Adaptive Estimation of the Distribution Function and its Density in Sup-Norm Loss

Evarist Giné and Richard Nickl

Department of Mathematics
University of Connecticut

$\rightarrow$ Let $X_1, ..., X_n$ be i.i.d. with completely unknown law P on $\mathbb{R}$.

$\rightarrow$ Define also $P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$, the measure consisting of point masses at the observations ('empirical measure').

$\rightarrow$ We want to find 'data-driven' functions $T(y, X_1, ..., X_n)$, $y \in \mathbb{R}$, that optimally estimate

(A) the distribution function $F(y) = \int_{-\infty}^{y} dP(x)$;

(B) its density function $f(y) = \frac{d}{dy}F(y)$;

in **sup-norm loss** on the real line.

Case (A): A classical minimax result is

$$\liminf_n \inf_{T_n} \sup_F \sqrt{n} E \sup_{y \in \mathbb{R}} |T_n(y) - F(y)| \geq c > 0.$$

$\rightarrow$ The natural candidate for T_n is the sample cdf $F_n(y) = \int_{-\infty}^{y} dP_n(t)$, which is an **efficient** estimator of F in $\ell^\infty(\mathbb{R})$.

Case (B): If f is contained in some Hölder space $C^t(\mathbb{R})$ with norm $\| \cdot \|_t$, then one has

$$\underline{\lim}_n \inf_{T_n} \sup_{\|f\|_t \leq D} \left(\frac{n}{\log n} \right)^{\frac{t}{(2t+1)}} E \|T_n - f\|_\infty \geq c(D) > 0$$

$\rightarrow$ Clearly, the step function F_n cannot be used to estimate the density f of F.

$\rightarrow$ Can one outperform F_n as an estimator for F in the sense that differentiable F can be estimated without knowing a priori that F is smooth?

$\rightarrow$ Somewhat suprisingly maybe, the answer is **yes**.

Theorem 1 (Giné, Nickl (2008, PTRF)) Let $X_1, ..., X_n$ be i.i.d. on $\mathbb{R}$ with unknown law P. Then there exists a purely-data driven estimator $\widehat{F}_n(s)$ that satisfies

$$\sqrt{n}\left(\widehat{F}_n - F\right) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P.$$

Furthermore, if P has a density $f \in C^t(\mathbb{R})$ for some $0 < t \leq T < \infty$ (where T is arbitrary but fixed), then $\widehat{F}_n$ has a density $\widehat{f}_n$ with pr. approaching one, and

$$\sup_{f:\|f\|_t \leq D} E \sup_{y \in \mathbb{R}} |\widehat{f}_n(y) - f(y)| = O\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right).$$

$\rightarrow$ This estimator can be explicitly written down (it is a nonlinear estimator based on kernel estimators with adaptive bandwidth choice), and we refer to the paper for details. Questions:

A) Can (and should) the estimator $\widehat{F}_n$ be implemented in practice?

B) Can one obtain reasonable asymptotic or even nonasymptotic risk bounds for the adaptive convergence rates? To which extent is this phenomenon purely asymptotic?

$\rightarrow$ To (partially) answer these questions, wavelets turned out to be more versatile than kernels.
If ϕ, ψ are father and mother wavelet and if

$$\widehat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} \phi(X_i - k), \widehat{\beta}_{\ell k} = \frac{1}{n}\sum_{i=1}^{n} 2^{\ell/2}\psi(2^{\ell}X_i - k),$$

then, for $j \in \mathbb{N}$, the (linear) **wavelet** density estimator is, with $\psi_{\ell k} = 2^{\ell/2}\psi(2^{\ell}x - k)$,

$$f_n^W(y, j) = \sum_k \widehat{\alpha}_k \phi(y - k) + \sum_{\ell=0}^{j-1}\sum_k \widehat{\beta}_{\ell k}\psi_{\ell k}(y).$$

$\rightarrow$ This estimator is a projection of the empirical measure P_n onto the space V_j spanned by the associated wavelet basis functions at resolution level j. If ϕ, ψ are the Battle-Lemarié wavelets, this corresponds to a projection onto the classical Schoenberg spaces spanned by (dyadic) B-splines.

$\rightarrow$ It was shown in Giné and Nickl (2007): If

$$2^{j_n} \simeq (n/\log n)^{1/(2t+1)}$$

and if $f \in C^t(\mathbb{R})$, then

$$E \sup_{y \in \mathbb{R}} |f_n^W(y) - f(y)| = O\left((n/\log n)^{t/(2t+1)}\right)$$

and, if $F_n^W(s) := \int_{-\infty}^{s} f_n^W(y)dy$, that

$$\sqrt{n}(F_n^W - F) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P.$$

$\rightarrow$ However, this is of limited practical importance, since $f \in C^t(\mathbb{R})$ is rarely known, and hence the choice $2^{j_n} \simeq (n/\log n)^{1/(2t+1)}$ is not feasible.

$\rightarrow$ A natural way to choose the resolution level j_n is to perform some model selection procedure on the sequence of nested spaces (or 'candidate models') $\mathcal{V}_j$.

HARD THRESHOLDING

The hard thresholding wavelet density estimator introduced by Donoho, Johnstone, Kerkyacharian and Picard (1996) is

$$f_n^T(y) = \sum_k \widehat{\alpha}_k \phi(y-k) +$$

$$\sum_{\ell=0}^{j_0-1} \sum_k \widehat{\beta}_{\ell k} \psi_{\ell k}(y) + \sum_{\ell=j_0}^{j_1-1} \sum_k \widehat{\beta}_{\ell k} 1_{[|\beta_{\ell k}| > \frac{l\tau}{\sqrt{n}}]} \psi_{\ell k}(y),$$

where $j_1 \simeq n/\log n$ and $j_0 \to \infty$ depending on the maximal smoothness up to which one wants to adapt.

Theorem 2 (Giné-Nickl (2007),Thm 8)

For a (reasonable) choice of τ, and under a moment assumption of arbitrary order on $f \in C^t(\mathbb{R})$, one can prove Theorem 1 with $\widehat{F}_n$ the hard thresholding estimator.

$\rightarrow$ This already gives an answer to the first question, since the hard thresholding estimator can be implemented without too much difficulties.

LEPSKI's METHOD

$\rightarrow$ In the model selection context, Lepski's (1991) method can be briefly described as follows:

a) Start with the smallest model $V_{j_{\mathsf{min}}}$; compare it to a nested sequence of larger models

$$\{V_j\}, \quad j_{\mathsf{min}} \leq j \leq j_{\mathsf{max}}$$

b) choose the *smallest* j for which *all* relevant blocks of wavelet coefficients between j and j_{max} are insignificant as compared to a certain threshold.

Formally, if $\mathcal{J}$ is the set of candidate resolution levels between j_{min} and j_{max}, we define $\widehat{j}_n$ as

$$\min\left\{j \in \mathcal{J} : \| f_n^W(j) - f_n^W(l) \|_\infty \leq T_{n,j,l}\, \forall l > j, l \in \mathcal{J}\right\},$$

where $T_{n,j,l}$ is a threshold discussed later.

$\rightarrow$ Note that, unlike hard thresholding procedures, Lepski's method does not discard irrelevant blocks at resolution levels that are *smaller* than $\widehat{j}_n$.

$\rightarrow$ The crucial point is of course the choice of the threshold $T_{n,j,l}$. The general principle behind Lepski's proof is that one needs a sharp estimate for the 'variance-term' of the linear estimator underlying the procedure.

$\rightarrow$ In the i.i.d. density model on $\mathbb{R}$ with sup-norm loss, this means that one needs exact exponential inequalities (involving constants!) for

$$\sup_{y \in \mathbb{R}} |f_n^W(y,j) - E f_n^W(y,j)|.$$

$\rightarrow$ In the Gaussian white noise model often assumed in the literature, exponential inequalities are immediate. Tsybakov (1998) for example works with a trigonometric basis and ends up with a stationary Gaussian process, and then one has the Rice formula at hand.

$\rightarrow$ Otherwise, one needs empirical processes: Talagrand's (1996) inequality, with sharp constants (Massart (2000), Bousquet (2003), Klein and Rio (2005)) can be used here.

$\rightarrow$ To apply Talagrand's inequality, one needs sharp moment bounds for suprema of empirical processes. The constants in these inequalities (Talagrand (1994), Einmahl and Mason (2000), Giné and Guillou (2001), Giné and Nickl (2007)) are not useful in adaptive estimation.

$\rightarrow$ To tackle this problem, we adapt an idea from machine learning due to Koltchinskii (2001, 2006), Bartlett, Boucheron and Lugosi (2002)), and use Rademacher processes.

$\rightarrow$ The following symmetrization inequality is well known: If ε_i's are i.i.d. Rademacher variables independent of the sample, then

$$E\left\|\sum_{i=1}^{n}(f(X_i)-Pf)\right\|_{\mathcal{F}} \leq 2E\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}},$$

and the r.h.s. can be estimated by the (supremum of the) "Rademacher-process"

$$\left\|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right\|_{\mathcal{F}},$$

which is 'purely data-driven' and concentrates (again by Talagrand) in a "Bernstein - way" nicely around its expectation.

$\rightarrow$ In our setup, if

$$K_l(x,y) = \sum_k 2^l \phi(2^l x - k)\phi(2^l y - k)$$

is a wavelet projection kernel, and if ε_i are i.i.d. Rademachers, we set

$$R(n,l) = 2 \sup_{y \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i K_l(X_i, y) \right|.$$

$\rightarrow$ We choose the threshold ($\|\Phi\|_2$ is a constant that depends only on ϕ):

$$T(n,j,l) = R(n,l) + 7\|\Phi\|_2 \|p_n(j_{\max})\|_\infty^{1/2} \sqrt{\frac{2^l l}{n}}.$$

Theorem 3 (GN 2008) Let $X_1, ..., X_n$ be i.i.d. on $\mathbb{R}$ with common law P and uniformly continuous density f. Let

$$\widehat{F}_n(s) = \int_{-\infty}^{s} \widehat{f}_n^W(y, \widehat{j}_n) dy.$$

Then

$$\sqrt{n}\left(\widehat{F}_n - F\right) \rightsquigarrow_{\ell^\infty(\mathbb{R})} G_P.$$

If, in addition, $f \in \mathsf{C}^t(\mathbb{R})$ for some $0 < t \leq r$ then also

$$\sup_{f:\|f\|_t \leq D} E \sup_{y \in \mathbb{R}} |\widehat{f}_n^W(y, \widehat{j}_n) - f(y)| = O\left(\left(\frac{\log n}{n}\right)^{t/(2t+1)}\right)$$

$\rightarrow$ The following theorem uses the previous proof, as well as the exact almost sure law of the logarithm for wavelet density estimators (GN (2007)).

Theorem 1 *Let the conditions of Theorem 3 hold. Then, if* $f \in \mathrm{C}^t(\mathbb{R})$ *for some* $0 < t \leq 1$, *and if* ϕ *is the Haar wavelet, we have*

$$\limsup_n \left(\frac{n}{\log n}\right)^{t/(2t+1)} E\|f_n^W(\hat{j}_n) - f\|_\infty \leq A(p_0)$$

where

$$A(p_0) = 26.6 \left[\frac{1}{\sqrt{2\log 2}(1+t)}\|f\|_\infty^t \|f\|_t\right]^{\frac{1}{2t+1}}$$

$\rightarrow$ For example if $t = 1$,

$$A(p_0) \leq 20 \|f\|_\infty^{1/3} \|Df\|_\infty^{1/3}.$$

$\rightarrow$ The best possible constant in the minimax risk is derived in Korostelev and Nussbaum (1999) for densities supported in $[0, 1]$, and our bound misses the one there by $\simeq 20$.

$\rightarrow$ Some loss of efficiency in the asymptotic constant of *any* adaptive estimator is to be expected in our estimation problem, cf. Lepski (1992) and also Tsybakov (1998).

$\rightarrow$ Our loss is still above that level. The reason behind this is most likely linked to the constant 2 in the Rademacher symmetrization inequality. Note though that without Rademacher symmetrization, one would inflate the constants by a factor of roughly 500.

$\rightarrow$ For densities that attain a critical Hölder singularity (e.g., Jaffard (1999)), one can also obtain finite-sample oracle inequalties in sup-norm. Let

$$\inf_{j \in \mathcal{J}} E\|f_n^W(j) - f\|_\infty = E\|f_n^W(j^H) - f\|_\infty.$$

Proposition 1 *Suppose* $f \in \mathsf{C}^1(\mathbb{R})$ *or assume* $f \in \mathsf{C}^t(\mathbb{R})$ *for some* $0 < t < 1$ *but* $f \notin \mathsf{C}^{t+\delta}(\mathbb{R})$ *for any* $\delta > 0$. *Then, for every* n,

$$E\|f_n^W(\widehat{j}_n) - f\|_\infty \leq \frac{52}{W(j^H, p_0)} E\|f_n^W(j^H) - f\|_\infty$$

$$+ O(n^{-1/2}) + O\left(\left(\frac{\log n}{n}\right)^{2t/(2t+1)}\right).$$

The constant $W(l, f)$ depends on the oscillation of the density at the point where it is least smooth. If a critical Hölder singularity is attained, $W(l, f) \to 0.5$. If f is 'self-similar' in the sense that

$$\sup_k |\beta_{lk}(p_0)| \geq 2^{-l(t+1/2)} w(l)$$

for some positive function $w(l)$, one can obtain simple lower bounds for $W(l, p_0)$. It is an interesting question whether such conditions are necessary?

This talk was based on

-) E. Giné and R. Nickl. An exponential inequality for the distribution function of the kernel density estimator, with applications to adaptive estimation. *Prob. Theory Relat. Fields*, to appear (2008).

-) E. Giné and R. Nickl. Uniform limit theorems for wavelet density estimators. preprint (2007).

-) E. Giné and R. Nickl. Adaptive estimation of the distribution function and its density in sup-norm loss using wavelet and spline projections. preprint (2008).