapidly as the dimension of the sample space, d, increases. For example, to get relative mean squared error at 0 to be less than 0.1 in estimating a multivariate normal density at zero using the optimal kernel requires $n = 4$ for $d = 1$, $n = 19$ for

$d = 2$, $n = 768$ for $d = 5$, and $n = 842,000$ for $d = 10$ (see Table 4.2 of *Silverman*, 1990). Partly, this is because of the behavior of the volume element in high dimensional spaces.

Consider the unit sphere in dimension d. As d grows, the volume of the sphere is increasingly concentrated in a thin shell near radius $r = 1$. As a result, regions of low density can contribute substantially to the probability in higher dimensions, and regions of high density can remain unsampled even for relatively large sample sizes when d is large. This makes details of the estimate matter increasingly as d grows, and makes it harder to estimate the density even where it is large as d grows.

1.4 Nearest neighbor estimates

This section follows *Silverman* (1990), § 5.2. We start with the d-dimensional case. For any point $t \in \mathbb{R}^d$, define $r_k(t)$ to be the Euclidean distance from t to the kth closest datum in the set $\{X_j\}_{j=1}^n$. Let $V_k(t)$ be the volume in $\mathbb{R}^d$ of a sphere of radius $r_k(t)$:

$$V_k(t) = c_d r_k^d(t), \tag{36}$$

where as before c_d is the volume of the unit ball in $\mathbb{R}^d$. The *nearest neighbor density estimate* is

$$\hat{f}_k(t) \equiv \frac{k}{nV_k(t)} = \frac{k}{nc_d r_k^d(t)}. \tag{37}$$

Usually k is chosen to be small compared with n; $k \approx \sqrt{n}$ is typical in dimension $d = 1$. Larger values of k produce smoother estimates, but the smoothness varies locally: the effective "window" is narrower where the local density of data is higher.

Why does the recipe for the nearest neighbor estimate make sense? If the density at t is $f(t)$, then in a sample of size n, we would expect there to be about $nf(t)V_k(t)$ observations in a small sphere of volume $V_k(t)$ centered at t. If we set the expected number equal to the observed number and solve for f, we get the nearest neighbor estimate:

$$\left\{ n\hat{f}(t)V_k(t) = n\hat{f}(t)c_d r_k^d(t) = k \right\} \Rightarrow \left\{ \hat{f}(t) = \frac{k}{nc_d r_k^d(t)} \right\} \tag{38}$$

At the point t, each datum within a distance $r_k(t)$ of t contributes $1/(nc_d r_k^d(t))$ to the density estimate—as if the density estimate at t were a kernel estimate with the kernel equal to the indicator function of the unit ball in $\mathbb{R}^d$ divided by the volume of the ball (so the kernel integrates to 1), with bandwidth $r_k(t)$. Of course, this bandwidth depends on t through $r_k(t)$, so the nearest neighbor

estimate can be thought of as a kernel density estimate with spatially varying kernel width. (The kernel width depends on the point t at which the estimate is sought, not just on the data $\{X_j\}$. This leads to some difficulties—see below.)

Nearest neighbor estimates are not smooth: although $r_k(t)$ is continuous in t, its derivative fails to exist at points where two or more data are at distance $r_k(t)$ from t. Moreover, the nearest neighbor estimate is not itself a density. Consider what happens as $\|t\|$ grows. When $\|t\|$ is larger than $\max_j \|X_j\|$, $r_k(t)$ grows linearly with $\|t\|$, so the density estimate falls off like $\|t\|^{-d}$—which has infinite integral. This rate of decay does not depend on how the tails of the sample fall off.

Nearest neighbor estimates can be generalized to kernels more complicated than indicator functions. The generalized nearest neighbor estimate using kernel K is

$$\hat{f}(t) = \frac{1}{n r_k^d(t)} \sum_{j=1}^{n} K\left(\frac{t - X_j}{r_k(t)}\right). \tag{39}$$

This reduces to the simple nearest neighborhood estimate when K is the indicator of the unit ball, scaled to have integral 1. The tail behavior of the generalized nearest neighbor estimate depends on details of the kernel K.

Even though at any fixed point, the nearest neighbor estimate is equivalent to a kernel estimate, it is a different kernel estimate at each point. The kernel estimate is a density because it is a linear combination of densities, with coefficients that sum to one. Just one "bump" is centered at each datum. In contrast, with the nearest neighbor estimate, a different bump is centered at each datum in finding the estimate for different values of t: the bandwidth associated with the contribution of the jth datum is a function of t, not just of X_j.

1.4.1 Variable Kernel Method

In contrast, the variable kernel method allows the bandwidth associated with each *datum* to be different, but holds those bandwidths fixed as t varies. Let d_{jk} be the distance from the X_j to its kth nearest neighbor; *i.e.*, $d_{jk} = r_{k+1}(X_j)$. Then the variable kernel estimate is

$$\hat{f} = \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{t - X_j}{h d_{jk}}\right)}{h^d d_{jk}^d}. \tag{40}$$

As h or k grows, the estimate gets smoother. This estimate centers one "bump" of mass $1/n$ at each datum, but the widths of the bumps depend on the local density of observations through d_{jk}^{-1}. Because of this, the estimate is itself a density if the basic kernel K is a density. When the distance

to the kth nearest neighbor is large, the width of the bump is large. Using d_{jk} is an attempt to adapt the bandwidth to the height of the underlying density. However, there are better estimates of the local density to use to adjust the bandwidth. Of course, one can allow the bandwidth to vary in ways other than through d_{jk}.

1.4.2 Adaptive Kernel estimates

This material is drawn from *Silverman* (1990, § 5.3). Instead of using d_{jk}^{-1} as (proportional to) an estimate of the local density for picking the bandwidth for the kernel centered at X_j, one could use a different density estimate. This approach leads to adaptive kernel estimates.

The idea is to make a pilot density estimate, usually highly smoothed, and to base the bandwidth choice for the final estimate on the pilot. Let $\tilde{f}(t)$ be a pilot density estimate for which $\min_j \tilde{f}(X_j) > 0$. Let $\alpha \in [0,1]$. Let g be the geometric mean of the values $\{\tilde{f}(X_j)\}_{j=1}^n$:

$$g = \left(\prod_{j=1}^{n} \tilde{f}(X_j) \right)^{1/n} = \exp\left\{ \frac{1}{n} \sum_{j=1}^{n} \log \tilde{f}(X_j) \right\}. \tag{41}$$

Define

$$\lambda_j = \left(\frac{\tilde{f}(X_j)}{g} \right)^{-\alpha}. \tag{42}$$

The adaptive kernel estimate is

$$\hat{f}(t) = \frac{1}{n} \sum_{j=1}^{n} \frac{K\left(\frac{t-X_j}{h\lambda_j}\right)}{h^d \lambda_j^d}. \tag{43}$$

Note that λ_j plays the role of d_{jk} of the variable kernel method. The estimate depends on a number of tuning constants: h, α, and the tuning constants of the method used to derive the pilot estimate. The overall bandwidth h plays the same role as before. The *sensitivity parameter* α controls how much the bandwidth varies as the pilot estimate varies—the rapidity of variation with $\tilde{f}$. For $\alpha = 0$, the method becomes the ordinary kernel estimate. *Silverman* (1990) says that "there are good reasons for setting $\alpha = 1/2$." (See his § 5.3.3 and reference to Abramson, 1982.)

Silverman (1990) says that the fine details and smoothness of the pilot estimate don't matter much for the final estimate, and recommends using an Epanechnikov kernel estimate with bandwidth chosen to perform well for a standard distribution, calibrated to have the same variance as the sample. He does not advocate using cross validation or other computationally intensive schemes for the pilot estimate.

Silverman reports simulation studies by Breiman, Meisel and Purcell (1977), showing that with the bandwidth chosen optimally, the adaptive kernel method performs remarkably better than the fixed kernel method, even for tame densities such as the normal.

The overall smoothing parameter h for the adaptive kernel estimate can be chosen by least squares cross validation.

1.4.3 Maximum Penalized Likelihood

This section follows § 5.4 of *Silverman* (1990); it is connected to an approach to solving inverse problems, which we will discuss later in the course.

Recall that the MLE of the distribution F is just a sum of point masses at the observations, each with mass $1/n$. This is not very satisfactory as a density estimate because it is so rough. The idea of maximum penalized likelihood is to give up some likelihood in favor of smoothness. We need a functional R that assigns a finite positive number to some subset of all density functions. For example, we might take

$$R(g) = \int (g'')^2 dt. \tag{44}$$

Let $\mathcal{F}$ be the set of probability density functions for which R is defined and finite. For a fixed positive number λ (the smoothing parameter), the penalized log-likelihood function of the density g is

$$\ell_\lambda(g) = \sum_{j=1}^{n} \log g(X_j) - \lambda R(g). \tag{45}$$

The maximum penalized likelihood density estimate $\hat{f}$ is any density in $\mathcal{F}$ for which

$$\ell_\lambda(\hat{f}) \geq \ell_\lambda(g) \ \ \forall g \in \mathcal{F}. \tag{46}$$

The maximum penalized log likelihood estimate is an "optimal" compromise between maximizing the likelihood and being as smooth as possible (in the sense of minimizing R). Estimators that optimize a tradeoff between data fit and simplicity are quite common in many settings; they are called *regularized* estimates. The functional R is called the *regularization functional* or *penalty functional*. Measures of fidelity to the data other than the likelihood are also common.

Finding the maximum penalized likelihood estimator can be made more tractable numerically in a variety of ways, depending on the choice of R. For example, imposing the constraint $g > 0$ is easier if one works with the square-root of the density or with the logarithm of the density, although imposing the other part of the constraint $g \in \mathcal{F}$—that the density integrates to one—is

harder then. Discrete approximations to the density (such as truncated expansions in orthogonal sets of functions) also can simplify the numerics of finding an approximate maximum penalized likelihood estimate. See *Silverman* (1990) for references and more detail.

The penalized maximum likelihood approach, using roughness penalties like those described, treats the underlying density as having homogeneous smoothness. We will talk more about maximum penalized likelihood in the context of nonparametric regression (function estimation).

1.5 Confidence sets for densities with shape restrictions; lower confidence interval for the number of modes

Reference: *Hengartner and Stark* (1995).

1.6 Wavelet shrinkage

Reference: *Donoho et al.* (1993).

1.6.1 Time-frequency localization: windowed Fourier transform and wavelets

1.6.2 Haar wavelets

1.6.3 Unconditional bases

1.6.4 Hard and soft thresholding

1.7 Inverse Problems

Reference: *Evans and Stark* (2002).

1.7.1 Nonparametric regression

1.7.2 Example: Abel's problem

Given a frictionless bowling ball with mass m, a stopwatch and the ability to roll the ball with any desired initial velocity $v_j = v_j(0)$, find the shape of a (2-dimensional) hill by rolling the ball with different velocities and measuring how long it takes the ball to return. The measurements have errors. You can think of this as a way to survey San Francisco on a foggy day.

The *forward problem* is to predict how long it takes the ball to return, if we know the shape of the hill. The *inverse problem* is to use a finite set of measurements to learn something about the shape of the hill.

Let's solve the forward problem. It is convenient to express the shape of the hill as the height $h(s)$ of the hill at an arc distance s along the surface of the hill from where the ball is launched. The initial kinetic energy of the ball is

$$E_j = mv_j^2(0)/2. \tag{47}$$

As the ball ascends, its energy is conserved (the ball is frictionless), but it is partitioned into a kinetic component and a potential component. The potential energy component at arc distance s is

$$E_{Pj}(s) = gmh(s), \tag{48}$$

so, by conservation of energy, the kinetic energy at arc distance s is

$$E_{Kj}(s) = mv_j^2(0)/2 - gmh(s). \tag{49}$$

We can find the velocity of the ball at arc distance s as follows

$$\begin{aligned} mv_j^2(s)/2 &= mv_j^2(0)/2 - gmh(s) \\ v_j^2(s) &= v_j^2(0) - 2gh(s) \\ v_j(s) &= \sqrt{v_j^2(0) - 2gh(s)}. \end{aligned} \tag{50}$$

The velocity of the ball goes to zero (and the ball starts to come back) when $v_j^2(0) = 2gh(s)$, provided the slope of the hill does not vanish there (then the ball would balance and never return). Let s_j satisfy $v_j^2(0) = 2gh(s_j)$. The time it takes the ball to return is equal to the time it takes the ball to ascend. The time it takes the ball to come back is thus

$$\tau_j = 2\int_{s=0}^{s_j} \frac{ds}{\sqrt{v_j^2(0) - 2gh(s)}}. \tag{51}$$

This is the solution to the forward problem. Each transit time τ_j is a nonlinear functional of the hill profile $h(s)$. The inverse problem is to learn something about $h(s)$ from measurements

$$d_j = \tau_j + \epsilon_j, \quad j = 1, \ldots, n, \tag{52}$$

where $\{\epsilon_j\}_{j=1}^n$ are stochastic errors whose joint distribution is assumed to be known—at least up to a parameter or two. (Rarely does anybody allow the joint distribution to be more general than a multivariate zero mean Gaussian with independent components whose variances are known.)

It turns out that this stylized surveying problem is related to inverse problems in seismology, helioseismology, and stereology.

1.7.3 General framework for inverse problems

Observe data X drawn from a distribution $\Pr_\theta$ where θ is unknown, but it is known that $\theta \in \Theta$. Use X and the constraint $\theta \in \Theta$ to learn about θ. For example, we might want to estimate a parameter $g(\theta)$. Assume that Θ contains at least two points; otherwise, we know θ and $g(\theta)$ perfectly even without data.

The parameter $g(\theta)$ is *identifiable* if

$$\{g(\theta) \neq g(\eta)\} \Leftarrow \{\Pr_\theta \neq \Pr_\eta\}, \quad \forall \theta, \eta \in \Theta. \tag{53}$$

In most inverse problems, θ is not identifiable. Little general is known about nonlinear inverse problems, although there are particular nonlinear inverse problems, like the surveying problem above, that are well understood.

(The "trick" to solving the surveying problem is to work with $s(h)$ instead of $h(s)$, on the assumption that $h(s)$ is strictly monotonic. Then the forward mapping τ is a linear functional of $s(h)$, but there are nonlinear constraints–$s(h)$ must also be monotonic. Although hills in San Francisco are not monotonic, in the seismic problem, there are thermodynamic arguments that the corresponding quantity—seismic velocity as a function of radius, divided by radius—is monotonic in Earth's core. See, *e.g.* Stark, P.B., 1992. Inference in infinite-dimensional inverse problems: discretization and duality, *J. Geophys. Res.*, *97*, 14,055–14,082.)

1.7.4 Linear forward and inverse problems

When the forward problem has more structure, more can be said. The best studied class of inverse problems are linear inverse problems.

A forward problem is *linear* if the constraint set Θ is a subset of a separable Banach space $\mathcal{T}$ and for some collection $\{\kappa_j\}_{j=1}^n$ of bounded linear functionals on $\mathcal{T}$,

$$X_j = \kappa_j \theta + \epsilon_j, \tag{54}$$

where $\{\epsilon_j\}_{j=1}^n$ are random errors whose distribution does not depend on θ. Usually, such a forward problem is written

$$X = K\theta + \epsilon, \quad \theta \in \Theta. \tag{55}$$

The corresponding *linear inverse problem* is to use the data X and the constraint $\theta \in \Theta$ to learn about $g(\theta)$. In a linear inverse problem, the distribution of X depends on θ through $K\theta$, so if there exist $\theta, \eta \in \Theta$ such that

$$K\theta = K\eta \text{ but } g(\theta) \neq g(\eta) \tag{56}$$

then $g(\theta)$ is not identifiable.

Let's simplify the setup even further—we assume that $\mathcal{T}$ is a Hilbert space, that $\Theta = \mathcal{T}$, and that g is a *linear parameter*; that is,

$$g(a\theta + b\eta) = ag(\theta) + bg(\eta) \tag{57}$$

for all $a, b \in \mathbb{R}$ and all $\theta, \eta \in \Theta$. The fundamental theorem of Backus and Gilbert says that then $g(\theta)$ is identifiable if and only if $g = \sum_{j=1}^{n} a_j \kappa_j$ for some constants $\{a_j\}$. In that case, if $\mathbb{E}\epsilon = 0$, $\sum_j a_j X_j$ is unbiased for $g(\theta)$, and if e has covariance matrix Σ, the MSE of this (linear) estimator is $a \cdot \Sigma \cdot a^T$. See *Evans and Stark* (2002) for more details and proofs.

1.8 Methods for inverse problems

There is a huge number of methods for "solving" inverse problems, although what qualifies as a solution is debatable. These solution methods can be analyzed using traditional statistical measures of performance, including bias and various loss criteria. Perhaps the most important message is that without constraints, little can be said. The issue is finding constraints justified by the science of the situation that still are helpful in reducing the uncertainty.

- Backus-Gilbert estimation. Finding linear functionals close, in some sense, to point evaluators.
- MLE and variants (regularization, maximum penalized likelihood, method of sieves, singular value truncation and weighting). Trading off fidelity to the data and a measure of complexity or roughness. With suitable assumptions, can show consistency and good rates of convergence if the tradeoff is tuned appropriately.
- Bayes estimation. Difficult to justify in infinite-dimensional problems, because prior probability distributions on infinite-dimensional spaces are strange—it's hard to capture constraints without injecting lots of additional information.
- Minimax estimation. Interesting papers by Donoho and others on estimating linear functionals or the entire model in the Hilbert space case. Connection between deterministic optimal

recovery and minimax statistical estimation in the case that the errors are Gaussian, Θ is convex, and the parameter is a linear functional.

- Shrinkage estimation. Shrinkage can improve MSE of estimates of high-dimensional means. Can help with multiple Backus-Gilbert estimates.

- Wavelet-vaguelette shrinkage estimation. Analogue of wavelet shrinkage density estimation we looked at earlier. Can outperform any linear method in some problems. Papers by Donoho, Johnstone, and others. Key idea is that the wavelet-vaguelette decomposition almost diagonalizes both the prior information and the forward problem.

- Strict bounds. Analog of the method for confidence bounds on shape-restricted densities we looked at earlier in the class. Can get conservative joint confidence sets for arbitrarily many parameters of the model by finding upper and lower bounds on functionals over a set of models that satisfies the constraints and is in an infinite-dimensional confidence set based on fit to the data. Not generally optimal for standard measures of misfit to the data.