A Statistician's View

on deconvolution and unfolding

Victor M. Panaretos

Institute of Mathematics
Swiss Federal Institute of Technology, Lausanne

victor.panaretos@epfl.ch

January 20, 2011

Estimation: Parametric and Nonparametric

One of the central problems of statistics is that of estimation.

$\hookrightarrow$ Can be formulated in general within the following framework:

- Can collect data in form of r.v.'s $(X_1, ..., X_n)$ (assume independent)
- These follow a probability model $F \in \mathscr{F}$
- $\mathscr{F}$ is a collection of probability models

The Problem of Estimation

1. Assume that $\mathscr{F}$ is known but F is unknown
2. Observe a realization $(x_1, ..., x_n)$ generated by F
3. Determine which model from $\mathscr{F}$ generated the sample on the basis of $(x_1, ..., x_n)$ (i.e. estimate F)

Estimation: Parametric and Nonparametric

Two broad categories of estimation problems:

- Parametric: $\mathscr{F} = \{F_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$
  - $\hookrightarrow$ The elements of $\mathscr{F}$ are completely functionally known, up to a finite-dimensional Euclidean parameter.
  - $\hookrightarrow$ e.g. $\mathscr{F} = \{\text{Gaussian distns with mean } \theta \in \mathbb{R} \text{ \& variance } 1\}$.
- Nonparametric: $\mathscr{F}$ is a subset of a general function space
  - $\hookrightarrow$ The functional form of the elements of $\mathscr{F}$ is completely unknown, except for some general regularity properties.
  - $\hookrightarrow$ e.g. $\mathscr{F} = \{\text{distributions on } \mathbb{R} \text{ with } C^2 \text{ density}\}$.

Essential difference: In first case, problem reduces to determining a finite-dimensional parameter. In second case, there is no finite-dimensional reduction of the estimation problem.

Inverse Problems in Estimation

Occasionally, problem is further perturbed by measurement limitations:

- Instead of $\mathbf{X} = (X_1, ..., X_n)$ can only observe a proxy $\mathbf{Y} = (Y_1, ..., Y_n)$
- Observations related $\mathbf{Y}$ related to $\mathbf{X}$ via some measurement error ("folding") mechanism:

$$Y_i = g(X_i, \varepsilon_i)$$

  - g is a known differentiable function
  - $g_{1,x}(\cdot) := g(x, \cdot)$ is invertible with differentiable inverse for fixed x (similarly $g_{2,\varepsilon} := g(\cdot, \varepsilon)$)
  - $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)$ have completely known probability distribution.
  - $\boldsymbol{\varepsilon}$ is unobservable

- Will suppose that all random variables are real-valued henceforth.

For parametric problems, this situation is "easy" to handle.

Things become more interesting (=mathematical term for challenging) for nonparametric models, though.

Inverse Problems in Estimation

Assuming that all random variables have probability density functions,

$$\begin{aligned} f_Y(y) &= \int_{\mathbb{R}} f_{Y|X}(y|x) f_X(x) dx \\ &= \int_{\mathbb{R}} f_{g(X,\varepsilon)|X}(y|x) f_X(x) dx \\ &= \int_{\mathbb{R}} h(x,y) f_X(x) dx \end{aligned}$$

where h is connected to the conditional density of ε given X:

$$h(x,y) = \left|\text{Jacobian}_{g_{1,x}^{-1}}(y)\right| \times f_{\varepsilon|X}(g_{1,x}^{-1}(y)|x)$$

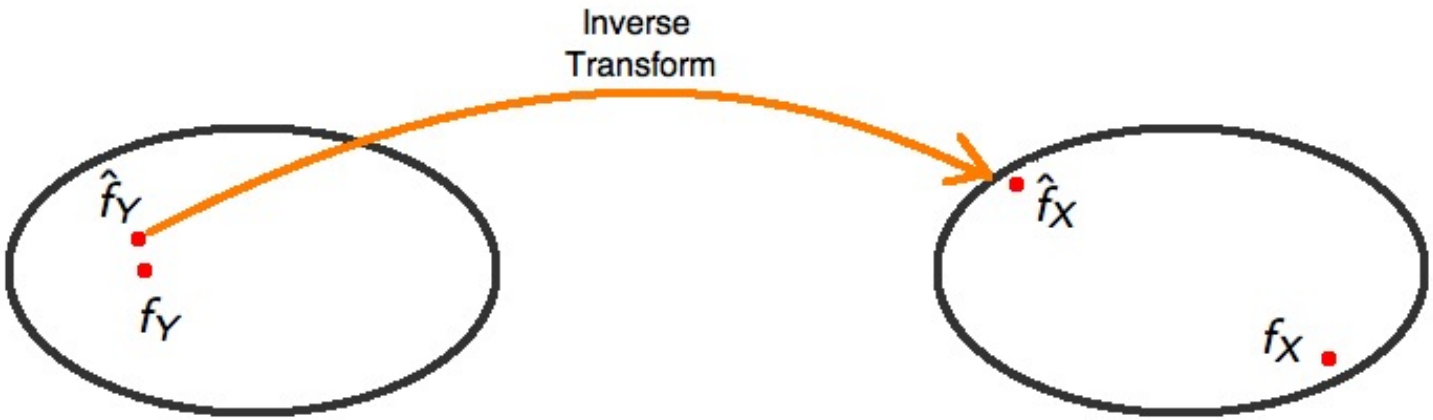
Conclusion: A Fredholm integral equation relates f_Y and f_X

(also see why in parametric case the problem is "essentially" not different)

Inverse Problems in Nonparametric Estimation

In nonparametric setting: no closed form solution for problem (functional form of f_Y is completely unknown)

- Need to first estimate f_Y from data, say $\hat{f}_Y$
- Then attempt to invert integral transformation using $\hat{f}_Y$ as proxy to obtain estimate of object of interest $\hat{f}_X$ (inverse problem)
- Difficulty: small errors in estimation of f_Y may be translated into large errors in estimation of f_X through the inversion process (ill-posedness)[1]



[1] Fredholm transform is bounded, hence inverse is unbounded

(Nonparametric) Deconvolution

Consider special case where $g(X, \varepsilon) = X + \varepsilon$, and ε independent of X,

$$Y_i = X_i + \varepsilon_i, \quad i = 1, ..., n.$$

Then Fredholm equation becomes a convolution equation

$$f_Y(y) = \int_{\mathbb{R}} f_\varepsilon(x - y) f_X(x) dx \qquad \Leftrightarrow \qquad f_Y = f_\varepsilon * f_X$$

Suggests inversion technique:

- Let ϕ_X, ϕ_ε, ϕ_Y be the corresponding characteristic functions (Fourier transforms)
- Then $f_Y = f_\varepsilon * f_X \implies \phi_Y = \phi_X \phi_\varepsilon$
- So if $\hat{\phi}_Y$ is an estimate of the characteristic function of Y, could estimate ϕ_X by $\hat{\phi}_Y / \phi_\varepsilon$, to obtain:

$$\hat{f}_X(t) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itu} \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} du$$

(Nonparametric) Deconvolution

Raises the question: how to estimate ϕ_Y?

$\hookrightarrow$ Answer: good estimates of f_Y give good estimates of ϕ_Y by Fourier transforming
(conversely, good estimates of ϕ_Y give good estimates of f_Y by back-Fourier transforming...)

Let $\widetilde{g}$ denote the Fourier transform of g. Then:

- Plancherel identity: $\int [f_1(y) - f_2(y)]^2 dy = \int [\widetilde{f}_1(u) - \widetilde{f}_2(u)]^2 du$
- So if $\hat{f}_Y$ is a good estimator of f_Y, then $\widetilde{\hat{f}}_Y$ is just as good an estimator of ϕ_Y.
- That is, estimate f_Y, and the Fourier transform to get an estimator of ϕ_Y.

But is it enough to try to estimate f_Y (or, equivalently ϕ_Y) well?

Estimation Error and Ill-Posedness

By the Plancherel theorem, we note that

$$\|f_X - \hat{f}_X\|^2 = \int |f_X(x) - \hat{f}_X(x)|^2 dx = \int \left| \phi_X(u) - \frac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du$$

and, since $f_Y = f_X * f_\varepsilon$, this reduces to

$$\|f_X - \hat{f}_X\|^2 = \int \left| \frac{\phi_Y(u) - \hat{\phi}_Y(u)}{\phi_\varepsilon(u)} \right|^2 du$$

- Typically $\phi_\varepsilon(u)$ decays "fast" as $u \to \pm\infty$ (e.g. Gaussian)
- Small discrepancy between $\hat{\phi}_Y$ and ϕ_Y for a large u, can be blown up to an arbitrarily large discrepancy between f_X and $\hat{f}_X$...
- Such discrepancies are guaranteed to occur by statistical variability

Regularization

In other words, $\dfrac{\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$ is potentially a bad estimator of $\phi_X(u)$

A closer look reveals the reason:

- Assuming that $\int |\phi_Y|^2$ and $\int |\hat{\phi}_Y|^2$ are finite, it must be that $|\phi_Y(u)|$ and $|\hat{\phi}(u)|$ are negligible as $u \to \pm\infty$
- But the decay of the $|\hat{\phi}_Y(u)|$ might not be fast enough to kill the $1/|\phi_\varepsilon(u)|$ term, which tends to infinity, yielding a terrible estimator of ϕ_X for large u.

Solution:

- Since ϕ_X is practically zero outside some domain $\left[-\frac{1}{b}, \frac{1}{b}\right]$, we can set our estimator to zero outside that domain to avoid the blow-up.
- i.e., instead of $\hat{\phi}_Y(u)/\phi_\varepsilon(u)$, use the estimator

$$\frac{\mathbf{1}\{-1/b \le u \le 1/b\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\mathbf{1}\{-1 \le bu \le 1\}\hat{\phi}_Y(u)}{\phi_\varepsilon(u)} = \frac{\widetilde{K}(bu)\hat{\phi}_Y(u)}{\phi_\varepsilon(u)}$$

Regularization

More generally, can use a weight function $\widetilde{K}(bu) := \widetilde{K}_b(u)$, where $\widetilde{K}(u)$ is:

- Supported on $[-1, 1]$.
- Bounded

Corresponds to estimating ϕ_Y by $\widetilde{K}(ub)\hat{\phi}_Y(u)$:

- Tames variation for larger frequencies, leaves estimate unaffected for small frequencies
- Kills variation completely beyond certain frequencies

Illustration: The Naive and Kernel Estimators

Choose the "plug-in" estimator of ϕ_Y, $\hat{\phi}_{naive}(u) = \frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$.

$\hookrightarrow$ A reasonable estimator for every u by the law of large numbers!

- Is unbiased, since $\mathbb{E}[\hat{\phi}_{naive}(u)] = \phi_Y(u)$
- For any u, $\mathbb{E}|\hat{\phi}_{naive}(u) - \phi_Y(u)|^2 = O(n^{-1})$

But is "totally affected" by ill-posedness when wishing to estimate f_X:

- The induced estimator $\hat{f}_X^{naive}(x) = (2\pi)^{-1}\sum_{j=1}^{n}\int \frac{e^{iu(Y_j - x)}}{n\phi_\varepsilon(u)}du$ is not even well-defined!

Regularized Estimator (Kernel Estimator)

But regularized estimator is well-defined, and with controllable variation:

$$\hat{f}_X^{kernel}(x) = \frac{1}{2\pi}\int e^{-iux}\frac{K(bu)\sum_{j=1}^{n}\frac{1}{n}e^{iuY_j}}{\phi_\varepsilon(u)}du$$

Error Properties of the Kernel Estimator

Theorem

Assume that f_X is square-integrable and that $\phi_\varepsilon(u) \neq 0$ everywhere. If $\widetilde{K}(u)$ is bounded and supported on a bounded interval, then

$$\mathbb{E}\|f_x - \hat{f}_X^{kernel}\|^2 = \frac{1}{2\pi n}\int |\widetilde{K}(ub)|^2 \left[\frac{1}{|\phi_\varepsilon(u)|^2} - |\phi_X(u)|^2\right] du + \\ + \frac{1}{2\pi}\int |\widetilde{K}(ub) - 1|^2 |\phi_X(u)|^2 du$$

Two error terms represent 'bias' and 'variance'

- First term (variance term): $\widetilde{K}$ controls blow-up caused by rapid decay of ϕ_ε
- Second term (bias term): $\widetilde{K}$ controls "how far we are on average" from estimating ϕ_X

There exists a fundamental tradeoff as we vary b (no free lunch)

Asymptotic Optimality of the Kernel Estimator

Can we do order of magnitude better than the kernel estimator?

- Answer question asymptotically (sample size $n \to \infty$)
- Need to specify which class of densities $\mathscr{F}$ we will be considering.

Assumptions

(A1) $\mathscr{F}_{\beta,C} = \mathscr{F}_\beta = \{\text{densities } f \text{ such that } \int |\widetilde{f}(u)|^2 |u|^{2\beta} du \leq C\}$

$\hookrightarrow$ equivalently, $\mathscr{F}_{\beta,C} = \{\text{densities } f \text{ such that } \|f^{(\beta)}\|^2 < c(C)\}$

(A2) $\widetilde{K}$ satisfies:

- is square integrable and bounded by 1
- it is supported on $[-1, 1]$
- $\sup_{u \neq 0} \frac{1}{|u|^\beta} |\widetilde{K}(u) - 1| < \infty$

(A3) The error characteristic function satisfies

$$C_1(1 + |u|)^{-\alpha} \leq |\phi_\varepsilon(u)| \leq C_2(1 + |u|)^{-\alpha}$$

Asymptotic Optimality of the Kernel Estimator

Under assumptions (A1), (A2), (A3), we have

Theorem (Upper Bound)

For $b \sim n^{-\frac{1}{2\beta+2\alpha+1}}$, *we have* $\sup_{f \in \mathscr{F}_{\beta,C}} \|\hat{f}_X^{kernel} - f\|^2 = O(n^{-\frac{2\beta}{2\beta+2\alpha+1}})$

Theorem (Lower Bound)

If in addition $|\frac{d}{du}\phi_\varepsilon(u)| \leq const \times |t|^{-\alpha}$, *and* $\beta > 1/2$, *we have*

$$\sup_{f \in \mathscr{F}_{\beta,C}} \|\hat{f}_X - f\|^2 \geq const \times n^{-\frac{2\beta}{2\beta+2\alpha+1}}$$

for any estimator $\hat{f}_X$ *of* f_X, *as* $n \to \infty$.

Analogous results hold for the supersmooth error case (exponential tail decay, e.g. Gaussian).

Time Domain Interpretation: Kernel Density Estimator

The kernel estimator of ϕ_Y amounts to replacing the naive estimator

$$\frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$$

by the regularized estimator

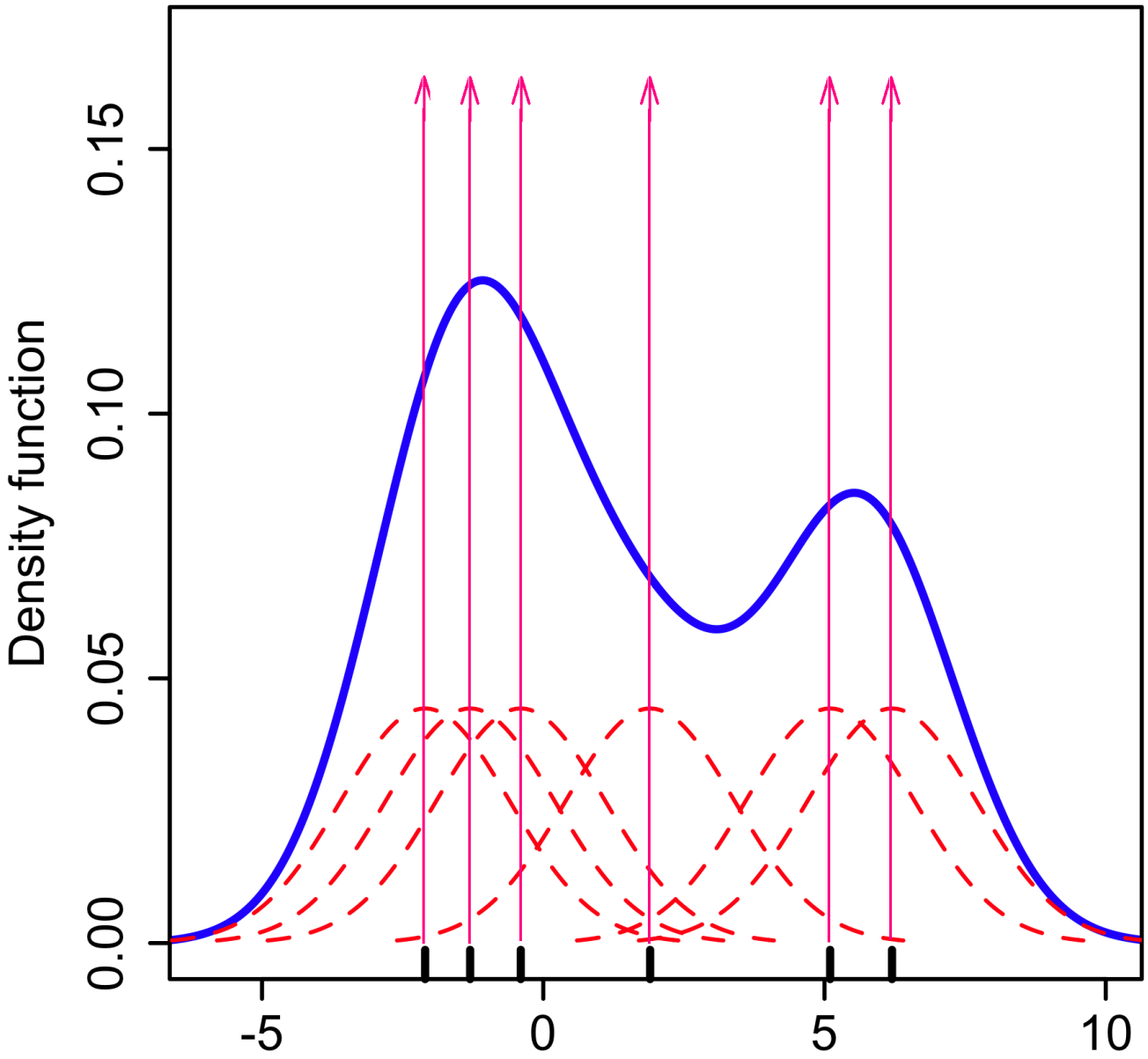
$$\widetilde{K}(ub)\cdot\frac{1}{n}\sum_{j=1}^{n} e^{iuY_j}$$

Applying the inverse Fourier transform, we see that this amounts to estimating f_Y by

$$\left[\frac{1}{b}K\left(\frac{x}{b}\right)\right] * \left[\frac{1}{n}\sum_{j=1}^{n}\delta(Y_j - x)\right] = \sum_{j=1}^{n}\frac{1}{nb}K\left(\frac{Y_j - x}{b}\right)$$

where $\widetilde{K}(u) = \int K(t)e^{itu}dt$ and δ is Dirac's delta.

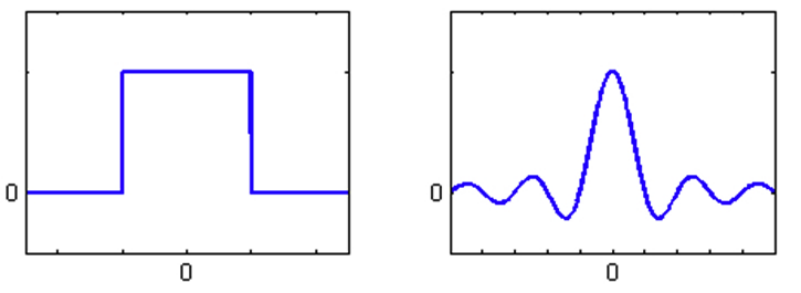
Time Domain Interpretation: Kernel Density Estimator



Time Domain Interpretation: Kernel Density Estimator

Example (Rectangular $\widetilde{K}$)

$$\widetilde{K}(u) = \mathbf{1}\{-1 \leq u \leq 1\} \iff K(x) = \frac{\sin x}{\pi x}$$



Instead of using the "empirical density", KDE uses a smoothed version.

- Empirical density 'fits best' to the observed data but is far too rough
- Roughness of density estimate $\hat{f}_Y \leftrightarrow$ Slow tail decay of $\widetilde{\hat{f}}$
- Reveals the nature of the bias-variance tradeoff seen earlier:
  - Attempt to fit data well
  - But should maintain smoothness to avoid ill-posedness (no overfitting)

Pointwise Error: Bias-Variance Decomposition

Instead of focusing on the whole function, we might be interested in its value at a particular point.

Theorem (Mean Squared Error)

Assume that f_X is bounded and continuous, with absolutely integrable Fourier transform and that ϕ_ε is non-vanishing. If the kernel $\widetilde{K}$ is absolutely and square integrable with bounded support, then

$$\begin{aligned}\mathbb{E}|\hat{f}_X^{kernel}(x) - f_X(x)|^2 &= |\mathbb{E}[\hat{f}_X^{kernel}(x) - f_X(x)]|^2 + Var[\hat{f}_X^{kernel}(x)] \\ &\leq |[K_b * f_x](x) - f(x)|^2 + \\ &\quad + \frac{\|f_X * f_\varepsilon\|_\infty}{2\pi n} \int \left| \frac{\widetilde{K}(bu)}{\phi_\varepsilon(u)} \right|^2 du\end{aligned}$$

for any $x \in \mathbb{R}$.

- Gives a pointwise view of the regularization compromise

Pointwise Asymptotic Optimality of the Kernel Estimator

Under conditions similar in spirit to (A1), (A2), (A3), we have similar properties for the pointwise error of estimation as for the integrated mean squared error, concerning both:

- Upper bounds
- Lower bounds

showing that, under these assumptions, the kernel estimator is asymptotically rate optimal for point estimation.

Pointwise Asymptotics

What about the statistical behaviour of the estimate at a point?

Theorem (Pointwise Central Limit Theorem)

Assume that $f_Y = f_X * f_\varepsilon$ is uniformly bounded. Then, under assumption (A2) on the kernel and (A3) on the error density, provided that b is selected so that $bn \overset{n\to\infty}{\longrightarrow} \infty$,

$$\frac{\hat{f}_X^{kernel}(x) - \mathbb{E}[\hat{f}_X^{kernel}(x)]}{\sqrt{Var[\hat{f}_X^{kernel}(x)]}} \xrightarrow{d} N(0,1).$$

for all $x \in \mathbb{R}$.

- Can be used to construct approximate confidence intervals
  - Hence, can be used for approximate tests
- This will employ the bound on the squared bias
- No CLT when error is supersmooth

Choice of Regularization (Bandwidth) Parameter b

What remains is the fine-tuning of the regularization parameter.

- For supersmooth error densities (e.g. Gaussian), the (asymptotically) optimal regularization parameter is independent of the unknown function f_X, and depends only on the error density
  - In fact we know its exact value and not up to a constant
- However, in the case of simply smooth error density [assumption (A3)] we saw that the bandwidth depends on the smoothness properties of the unknown f_X
  - In fact, even if we knew these exactly, we would still only know how to choose a bandwidth up to a constant.

Choice via Cross-Validation

An idea is to choose b using the data as a guide.

- Can we choose b to minimize $\mathbb{E}\|\hat{f}^{\text{kernel}}_{b,X} - f_X\|^2$?
- Squared norm admits the decomposition:

$$\mathbb{E}\|\hat{f}^{\text{kernel}}_{b,X} - f_X\|^2 = \|\hat{f}^{\text{kernel}}_{b,X}\|^2 - 2\Re\underbrace{\left[\langle \hat{f}^{\text{kernel}}_{b,X}, f_X\rangle\right]}_{\mathbb{E}\left[\overline{\hat{f}^{\text{kernel}}_{b,X}}(X)\right]} + \|f_X\|^2$$

- Notice that:
  - First term depends only on data
  - Second term is estimable from data
  - Third term independent of b
- Idea: Leave-one-out cross validation to estimate $\mathbb{E}\left[\overline{\hat{f}^{\text{kernel}}_{b,X}}(X)\right]$

Adaptivity of Cross-Validation

The problem with cross-validation is that b becomes a random variable, dependent on $\hat{f}_{b,X}^{\text{kernel}}$:

- We use the data twice
- Hence our error and optimality results are not guaranteed to hold true

However, CV asymptotically 'does the right thing':

Under some additional smoothness assumption of f_X and assumption (A3) on f_ε, if we use the sinc kernel to construct $\hat{f}_{b,X}^{\text{kernel}}$ and optimize the empirical integrated squared error on a fine enough grid (depending on n and (A3), we obtain:

$$\limsup_{n\to\infty} \frac{\inf_{b\in \text{Grid}} \text{EmpISE}(b)}{\inf_{b>0} \mathbb{E}\|\hat{f}_{b,X}^{\text{kernel}} - f_X\|^2} \leq 1, \qquad \text{with probability 1.}$$

Estimation of the Distribution Function

Suppose that we also wish to estimate

$$F_{c,d} = F_X(d) - F_X(c) = \mathbb{P}[c \leq X \leq d] = \int_c^d f_X(x)dx.$$

- This is an 'easier' problem (estimate a smooth functional).
- Plug-in idea: have done the hard job of estimating f_X
- So use estimator:

$$\hat{F}_{c,d} = \int_a^b \hat{f}_X^{\text{kernel}}(x)dx = \int \mathbf{1}\{c \leq x \leq d\} f_X^{\text{kernel}}(x)dx$$

- Plancherel's identity now shows that

$$\hat{F}_{c,d} = \frac{1}{n\pi} \sum_{j=1}^{n} \int e^{iu(c+d)/2} \sin[(u(d-c)/2] \frac{\widetilde{K}(ub) e^{itY_j}}{u\phi_\varepsilon(u)} du$$

Estimation of the Distribution: Exact Error Bounds

(B1) Let $\mathscr{H}_{C,\beta}$ be the collection of uniformly bounded densities $\|f\|_\infty = C < \infty$, that are $\lfloor\beta\rfloor$-fold differentiable and globally Hölder continuous, satisfying $|f^{(\lfloor\beta\rfloor)}(x) - f^{(\lfloor\beta\rfloor)}(y)| \leq \|f\|_\infty |x-y|^{\beta-\lfloor\beta\rfloor}$.

(B2) Assume the kernel is absolutely and square integrable, with Fourier transform supported on $[-1,1]$, with $\int |K(t)||t|^{\beta+1}dt < \infty$ and $\int K(t)t^m dt = 0$ for $k \leq \lfloor\beta+1\rfloor$.

Theorem

Under (A1) and (A2) the plug-in estimator satisfies

$$\mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 \leq \|f_Y\|_\infty \frac{2}{\pi n} \int \frac{\left|\sin[t(d-c)/2]\widetilde{K}(tb)\right|^2}{|t|^2|\phi_\varepsilon|^2} dt + \\ +4c^2b^2 \left(\int |K(t)||t|dt\right)^2$$

Estimation of the Distribution: Asymptotic Error Bounds

Theorem

- *Under assumption (A3) with* $0 < \alpha < 1/2$, *if we simply put* $\widetilde{K} = 1$

$$\sup_{f \in \mathscr{H}_{C,\beta}} \mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 = O(1/n)$$

- *Under assumptions (B1), (B2) and (A3) with* $\alpha = 1/2$ *and* $b \sim n^{-1}$

$$\sup_{f \in \mathscr{H}_{C,\beta}} \mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 = O\left(\frac{\log n}{n}\right)$$

- *Under assumptions (B1), (B2) and (A3) with* $\alpha > 1/2$ *and* $b \sim n^{-1/(2\beta+2\alpha+1)}$

$$\sup_{f \in \mathscr{H}_{C,\beta}} \mathbb{E}|\hat{F}_{c,d} - F_{c,d}|^2 = O\left(n^{-(2\beta+2)/(2\beta+2\alpha+1)}\right)$$

More General Measurement Error (Unfolding)

Consider more general $g(X, \varepsilon)$, with ε independent of X,

$$Y_i = g(X_i, \varepsilon_i), \quad i = 1, ..., n.$$

Fredholm problem is now:

$$f_Y(y) = \int h(x, y) f_X(x) dx \implies f_Y = \mathcal{L} f_X,$$

where $\mathcal{L}$ is a more general integral operator.

Further, assume for simplicity that

- The densities involved are square integrable, supported on $[a, b]$
- h is square integrable and symmetric

Same basic strategy and considerations apply

More General Measurement Error (Unfolding)

Then operator $\mathcal{L}$ possesses useful properties:

- Has real eigenvalues $\{\lambda_k\}_{k\geq 1}$ with $\sum_k \lambda_k^2 < \infty$
- Has eigenfunctions $\mathcal{L}\varphi_k = \lambda_k \varphi_k$ forming an orthonormal basis,

$$f(x) = \sum_{k=1}^{\infty} a_k \phi_k$$

for all square integrable f, with

$$a_k = \langle f, \varphi_k \rangle = \int_a^b f(t)\varphi_k(t)dt.$$

Hence,

$$f_Y = \mathcal{L} f_X = \mathcal{L}\left[\sum_{k=1}^{\infty} \langle f_X, \varphi_k \rangle \varphi_k\right] = \sum_{k=1}^{\infty} \langle f_X, \varphi_k \rangle \mathcal{L}[\varphi_k] = \sum_{k=1}^{\infty} \lambda_k \langle f_X, \varphi_k \rangle \varphi_k$$

Inversion and Ill-Posedness

On the other hand, we must also have $f_Y = \sum_{k=1}^{\infty} \langle f_Y, \varphi_k \rangle \varphi_k$ so that

$$\langle f_Y, \varphi_k \rangle = \lambda \langle f_X, \varphi_k \rangle \varphi_k$$

Going backwards, we can invert the transform by taking:

$$f_X = \sum_{k=1}^{\infty} \frac{\langle f_Y, \varphi_k \rangle}{\lambda_k} \varphi_k$$

Suggests strategy similar to deconvolution:

- Estimate f_Y by some reasonable estimator $\hat{f}_Y$
- Then apply inverse transform to $\hat{f}_Y$, to obtain

$$\hat{f}_X = \sum_{k=1}^{\infty} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k$$

Ill-posedness reveals itself: $\lambda_k \downarrow 0$!

Ill-Posedness and Regularization

Note that since $\{\varphi_k\}$ is a basis,

$$\|f_X - \hat{f}_X\|^2 = \sum_{k=1}^{\infty} \left(\langle f_X, \varphi_k \rangle - \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \right)^2 = \sum_{k=1}^{\infty} \frac{\left(\langle f_Y, \varphi_k \rangle - \langle \hat{f}_Y, \varphi_k \rangle \right)^2}{\lambda_k^2}$$

Situation similar to deconvolution:

- Small estimation errors of f_Y magnify to large estimation errors of f_X
- $\langle \hat{f}_Y, \varphi_k \rangle$ potentially not a good enough estimator of $\langle f_Y, \varphi_k \rangle$ in order to beat the blow-up of inverse eigenvalues.
- As in deconvolution, will need to 'tame' this blow-up by regularization.

Spectral Truncation

As with deconvolution, we can argue that:

- Since $\|f_X\|^2 = \sum_{k=1}^{\infty} \langle f_X, \varphi_k \rangle^2 < \infty$, it must be that $\langle f_X, \varphi_k \rangle \to 0$
- Therefore, pick a truncation level B, and enforce:

$$\langle \hat{f}_X, \varphi_k \rangle = \begin{cases} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} & \text{if } k \leq B, \\ 0 & \text{otherwise.} \end{cases}$$

Equivalently, estimate f_Y by

$$\sum_{k=1}^{B} \langle \hat{f}_Y, \varphi_k \rangle$$

Yields estimator:

$$\hat{f}_X = \sum_{k=1}^{B} \frac{\langle \hat{f}_Y, \varphi_k \rangle}{\lambda_k} \varphi_k$$

Choice of truncation point B similar to choice of bandwidth parameter b

Illustration: The Naive and Truncated Estimators

As before, can choose a 'plug-in' estimator of f_Y:

- Note that $\langle f_Y, \varphi_k \rangle = \int_a^b \varphi_k(u) f_Y(u) du = \mathbb{E}[\varphi_k(Y)]$
- Consequently $f_Y(u) = \sum_{k=1}^{\infty} \mathbb{E}[\varphi_k(Y)] \varphi_k(u)$
- Again, law of large numbers seems to suggest naive estimator

$$f_Y^{\text{naive}}(u) = \sum_{k=1}^{\infty} \left[\frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) \right] \varphi_k(u)$$

Problem: This is not even a well-defined estimator!

- The series "converges" to the expansion of the empirical density $\frac{1}{n} \sum_{j=1}^{n} \delta(u - Y_j)$ which is not square integrable:

$$\langle f_Y^{\text{naive}}, \varphi_k \rangle = \frac{1}{n} \sum_{j=1}^{n} \varphi_k(Y_j) = \frac{1}{n} \sum_{j=1}^{n} \int \delta(u - Y_j) \varphi_k(u) du$$

But truncated series, $\hat{f}_Y^{\text{trunc}}$, is well-defined!

The Bias-Variance Decomposition Revisited

It is not hard to see that the truncated estimator satisfies:

$$\mathbb{E}\|f_X - \hat{f}_X^{\text{trunc}}\|^2 = \sum_{i=1}^{B} \frac{\text{Var}[\varphi_k(Y)]}{n\lambda_k^2} + \sum_{k=B}^{\infty} \langle f_X, \varphi_k \rangle^2.$$

Essentially, a bias-variance tradeoff similar with before:

- First term (variance term): B controls the blow-up cause by rapid decay of eigenvalues as compared to the behaviour of $\text{Var}[\varphi_k(Y)]/n$
- Second term (bias term): B controls our systematic deviation from f_X, expressed in terms of what part we are missing due to the truncation.

These are two 'competing' errors that need to be balanced.

Relation to Kernel Estimator

There is a very simple connection between the spectrally truncated naive estimator and a kernel estimator.

There exists a "kernel" such that they coincide

- Let $K_B(x,y) = \sum_{k=1}^{B} \varphi_k(x)\varphi_k(y)$
- Note that

$$\hat{f}_Y(x) = \int_a^b K_B(x,y)\hat{f}_Y^{\text{naive}}(y)dy = \sum_{k=1}^{B} \varphi_k(x) \int_a^b \varphi_k(y)\hat{f}_Y^{\text{naive}}(y)dy$$

$$= \sum_{k=1}^{B} \langle \hat{f}_Y^{\text{naive}}, \varphi_k \rangle \varphi_k(x)$$

- Yields the spectrally truncated estimator for f_X.
- E.g. in bounded deconvolution: spectral truncation $\leftrightarrow$ Dirichlet kernel

A Different Formulation

Instead of observing an iid sample from f_Y, we might assume that we observe f_Y itself, subject to some error:

$$f_Y = \mathcal{L} f_X + \epsilon$$

- For example, ϵ can be thought of as white noise
- In essence this means that we observe the LHS of

$$\langle f_Y, \varphi_k \rangle = \lambda_k \langle f_X, \varphi_k \rangle + \langle \epsilon, \varphi_k \rangle, \qquad k = 1, 2, ...$$

- $\langle \epsilon, \varphi_k \rangle$ is then iid white Gaussian noise
- Leads to the more classical inverse problems framework
- Similar considerations apply
- Regularization: spectral truncation, Tikhonov,...
- Regularizaiton tuning: cross-validation, Stein risk, risk hull...


Bissantz, N., Dümbgen, L., Holzmann, H. & Munk, A. (2007). Nonparametric Confidence bands in deconvolution density estimation. *J. Roy. Stat. Soc. Ser. B,* **69***: 483–506.*


Carrol, R.J. & Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Am. Statist. Assoc.,* **83***: 1184–1186.*


Cavalier, L. (2008). Nonparametric statistical inverse problems. *Inverse Problems,* **24***: 034004.*


Delaigle, A. & Hall, P. (2006). On the optimal kernel choice for deconvolution. *Stat. Prob. Lett.,* **76***: 1594–1602.*


Fan, J. (1991). On the optimal rates of convergence for non-parametric deconvolution problems. *Ann. Stat.* **19***: 1257–1272.*


Hall, P. & Lahiri, S.N.(2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Stat.* **36***: 2110–2134.*


Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer.


Silverman, B.W. (1998). *Density Estimation*. Chapman & Hall.


Stefanski, L.A. & Carrol, R.J. (1990). Deconvoluting kernel density estimators. *Statistics,* **20***: 169–184.*