Statistical Inference when

$p >> n$ or $p = \infty$

Richard Nickl

—

University of Cambridge

–

Department of Pure Mathematics and
Mathematical Statistics
Statistical Laboratory

I. SOME EXAMPLES

Nonparametric Function Estimation.

$\rightarrow$ *Distribution Function Estimation.*

$$X,X_1,\ldots,X_n \sim^{i.i.d.} F,$$

where

$$F(t) = P(X \leq t), t \in \mathbb{R}^p,$$

is an uknown distribution function.

$\rightarrow$ The obvious estimator is

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(X_i), t \in \mathbb{R}^p.$$

$\rightarrow$ *Density Estimation:*

$$X_1, \ldots, X_n \sim^{i.i.d.} f,$$

where f is an unknown probability density function $f = F'$.

$\rightarrow$ The goal is to estimate f, or a functional of it, examples include

$$\int (D^\alpha f)^2(x)dx, \quad \int f(x)\log(f)(x)dx,$$

$$\frac{f(t)}{1-F(t)} \; etc.$$

$\rightarrow$ Natural density estimators are

$$f_n^K(h,x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$$

for K a kernel, h a bandwidth, or the wavelet estimator at resolution level j,

$$f_n^W(j,x) = \sum_{l=-1}^{j-1}\sum_{k}\widehat{\beta}_{lk}\psi_{lk}(x)$$

where the $\psi_{lk} = 2^{l/2}\psi(2^l x - k)$'s form a wavelet basis and where

$$\widehat{\beta}_{lk} = \frac{1}{n}\sum_{i=1}^{n}\psi_{lk}(X_i).$$

$\rightarrow$ *Dimension Reduction by Differential Geometry.*

$\rightarrow$ Consider a sample $X_1, \ldots, X_n$ of i.i.d. random variables on a compact homogeneous manifold $\mathbf{M}$ of dimension d, with density f.

$\rightarrow$ Examples:
-) (half-) spheres in $\mathbb{R}^d$ (astrophysical data),
-) Grassmann and Stiefel manifolds (directional data),
-) projective spaces (transformed data).

$\rightarrow$ In recent years, wavelets have been constructed on manifolds, known as 'needlets', and they can be used to estimate densities efficiently by localised procedures

$\rightarrow$ These are tight frames of $L^2(\mathbb{M})$ build on the eigenfunctions of the Laplacian on $\mathbb{M}$

$\rightarrow$ In contrast to non-localised (spherical) harmonics, needlets are localised around their center of support, and this can be used for statistical inference.

$\rightarrow$ *Nonparametric Regression.* Sample $(Y_i, x_i)_{i=1}^n$ and postulate a functional relationship

$$Y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \ i.i.d., E(\epsilon_i) = 0$$

where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is an unknown function, and where the x_i are design points/covariates.

$\rightarrow$ *Nonparametric Prediction*: If X_i is random, $E(\epsilon_i|X_i) = 0$ then automatically the regression function equals

$$f(x) = E(Y|X = x).$$

$\rightarrow$ If $f(x, y)$ is the joint density of (X, Y) and if $f^X(x)$ the marginal density of X, then

$$E(Y|X = x) = \int y \frac{f(x, y)}{f^X(x)} dy$$

which motivates the simple estimator

$$\widehat{f}_n(h, x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}$$

for K a kernel, h a bandwidth. More sophisticated techniques exist, but this already works fairly well if we can choose $h > 0$.

$\rightarrow$ *Inverse Problems* The simplest statistical inverse problem is the *deconvolution problem* where one observes corrupted replications $Y_1, \ldots, Y_n$ of Y according to

$$Y = X + \varepsilon,$$

where X is an unobserved random variable and ε a (at least partially) known error distribution, independent of X (so that the law of Y is the convolution of the laws of X, ε).

$\rightarrow$ Nonparametric techniques work in these situations as well, by deconvolving the error distribution, for instance using Meyer wavelets.

$\rightarrow$ A *nonlinear inverse problem* that is often of interest in applications (finance, genetics, biology etc.) is that of estimating the jump measure of a Lévy process. A Lévy process is a stochastic process L_t with independent increments which is, in simplified terms, the independent sum

$$L_t = \sigma B_t + \gamma t + N_t$$

where B_t is a Brownian motion, σ a scalar variance parameter, γ a scalar drift parameter, and N_t a jump (e.g., Poisson or compound Poisson) process.

$\rightarrow$ The simplest example is a compound Poisson process: We observe

$$L_t = \sum_{i=1}^{N_t} X_i$$

where the X_i are i.i.d. with unknown distribution f and N_t is a Poisson process.

$\rightarrow$ So this process jumps at random times with jump size drawn at random from distribution f.

$\rightarrow$ The goal is to estimate f from observing L_t along a discrete trajectory.

$\rightarrow$ In general, the characteristics of the jump component of a Lévy process are described by the *Lévy measure* ν, which in itself is a nonparametric object describing the frequency and size of the random jumps occuring in the jump component of the process.

$\rightarrow$ Using Fourier inversion techniques, ν can be estimated from a discrete observation of the Lévy process. The linearised problem is a deconvolution problem.

$\rightarrow$ **Sparse Linear Regression.** Consider the high-dimensional linear model

$$Y = X\beta + \epsilon$$

where X is a $n \times p$ matrix, $\beta \in \mathbb{R}^p$, $p >> n$, but β is sparse, i.e.,

$$p_0 = \|\beta\|_0 = card\{i : \beta_i \neq 0\} < n$$

so the number p_0 of nonzero coefficients is smaller than n.

$\rightarrow$ So despite $p >> n$ we can still hope to estimate β if $p_0 << n$.

$\rightarrow$ We solve the minimisation problem

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|Y - X\beta\|_2^2 \quad s.t. \quad \sum_{k=1}^{p} |\beta_k| \leq t$$

for some $t > 0$, which is the classical formula to obtain the Lasso-estimator.

$\rightarrow$ Many variations of this idea of ℓ_1-minimisation have been introduced in the last years, see Bühlmann and van de Geer (2011, Springer) for an excellent recent monograph.

$\rightarrow$ **Matrix Recovery.** Consider

$$Y_i = tr(X_i^T A_0) + \epsilon_i, \quad i = 1, \ldots, n$$

where X_i is a known $m_1 \times m_2$ (possibly random) matrix and where A_0 is an unknown $m_1 \times m_2$ matrix, m_1 and/or m_2 are potentially large but

$$rank(A_0) << n$$

is small.

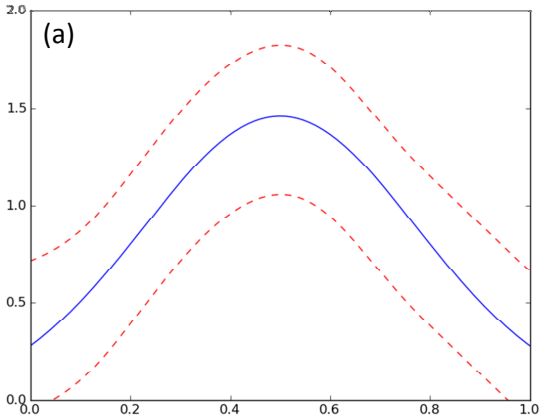
$\rightarrow$ Related problems: Matrix completion, Netflix problem etc... Compressed sensing, nuclear norm penalisation etc...

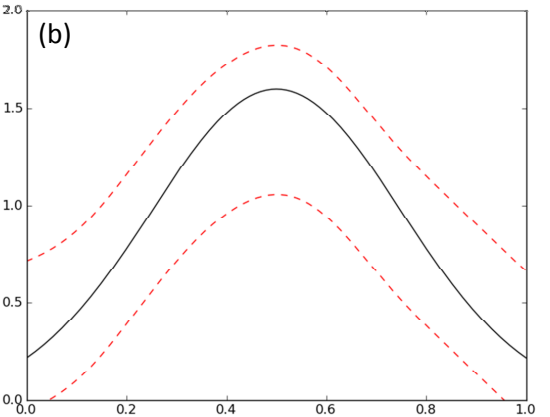
$\rightarrow$ The crucial challenge is clearly an *adaptive choice* of bandwidth h, resolution level j, regularisation parameter t etc.

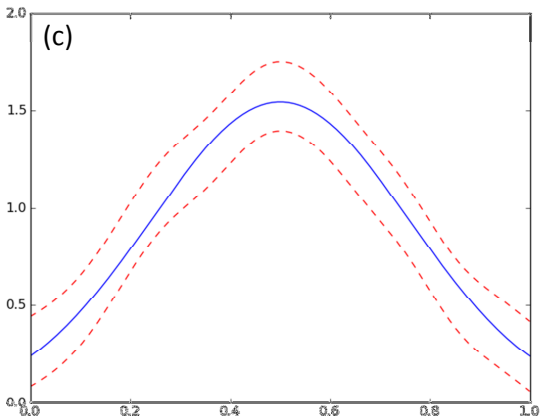
$\rightarrow$ Many methods have been suggested, based on heuristics, theory etc: cross-validation, wavelet thresholding, model selection, Lepski's method, etc...

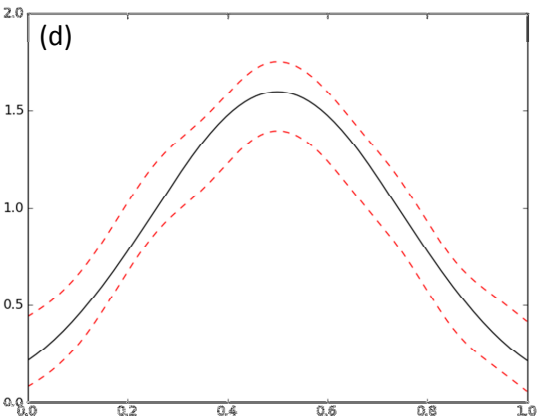
$\rightarrow$ From a heuristic, methodological point of view, one uses the adaptive bandwidth for inference. If possible one should use different samples to estimate h and f.

II. CONFIDENCE BANDS FOR NON-PARAMETRIC FUNCTIONS


(a)
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


(b)
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


(c)
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0


(d)
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0

$\rightarrow$ Given sample and a significance level $0 < \alpha < 1$, we would like to to find a confidence band $\{C_n(y) : y \in [a, b]\}$ for f such that

$$\mathsf{Pr}_f(f(y) \in C_n(y) \quad \forall y \in [a, b]) = 1 - \alpha.$$

$\rightarrow$ This allows to test hypotheses

$$f \in H_0$$

at level α simply by checking if $H_0 \subset C_n$, but is also useful for general goodness of fit tests and graphical purposes etc.

$\rightarrow$ Heuristically, if $C_n = [f_n \pm c_n]$ then for confidence sets we need to understand

$$\mathsf{Pr}_f(|f_n(y) - f(y)| \leq d_n \ \forall y \in [a, b])$$

$$= \mathsf{Pr}_f(\sup_y |f_n(y) - f(y)| \leq d_n),$$

which can be a delicate probabilistic problem.

$\rightarrow$ Distribution Function Estimation. The classical **Kolmogorov-Smirnov theorem** says

$$\sqrt{n} \sup_{t \in \mathbb{R}} |F_n - F|(t) \rightarrow^d \max_{t \in [0,1]} |G(t)|,$$

as $n \rightarrow \infty$ where G is a standard Brownian bridge.

$\rightarrow$ Using the limit distribution one obtains 95-percent confidence bands of width $1.357/\sqrt{n}$, and this works in fact for any n.

$\rightarrow$ **[Bickel and Rosenblatt, 1973]:** Let $f_n^K(h,x)$ be the kernel density estimator with bandwidth h_n. Then one can find sequences of constants

$$A_n \simeq \sqrt{\log n} \simeq B_n$$

such that, under mild conditions on K, h_n, f,

$$A_n \left(\sqrt{nh_n} \sup_{x \in [0,1]} \left| \frac{f_n^K(h,x) - E f_n^K(h,x)}{\sqrt{f(x)}} \right| - B_n \right) \rightarrow^d Z$$

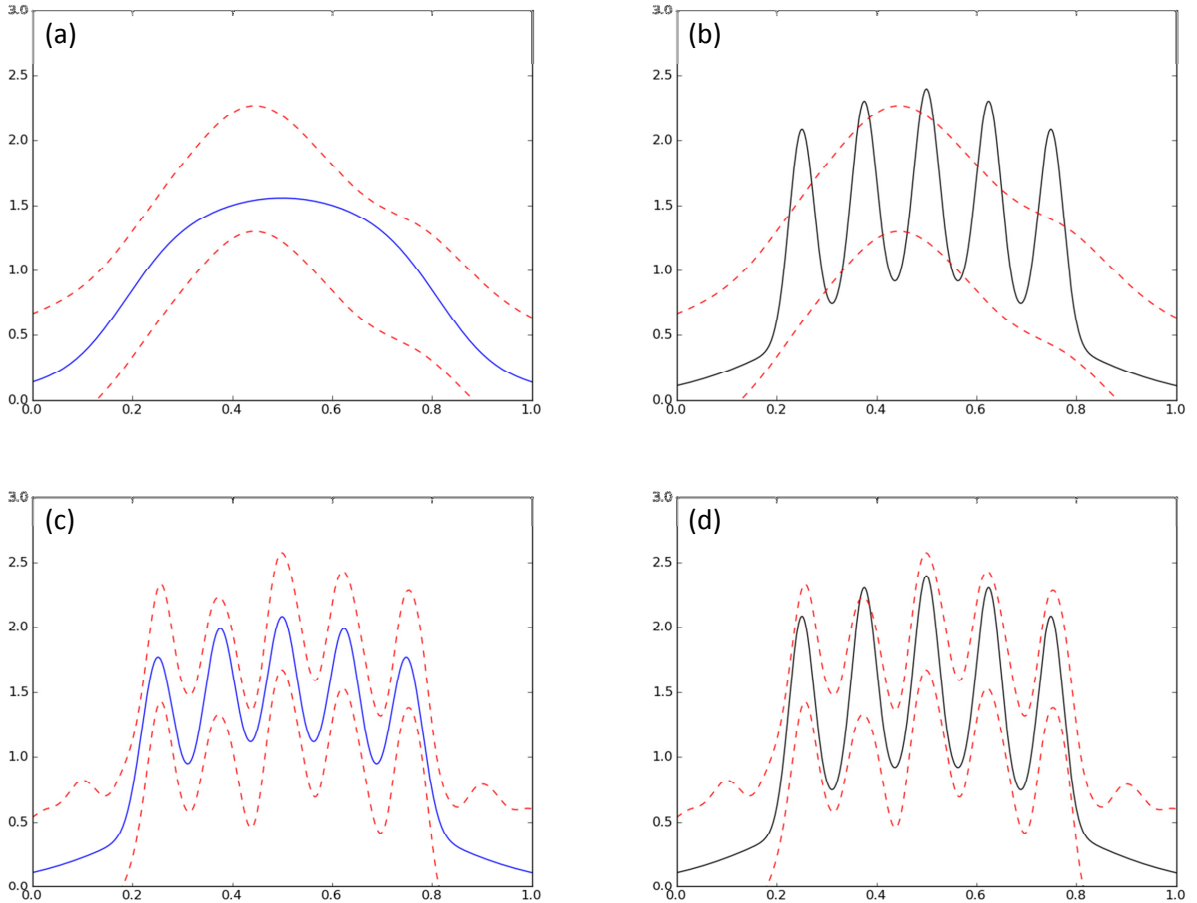
where Z is a Gumbel random variable.

$\rightarrow$ This gives confidence bands combined with slight undersmoothing of h_n.

$\rightarrow$ **[Giné-Nickl-Bull, 2010, 11]**. Let $\widehat{f}_n^W(j,x)$ be the wavelet density estimator with resolution level j_n based on a suitable regular wavelet basis (including Daubechies, splines, etc). Then there exists $A_n \simeq \sqrt{\log n} \simeq B_n$ s.t. under certain conditions on K, h_n, f, for Z a Gumbel variable,

$$A_n \left(\sqrt{n 2^{-j_n}} \sup_{x \in [0,1]} \left| \frac{f_n^W(j,x) - E f_n^W(j,x)}{\sqrt{f(x)}} \right| - B_n \right) \rightarrow^d Z$$

$\rightarrow$ This can be used for statistical inference with wavelets.


(a)
(b)
(c)
(d)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8

$\rightarrow$ This extends naturally to regression, inverse problems etc...

$\rightarrow$ There are also recent nonasymptotic techniques, which work particularly well in inverse problems and on sample spaces with nonstandard geometric structure.

$\rightarrow$ Other techniques involve bootstrap-based methods. Bayesian methods are also of interest, but theory is more difficult here.

For references on these techniques, models, please ask now, check my website, or send me an email

r.nickl@statslab.cam.ac.uk

www.statslab.cam.ac.uk/∼clinic