Nonparametric Inference and Smoothed Empirical Processes

Richard Nickl
University of Cambridge
DPMMS, Statistical Laboratory

March 2013
Spring School "Structural Inference"
Bad Belzig

CONTENTS

I. Introduction (p.4)

II. $1/\sqrt{n}$-problems in Gaussian White Noise (p.21)

III. Smoothed Empirical Processes (p.44)

IV. Smoothed ... indexed by non-Donsker Classes (p.68)

V. Efficient Inference for Inverse Problems I – Deconvolution (p.78)

VI. Efficient ... II – Lévy Processes (p.85)

VII. Weak Limit Theory and Confidence Sets for Likelihood Based Procedures (p.121)

VIII. References (p.154)

Collaborators (chronological):

Evarist Giné, University of Connecticut

Markus Reiß, Humboldt Universität Berlin

Ismaël Castillo, Université Paris VI & VII

I. INTRODUCTION – NONPARAMETRIC STATISTICAL MODELS

$\rightarrow$ *Random Sample.* Consider variables

$$X, X_1, \ldots, X_n, \ i.i.d. \quad X \sim P$$

where P is a probability measure on a subset of $\mathbb{R}^d$. Often we assume P has a probability density function $f : \mathbb{R}^d \rightarrow [0, \infty)$.

$\rightarrow$The empirical measure is $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ and for $g \in L^2(P)$ we know from the CLT

$$(P_n - P)(g) = \frac{1}{n} \sum_{i=1}^n (g(X_i) - Eg(X))$$

$$\approx N\left(0, \frac{\|g - Eg\|_{2,P}}{n}\right).$$

$\rightarrow$ *Gaussian Regression.* Consider observing a signal $f \in L^2([0,1])$ in white noise,

$$dX^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t),$$

where $\sigma/\sqrt{n}, \sigma > 0$ is the signal to noise ratio, dW a standard white noise.

$\rightarrow$ dW arises from the isonormal Gaussian process on L^2, i.e., the (cylindrical) Gaussian probability measure $\mathcal{N}$ on L^2 with marginal laws

$$\int_0^1 g dW \sim N(0, \|g\|_2^2), \ g \in L^2.$$

$\rightarrow$ The equivalent of the empirical measure P_n is simply the observed trajectory $X^{(n)}$, which has marginal distribution

$$X^{(n)}(g) = \int_0^1 g dX^{(n)} \sim N\left(\langle g, f\rangle, \frac{\|g\|_2^2}{n}\right), \ \ g \in L^2.$$

$\rightarrow$ Equivalently, if we test against all $g = e_k$ from an orthonormal basis $\{e_k : k \in \mathbb{Z}\}$ of L^2 we observe a vector

$$X_k = \langle f, e_k\rangle + \frac{1}{\sqrt{n}} g_k, \quad g_k \sim N(0, \sigma^2), \ k \in \mathbb{Z}$$

in infinite sequence space.

$\rightarrow$ Closely related to this model is standard fixed design nonparametric regression

$$Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \varepsilon_i \sim N(0, \sigma^2).$$

$\rightarrow$ The white noise model is in fact asymptotically equivalent to many 'reasonable' nonparametric statistical models, in a similar way as a simple Gaussian shift experiment is asymptotically equivalent to standard parametric 'locally asymptotically normal' models.

$\rightarrow$ The probabilistic structure of the white noise model is an idealised nonparametric Gaussian 'limit experiment'.

$\rightarrow$ *Inverse Problems.* Instead of observing a sample $X_1, \ldots, X_n \sim P$ directly, we observe data corrupted with noise ε_i independent of X, formally

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

where $\varepsilon_i \sim \varphi$, φ a known p.m.

$\rightarrow$ Since the probability distribution of Y_i is

$$P^Y = P * \varphi,$$

and the goal is to estimate P, this problem is often called the *deconvolution problem*.

$\rightarrow$ The empirical measure $P_n^Y = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$ of the observations does now not have the 'right' centering P, but rather P^Y.

$\rightarrow$ Deconvolution is a toy model for general 'operator' inverse problems

$$Y_k = \kappa_k \cdot \langle f, e_k \rangle \quad (+g_k/\sqrt{n}),$$

where the e_k are the SVD-eigenfunctions of some compact operator $\mathcal{K}$ on L^2, with eigenvalues κ_k.

$\rightarrow$ In the presence of additive Gaussian noise we see a nonparametric regression model, but with κ_k corrupting the signal $\langle f, e_k \rangle$.

$\rightarrow$ *Lévy Processes.* We observe n discrete realisations

$$L_{k\Delta}, k = 1, \ldots n,$$

of a Lévy process

$$\{L_t : t \geq 0\}$$

at (equally spaced) time intervals of length $\Delta > 0$ (precise definitions later).

$\rightarrow$ The goal is to estimate the Lévy measure ν that describes the jump part of the process. Here a natural analogue of the 'empirical measure' is not immediately available.

$\rightarrow$ In all the above situations we would like to make inference upon P, f, ν from the observations. Often P is indexed by a parameter/function f, in some parameter space Σ, here typically infinite-dimensional.

$\rightarrow$ As Σ is infinite-dimensional, the choice of 'distance', 'loss', or 'metric' d on Σ can make a crucial difference. Let us illustrate this with a simple example.

$\rightarrow$ **Distribution Function Estimation.**

$$\Sigma = \{F : \mathbb{R} \to [0,1],\ F(-\infty) = 0, F(\infty) = 1,$$

$$F \text{ increasing, right-continuous.}\}$$

then the natural estimator for an unknown distribution function F is

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

$\rightarrow$ **Density Estimation.** The space

$$\Sigma = \left\{ f : \mathbb{R} \to [0, \infty), \int_{\mathbb{R}} f(x)dx = 1 \right\}$$

then

$$\left\{ P_f : f \in \Sigma \right\}$$

is the natural nonparametric model for density estimation on $\mathbb{R}$. A typical estimator is the kernel density estimator

$$\widehat{f}_n(x, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h} \right), h > 0,$$

where K is a nonnegative kernel that integrates to one, and $h > 0$ is a bandwidth.

$\rightarrow$ In both cases we are in some sense estimating P, but with different loss functions.

$\rightarrow$ For cumulative distribution functions, uniformly in $F \in \Sigma$

$$\sqrt{n}\|F_n - F\|_\infty \rightarrow^d \|\mathbb{G}_F\|_\infty.$$

$\rightarrow$ In density estimation, for $\mathcal{C}_s, s > 0$, any Hölder ball, the minimax risk is

$$\inf_{\tilde{f}_n} \sup_{f \in \mathcal{C}_s \cap \Sigma} E_f \|\tilde{f}_n - f\|_\infty \simeq (n/\log n)^{-\frac{s}{(2s+1)}} >> \frac{1}{\sqrt{n}}.$$

Let us divide all statistical problems into 2 zones:

$\rightarrow$ (i) (Σ, d) admits $\sqrt{n}$-consistent estimators ("statistically finite-dimensional models")

$\rightarrow$ (ii) (Σ, d) does NOT admit $\sqrt{n}$-consistent estimators

$\rightarrow$ Problem (i) involves interesting probabilistic problems. The statistical interpretation of the results often proceeds along the lines of the classical parametric theory.

$\rightarrow$ Problem (ii) is statistically more involved: the rate of estimation typically depends on specific geometric/analytic properties of Σ. 'Adaptation' problems arise (pandora's box).

$\rightarrow$ In (i) statistical inference (tests/confidence sets) can often be made via asymptotic theory/bootstrap etc. without serious difficulty

$\rightarrow$ In (ii) honest statistical inference is a highly nontrivial problem. In particular the transition from a good estimator to a good confidence set/test is nonobvious and fundamentally different from the $1/\sqrt{n}$-situation.

$\rightarrow$ In these lectures we do something that looks at first (and perhaps also afterwards!) unconventional. We take the common non-parametric estimators designed for problems of type ii), such as kernel or series estimators, and study them in loss functions from the world i). The main results are:

$\rightarrow$ A) We show that there exist $\sqrt{n}$-problems where naive empirical measures cannot be used, but where world ii)-estimators are optimal. Thus we add to the list of 'nice' $\sqrt{n}$-problems, particularly in the field of statistical inverse problems.

$\rightarrow$ B) We show that typically the standard nonparametric estimators are optimal in 'both worlds' i) and ii) simultaneously, a fact sometimes coined the 'plug-in property' (Bickel and Ritov, 2003) of these estimators. When used carefully this allows to construct valid inference procedures in the difficult world ii) from ideas inspired by i).

$\rightarrow$ [C) The phenomenon in B) extends to 'adaptive' estimators to a large extent. See Giné & N (2009ab) if time here does not permit.]

OUTLINE

$\rightarrow$ Gaussian White Noise Model

$\rightarrow$ Sampling model/smoothed empirical processes.

$\rightarrow$ Application to inverse problems/inference for Lévy processes.

$\rightarrow$ Theory for likelihood based inference procedures (Bayesian/ MLE etc.), and perhaps to adaptation.

II. $1/\sqrt{n}$*-Problems in Gaussian White Noise*

$\rightarrow$ We return to observing, for $f \in L^2$,

$$dX^{(n)}(t) = f(t)dt + \frac{\sigma}{\sqrt{n}} dW(t).$$

As before, for any $g \in L^2$ we have

$$X^{(n)}(g) - \langle f, g \rangle = \int_0^1 g dX^{(n)} - \langle g, f \rangle = \frac{\sigma}{\sqrt{n}} \int_0^1 g dW.$$

$\rightarrow$ So if our aim was to estimate the linear functional $f \mapsto \langle f, g \rangle$ then

$$E_f |X^{(n)}(g) - \langle f, g \rangle| = E_f \left| \frac{\sigma}{\sqrt{n}} \int g dW \right| \leq \frac{\sigma \|g\|_2}{\sqrt{n}}.$$

$\rightarrow$ Hence any such linear functional can be estimated at rate $1/\sqrt{n}$ simply by the plug-in estimator $X^{(n)}(g)$. Note that this includes any continuous linear functional on L^2 (by the Riesz-representation theorem).

$\rightarrow$ One can show that this estimator is in fact (Cramer-Rao) efficient for the usual nonparametric models for f, such as Sobolev/Hölder balls of functions.

$\rightarrow$ It is of interest to make such results uniform in $g \in \mathcal{G}$ bounded in L^2. Define

$$\ell^\infty(\mathcal{G}) = \left\{ H : \mathcal{G} \to \mathbb{R}, \|H\|_{\mathcal{G}} \equiv \sup_{g \in \mathcal{G}} |H(g)| < \infty \right\}.$$

The statistical parameter space embeds

$$\Sigma \subset \ell^\infty(\mathcal{G})$$

via the action, for $f \in \Sigma$,

$$g \mapsto \int_0^1 fg \in \ell^\infty(\mathcal{G}), \;\; \|f\|_{\mathcal{G}} \leq \|f\|_2 \sup_{g \in \mathcal{G}} \|g\|_2 < \infty.$$

$\rightarrow$ In other words, we endow Σ with the metric coming from the norm $\|\cdot\|_{\mathcal{G}}$.

$\rightarrow$ We can then estimate f again by $X^{(n)}$, now in $\|\cdot\|_{\mathcal{G}}$-loss, and as before we have, for $f \in L^2$

$$\|X^{(n)} - f\|_{\mathcal{G}} = \left\| f - f + \frac{\sigma}{\sqrt{n}} W \right\|_{\mathcal{G}} = \frac{\sigma}{\sqrt{n}} \sup_{g \in \mathcal{G}} \left| \int_0^1 g dW \right|.$$

$\rightarrow$ We thus are in world i) if the last supremum is a.s. finite. We say by definition that $\mathcal{G}$ is *pregaussian* if it is the case.

$\rightarrow$ Clearly taking $\mathcal{G}$ equal to a ball of L^2 is too optimistic: Taking

$$\mathcal{G} = \{e_k : k \in \mathbb{Z}\},$$

where the e_k's are orthonormal basis functions of L^2, we see that, for X_k i.i.d.$N(0,1)$,

$$\sup_{k\in\mathbb{Z}} \left| \int_0^1 e_k dW \right| = \sup_{k\in\mathbb{Z}} X_k = \infty \ a.s.$$

$\rightarrow$ This should be no surprise, since $\|f\|_{\mathcal{G}} \simeq \|f\|_2$ if $\mathcal{G}$ is a ball in L^2, and the minimax rates in L^2 are slower than $1/\sqrt{n}$ in general.

$\rightarrow$ A very useful sufficient condition to check finiteness of suprema of Gaussian processes is via the *Dudley integral*.

$\rightarrow$ Let $N(\varepsilon, \mathcal{G}, \|\cdot\|_2)$ be the minimal number of balls of radius ε needed to cover $\mathcal{G}$.

Theorem 1 (Dudley) *If*

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{G}, \|\cdot\|_2)} d\varepsilon < \infty$$

then

$$E \sup_{g \in \mathcal{G}} \left| \int_0^1 g dW \right| < \infty,$$

in particular for every $\delta > 0$, *some* $c > 0$,

$$E \sup_{g \in \mathcal{G}, \|g\|_2 \leq \delta} \left| \int_0^1 g dW \right| \lesssim \int_0^{c\delta} \sqrt{\log N(\varepsilon, \mathcal{G}, \|\cdot\|_2)} d\varepsilon.$$

$\rightarrow$ For $\{e^{ik\cdot} : k \in \mathbb{Z}\}$ the trigonometric basis and $s > 0$, consider a Sobolev ball over $[0, 1]$

$$\mathcal{G}(s, B) = \left\{ g \in L^2 : \sum_{k \in \mathbb{Z}} (1 + k^2)^s |\langle g, e_k \rangle|^2 \leq B^2 \right\},$$

which for $s \in \mathbb{N}$ contains, for some B',

$$\left\{ g \in L^2 : \|g\|_2 + \|D^s g\|_2 \leq B' \right\}.$$

Then for some $0 < A < \infty$,

$$\log N(\varepsilon, \mathcal{G}(s, B), \|\cdot\|_2) \lesssim \left(\frac{AB}{\varepsilon}\right)^{1/s}, \quad \varepsilon > 0$$

so Dudley's theorem applies if $s > 1/2$.

$\rightarrow$ The boundary occurs at $s = 1/2$: Define

$$\mathcal{G}_\delta = \left\{ g : \sum_{k \in \mathbb{Z}} (1 + k^2)^{1/2} \log(e + k)^{2\delta} |\langle g, e_k \rangle|^2 \leq B^2 \right\}.$$

Then one may show that $\mathcal{G}_\delta$ is pregaussian if $\delta > 1/2$ and not otherwise.

$\rightarrow$ These assertions can be proved directly (without Dudley's theorem), using the theory of Hilbert-Schmidt embeddings and some basic Gaussian process theory.

$\rightarrow$ Note that such Sobolev balls contain balls of s-Hölderian functions on $[0, 1]$, $s > 1/2$.

$\rightarrow$ An interesting non-Hilbertian example: a ball $\mathcal{B}_s, s < 1$, in the Besov space over $[0, 1]$

$$B^s_{1\infty} = \left\{ f \in L^1 : \sup_{h>0} \int_{A_h} \frac{|f(x+h) - f(x)|}{h^s} dx \leq B \right\},$$

where $A_h = \{x : x + h \in [0, 1]\}$. For $s = 1$ the space is defined via second differences.

$\rightarrow$ These spaces contain all indicator functions of intervals,

$$\{1_{[0,t]} : t \in [0, 1]\} \subset \mathcal{B}_s \quad \forall s \leq 1,$$

but also much less regular functions, such as 'fractional derivatives' of such indicator functions.

$\rightarrow$ For $s > 1/2$ one shows (Birman & Solomyak)

$$\log N(\varepsilon, \mathcal{B}_s, \|\cdot\|_2) \lesssim \left(\frac{A}{\varepsilon}\right)^{1/s}, \quad \varepsilon > 0,$$

so that this class is still pregaussian. [This result is not trivial, related to sharp estimates of ℓ^2 entropy of ℓ^1-balls, combined with non-linear approximation theory.]

$\rightarrow$ Again this could be refined to cover the limiting case $s = 1/2$, with a logarithmic correction, replacing $\sqrt{h}$ by $\sqrt{h}(\log(1/h))^{\delta}$, $\delta > 1$, in the Hölder type condition above.

$\rightarrow$ Returning to the statistical setting, for $\mathcal{G}$ such a pregaussian Sobolev/Besov ball

$$\|X^{(n)} - f\|_{\mathcal{G}} = \sup_{g \in \mathcal{G}} \left| \int g(x)(dX^{(n)}(x) - f(x)dx) \right|$$
$$= \sup_{g \in \mathcal{G}} \left| \int g dW \right| = O_p(1/\sqrt{n})$$

so that the rate of estimation in these settings is $1/\sqrt{n}$, in the metric $\|\cdot\|_{\mathcal{G}}$.

$\rightarrow$ We have, for example, a white noise version of Donsker's theorem in $C([0,1])$

$$\sqrt{n}\left(X^{(n)}(1_{[0,t]}) - \int_0^t f\right) = \int_0^t dW.$$

$\rightarrow$ Given that we assume $f \in L^2$ one may wonder whether $X^{(n)}$ isn't too crude an estimate. Note $X^{(n)} \notin L^2$ as

$$\|W\|_2 = \sup_{g \in L^2-\text{ball}} \left| \int g(x) dW(x) \right| = \infty \quad a.s.$$

$\rightarrow$ If we want to estimate f in L^2-loss, a more natural estimator of f is for instance

$$\widehat{f}_{n,H} = \sum_{|k| \leq H} \langle e_k, X^{(n)} \rangle e_k, \quad H \in \mathbb{N},$$

a simple orthogonal series estimator, truncated at the H-th frequency, so that $\widehat{f} \in L^2$.

$\rightarrow$ One sees immediately, from the Parseval isometry of L^2 with ℓ^2 that

$$E_f\|\widehat{f}_{H,n} - f\|_2^2 = E\frac{1}{n}\sum_{|k|\leq H}|\langle W, e_k\rangle|^2 + \sum_{|k|>H}|\langle f, e_k\rangle|^2$$

$$\leq \frac{\sigma^2 H}{n} + H^{-2t}\|f\|_{W_2^t}^2$$

where $\|\cdot\|_{W_2^t}$ is the t-Sobolev norm, $t \geq 0$. Thus if we choose $H \sim n^{1/(2t+1)}$ we obtain the rate

$$\|\widehat{f}_{H,n} - f\|_2 = O_P\left(n^{-t/(2t+1)}\right),$$

which is minimax optimal for estimating f in a t-Sobolev ball.

$\rightarrow$ How does $\widehat{f}_{H,n}$ perform in the 'nice' loss function $\|\cdot\|_{\mathcal{G}}$? Let us consider first the simplest case where $\mathcal{G} = \mathcal{G}_s$ is itself a s-Sobolev ball, $s > 1/2$. Then for $g \in \mathcal{G}$ fixed, let us study the distance

$$\|\widehat{f}_{H,n} - X^{(n)}\|_{\mathcal{G}}$$

instead of the distance of $\widehat{f}_{H,n}$ to f. Formally $\pi_H(X^{(n)}) = \widehat{f}_{H,n}$ so

$$\widehat{f}_{H,n} - X^{(n)} = (\pi_H - id)(X^{(n)})$$

where π_H is the L^2-projector onto the span of $\{e_k : |k| \leq H\}$.

We thus have, by standard mean square arguments,

$$\left|\int (\widehat{f}_{H,n} - X^{(n)})g\right| = \left|\int (\pi_H - id)(X^{(n)})g\right| \leq$$

$$\sum_{|k|>H} |\langle f, e_k\rangle||\langle e_k, g\rangle| + \frac{1}{\sqrt{n}}\left|\int \left(\sum_{|k|>H} \langle g, e_k\rangle e_k\right) dW\right|$$

Using the Cauchy Schwarz inequality, the first term is bounded, uniformly in $\mathcal{G}_s$, by a constant multiple of

$$\|f\|_{W_2^t} H^{-s-t} = O\left(n^{\frac{-t-s}{2t+1}}\right) = o(n^{-1/2}),$$

for the above choice of H_n. For the second

term we have, for some sequence $\delta_n \to 0$, noting $\sup_{g \in \mathcal{G}_s} \|g - \pi_H(g)\|_2 \to 0$ as $H \to \infty$,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} E \sup_{g \in \mathcal{G}_s} \left| \int \sum_{|k|>H} \langle g, e_k \rangle e_k dW \right| \\
&\leq \frac{1}{\sqrt{n}} E \sup_{g \in \mathcal{G}_s, \|g\|_2 \leq \delta_n} \left| \int g dW \right| \\
&\lesssim \frac{1}{\sqrt{n}} \int_0^{\delta_n} (1/\varepsilon)^{1/2s} d\varepsilon \\
&= o\left(\frac{1}{\sqrt{n}} \right)
\end{aligned}$$

by Dudley's theorem and since $s > 1/2$.

$\rightarrow$ We can thus conclude that such $\widehat{f}_{H,n}$ is asymptotically linear in $X^{(n)}$,

$$\|\widehat{f}_{H,n} - X^{(n)}\|_{\mathcal{G}} = o_P(1/\sqrt{n})$$

and thus in particular

$$\|\widehat{f}_{H,n} - f\|_{\mathcal{G}} = O_P(1/\sqrt{n})$$

in fact

$$\sqrt{n}(\widehat{f}_{H,n} - f) \rightarrow^d W \quad \text{in } \ell^\infty(\mathcal{G})$$

where W is the standard white noise.

$\rightarrow$ Precisely the same arguments go through for the 'sharp' 1/2-δ-Sobolev ball $\mathcal{G}$.

$\rightarrow$ Note that any bandwidth H such that $H \to \infty$, $H^{-s-t} = o(1/\sqrt{n})$ is satisfied is admissible. In particular $t = 0, H \simeq n$ would always work.

$\rightarrow$ When t is known, most interesting is still $\widehat{f}_{H,n}$ with $H \sim n^{1/(2t+1)}$ where, from the above,

$$\|\widehat{f}_{H,n} - f\|_2 = O_P\left(n^{-t/(2t+1)}\right)$$

and, simultaneously,

$$\sqrt{n}(\widehat{f}_{H,n} - f) \to^d W \sim \mathcal{N} \quad \text{in } \ell^\infty(\mathcal{G}).$$

One may then use the weak asymptotics to construct a set

$$C_n = \{f : \|\widehat{f}_{H,n} - f\|_{\mathcal{G}} \leq z_\alpha/\sqrt{n}\}$$

where z_α are the α quantiles of the distribution of $\|W\|_{\mathcal{G}}$. If we intersect this set with a W_2^t-ball of radius

$$\|\widehat{f}_{H,n}\|_{W_2^t}(1 + c_n), c_n \to \infty,$$

to define a confidence set C_n one can show

$$P_f(f \in C_n) \to 1 - \alpha$$

as $n \to \infty$ whenever $f \in W_2^t$ (honestly).

$\rightarrow$ Moreover, the diameter of this confidence set can be shown to be of order

$$|C|_2^2 = O_P\left(n^{-2t/(2t+1)} \log n\right)$$

uniformly in balls of W_2^t.

$\rightarrow$ The proof of this is not difficult, using interpolation arguments. We postpone a proof to the last lecture, where we shall use this in the setting of exact asymptotics of nonparametric posterior distributions

$\rightarrow$ Similar constructions give us confidence sets that contract in $\|\cdot\|_\infty$-diameter, using the Besov-ball theory (plus duality arguments)

$\rightarrow$ This gives an alternative to 'classical' methods for nonparametric confidence sets, based on the exact normal asymptotics of

$$\|\widehat{f}_{H,n} - E\widehat{f}_{H,n}\|_2^2 - E\|\widehat{f}_{H,n} - E\widehat{f}_{H,n}\|_2^2$$

or the exact Gumbel asymptotics of

$$\|\widehat{f}_{H,n} - E\widehat{f}_{H,n}\|_\infty - E\|\widehat{f}_{H,n} - E\widehat{f}_{H,n}\|_\infty$$

and 'undersmoothing', i.e., ignoring the bias

$$\|E\widehat{f}_{H,n} - f\|,$$

which incurs a similar $\log n$-penalty in the rate of $|C|$.

$\rightarrow$ The classical asymptotics perhaps give geometrically more intuitive confidence sets, but effectively completely ignore the bias, whereas the above confidence sets really are built for the full parameter. The proofs show that the underlying ideas are similar, but the above asymptotics are drawn from a different probabilistic approach.

$\rightarrow$ The new approach is particularly interesting in situations where frequentist methods are difficult to deal with mathematically: e.g., adaptive estimators, nonparametric MLEs and Bayes methods (see last lecture).

III. Smoothed Empirical Processes

$\rightarrow$ We now leave the idealised world of the Gaussian white noise model. Consider i.i.d. $X_1, \ldots, X_n$ drawn from law P.

$\rightarrow$ The obvious estimator of P is the empirical measure $P_n = \frac{1}{n}\sum_i \delta_{X_i}$. For any $g \in L^2(P)$ we have from the CLT

$$\sqrt{n}(P_n - P)(g) \rightarrow^d N(0, \|g - Pg\|_{2,P}^2)$$

so that $P_n g$ estimates the linear functional Pg at rate $1/\sqrt{n}$. This estimator is efficient for usual nonparametric models.

$\rightarrow$ As before it is interesting to ask in which sense this is uniform in $g \in \mathcal{G}$.

$\rightarrow$ By definition a class $\mathcal{G}$ of functions is called P-Donsker iff

$$\sqrt{n}(P_n - P) \rightarrow^d \mathbb{G}_P \text{ in } \ell^\infty(\mathcal{G})$$

where $\mathcal{G}_P$ is the P-Brownian bridge process indexed by $\mathcal{G}$. [..measurability issues $\rightarrow$ next spring school...]

$\rightarrow$ For this to be true a fortiori $\mathbb{G}_P$ needs to be pregaussian, particularly that

$$\sup_{g \in \mathcal{G}} |\mathbb{G}_P(g)| < \infty \ a.s. \ \Rightarrow \mathbb{G}_P \in \ell^\infty(\mathcal{G}).$$

$\rightarrow$ As in Dudley's theorem, a sufficient condition for G_P to be pregaussian is that

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{G}, \|\cdot\|_{2,P})} d\varepsilon < \infty,$$

where the covering now is wrt the $L^2(P)$-norm (the intrinsic covariance metric of $\mathbb{G}_P$.)

$\rightarrow$ How does the P-Donsker property of $\mathcal{G}$ relate to the P-pregaussian property??

$\rightarrow$ Well-known sufficient (but not necessary) conditions for the Donsker-property are uniform entropy conditions

$$\sup_Q \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{G}, L^2(Q))} d\varepsilon < \infty$$

and bracketing metric entropy conditions

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{G}, L^2(P))} d\varepsilon < \infty.$$

Examples: s-Hölder/Sobolev balls on $[0,1]^d$, $s > d/2$. Indicators of classes of sets in $\mathbb{R}^d$ of differentiable boundaries, VC-classes, ...etc.

$\rightarrow$ A deep result of Giné and Zinn (1992, AoP) is that any *uniform* in P pregaussian class $\mathcal{G}$ is Donsker (uniformly in P). They actually prove an equivalence between these two (uniform) properties.

$\rightarrow$Balls in Sobolev space W_2^s over $\mathbb{R}^d$ are P-Donsker for any P iff $s > d/2$, and the same is true for the pregaussian property.

$\rightarrow$ However, for *specific* P, there might be quite a substantial gap between the P pregaussian and the P-Donsker property.

$\rightarrow$ For instance for P with a bounded density p, a ball in $B^s_{1\infty}$ with $s \leq 1$ is P-pregaussian (in view of the entropy estimate above), but it is **not** P-Donsker.

$\rightarrow$ Why? On $\mathbb{R}$, such a ball contains all translates of Gamma-densities of parameter $\alpha < 1$, so in particular functions that are unbounded at any point x. The X_i's thus a.s. take values where some g is unbounded.

$\rightarrow$ Since $\mathcal{G}$ is bounded in $L^1(P)$ we conclude

$$\begin{aligned}\|P_n - P\|_{\mathcal{G}} &\geq \|P_n\|_{\mathcal{G}} - \|P\|_{\mathcal{G}} \\ &\geq \|P_n\|_{\mathcal{G}} - \|p\|_\infty \sup_{g\in\mathcal{G}} \|g\|_1 \\ &= \infty \ a.s.\end{aligned}$$

$\rightarrow$ A closely related example we shall study later on is the action of a deconvolution operator on indicators of intervals, i.e., for φ the characteristic function of a probability density

$$\{\mathcal{F}^{-1}[1/\varphi] * 1_{(-\infty,t]} : t \in \mathbb{R}\}.$$

$\rightarrow$ Intuitively speaking, for a class $\mathcal{G}$ to be compact (have finite entropy) in L^2, one may well consider all translates of a fixed unbounded function that decays nicely at $\pm\infty$ and is L^1-Hölder.

$\rightarrow$ On the other hand, for the discrete empirical measure to be well behaved one needs *some kind* of pointwise control of $g \in \mathcal{G}$, as otherwise $\|P_n\|_{\mathcal{G}} = \infty$ a.s. despite $\|P\|_{\mathcal{G}}$ being well-defined. This is, in a certain sense, a pecularity of the fact that we restrict ourselves to P_n as the estimator of our choice, which consists of Dirac point masses.

$\rightarrow$ When assuming that P has a bounded density, restricting to discrete P_n is unnatural, and we may need to consider a smoothed version of P_n instead.

$\rightarrow$ To mind comes the convolution with

$$K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right), \quad K \in L^1(\mathbb{R}^d), \ \int K = 1,$$

to smear out the singularities of the δ_{X_i}'s.

$\rightarrow$ We define the *smoothed empirical measure* $P_n * K_h$ as

$$dP_n * K_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) dx, \ x \in \mathbb{R}^d.$$

$\rightarrow$ We immediately notice that this smoothed empirical measure simply equals the probability measure that has probability density equal to the *kernel density estimator* with kernel K and bandwidth h.

$\rightarrow$ Other smoothed empirical measures come to mind, such as projections of P_n on orthonormal bases, particularly wavelets. The theory there is similar so we restrict to convolution kernels.

$\rightarrow$ When do we have

$$\sqrt{n}(P_n * K_h - P) \rightarrow^d \mathbb{G}_P \text{ in } \ell^\infty(\mathcal{G})?$$

$\rightarrow$ We now have at least two reasons to study such problems:

$\rightarrow$ A) To obtain Donsker-type theorems in the case where $\mathcal{G}$ is P-pregaussian but not P-Donsker.

$\rightarrow$ B) To construct nonparametric confidence sets using the 'plug-in property' ideas laid out above in the Gaussian white noise model.

$\rightarrow$ We shall be mostly interested in the *non*-Donsker case, but let's start simple, with $\mathcal{G}$ a uniformly bounded P-Donsker class.

$\rightarrow$ If $\mathcal{G}$ is P-Donsker then, since

$$P_n * K_h - P = P_n * K_h - P_n + P_n - P$$

it again pays off to study $P_n * K_h - P_n$ first. For $g \in \mathcal{G}$ we have, from Fubini,

$$(P_n * K_h - P_n)g = (P_n - P)(K_h * g - g) + (P * K_h - P)g.$$

$\rightarrow$ In most situations

$$\sup_{g \in \mathcal{G}} \|K_h * g - g\|_{2,P} \rightarrow 0,$$

so that the finite-dimensional distributions of the first process converge to zero.

$\rightarrow$ If $\mathcal{G}$ is translation-invariant, i.e., $g \in \mathcal{G} \Rightarrow g(\cdot - y) \in \mathcal{G}$, then the class

$$\cup_{h>0}\{g * K_h : g \in \mathcal{G}\}$$

is contained in a suitable closure of the symmetric convex hull of $\mathcal{G}$, and thus is itself P-Donsker. As a consequence

$$\{K_h * g - g : g \in \mathcal{G}\}$$

is also contained in a fixed Donsker class. $\sqrt{n}$ times the first process indexed by g is thus asymptotically equicontinuous, from which we conclude

$$\sup_{g \in \mathcal{G}} |(P_n - P)(K_h * g - g)| = o_P(1/\sqrt{n}).$$

$\rightarrow$ The second, 'bias', term is nonrandom, and rewriting it, using Fubini, we have

$$(P * K_h - P)g = \int_{\mathbb{R}} K(u)[p * g(-uh) - p * g(0)]$$

so that the combined smoothness of p, g can (and should) be used.

$\rightarrow$ If $p \in C^1(\mathbb{R}), g \in BV(\mathbb{R})$, then

$$D^2(p * g) = Dp * Dg \in C(\mathbb{R})$$

by distributing one derivative on each convolution product.

$\rightarrow$ By induction, if $p \in C^t$, $g \in BV^s$, the bias term is, by standard Taylor expansions, of order h^{t+s} which for the t-optimal choice of h and $s > 1/2$ is

$$O(n^{-(t+s)/(2t+1)}) = o(n^{-1/2}).$$

In the multidimensional case one needs the constraint $s > d/2$ for this bound.

$\rightarrow$ For example, if $p \in C^t$ and $g = 1_{(-\infty,t]} \in BV$, then this term is of order $h^{t+1} = o(1/\sqrt{n})$, which corresponds to the case where one is estimating the distribution function of F.

$\rightarrow$ In the Sobolev case one proves easily that

$$p \in W_2^t, g \in W_2^s \Rightarrow p * g \in C^{s+t}$$

since

$$\begin{aligned}
\|D^{s+t}(p*g)\|_\infty &\leq \int |u|^t |\widehat{p}(u)||u|^s |\widehat{g}(u)| du \\
&\leq \sqrt{\int |u|^{2t} |\widehat{p}(u)|^2 du \int |u|^{2s} |\widehat{g}(u)|^2 du} \\
&\leq \|p\|_{W_2^t} \|g\|_{W_2^s}.
\end{aligned}$$

$\rightarrow$ The above Fourier-proof generalises to general Besov spaces, using Fourier-multipliers. See Lemma 12 in Giné and N(08).

$\rightarrow$ Thus in these situations

$$\|P * K_h - P_n\|_{\mathcal{G}} = o_P(1/\sqrt{n})$$

and by the above decompositions

$$\sqrt{n}(P_n * K_h - P) \rightarrow^d \mathbb{G}_P \text{ in } \ell^\infty(\mathcal{G}),$$

and, for suitable choice of h and $f \in W_2^t$, also

$$\|\widehat{f}_n - f\|_2 = O_P(n^{-t/(2t+1)})$$

if $\widehat{f}_n$ is the density of $P_n * K_h$.

$\rightarrow$ This includes Sobolev-, Hölder and bounded variation balls when they are P-Donsker.

$\rightarrow$ In particular, if F_n^K is the cumulative distribution function of $P_n * K_h$, then

$$\sqrt{n}(F_n^K(h) - F) \rightarrow \mathbb{G}_P$$

for all bandwidths $h_n \lesssim n^{1/(2t+1)}$. Note that h_n very small is no problem in the case where $\mathcal{G}$ is P-Donsker, as $P_n * K_h$ then only gets closer to P_n.

$\rightarrow$ Translation invariance can be dispensed with but is often natural and provides efficient proofs. To summarise:

Proposition 1 *Let $P_n * K_h$ be the random p.m. from kernel K with bandwidth $h = h_n \to 0$. Let $\mathcal{G}$ be a translation invariant uniformly bounded P-Donsker class on $\mathbb{R}^d$ such that*

$$\sup_{g \in \mathcal{G}} \|K_h * g - g\|_{2,P} \to 0,$$

$$\sup_{g \in \mathcal{G}} \left| \int K(u)[p * g(-uh) - p * g(0)] \right| = o(n^{-1/2})$$

Then

$$\|P_n * K_h - P_n\|_{\mathcal{G}} = o_P(/1\sqrt{n})$$

and in particular

$$\sqrt{n}(P_n * K_h - P) \to^d \mathbb{G}_P \quad in\ \ell^\infty(\mathcal{G}).$$

$\rightarrow$ *Application to Estimating Self-Convolutions:* For $\mathcal{G}_s$ an s-Sobolev ball, and P with unknown density $f \in W_2^s$, the mapping

$$h \mapsto h * f$$

from $\ell^\infty(\mathcal{G}_s)$ to $C(\mathbb{R})$ is continuous since

$$\begin{aligned}
\|h * f\|_\infty &= \sup_x |\int h(x-y) f(y) dy| \\
&\leq \sum_k |\langle h, e_k \rangle| |\langle f, e_k \rangle| \\
&= \sum_k \frac{(1+k^2)^{s/2}}{(1+k^2)^{s/2}} |\langle h, e_k \rangle| |\langle f, e_k \rangle| \\
&\leq \|h\|_{\mathcal{G}} \|f\|_{W_2^s}.
\end{aligned}$$

$\rightarrow$ Using the decomposition

$$f * f - g * g = 2(f - g) * g + (f - g) * (f - g)$$

we conclude that, for $\widehat{f}_n$ the ordinary kernel density estimator of $f \in W_2^s, s > 1/2$,

$$\begin{aligned} \widehat{f}_n * \widehat{f}_n - f * f &= 2(\widehat{f}_n - f) * f + O(\|\widehat{f}_n - f\|_2^2) \\ &= 2(\widehat{f}_n - f) * f + o(1/\sqrt{n}) \end{aligned}$$

so $\sqrt{n}$ times the above converges in law in the space $C(\mathbb{R})$ by the continuos mapping theorem, to a generalised Brownian bridge.

$\rightarrow$ Note that $P_n * P_n$ does not estimate $f * f$ consistently, as it is a.s. a discrete measure.

$\rightarrow$ In other words, the law of the sum

$$X + X' \sim f * f$$

of two independent copies $X, X' \sim f$ can be estimated efficiently at $\sqrt{n}$-rate in sup-norm loss even when the rate for f itself is only of order $\|\hat{f}_n - f\|_2 = O_P(n^{-1/4-\epsilon})$ only, and when P_n cannot be used.

$\rightarrow$ Other applications include confidence sets (as discussed above), or estimation of other integral functionals, such as $f \mapsto \int f^2$.

$\rightarrow$ Using Talagrand's inequality one can prove tight concentration inequalities for

$$\|P_n * K_h - P_n\|_{\mathcal{G}},$$

which combined with concentration inequalities for $\|P_n - P\|_{\mathcal{G}}$ (again Talagrand) then gives inequalities for

$$\|P_n * K_h - P\|_{\mathcal{G}}$$

(by the triangle inequality), see Giné and N (08). For instance one can prove a Dvoretzky-Kiefer-Wolfowitz inequality for cdfs

$$\sqrt{n}\|F_n^K - F\|_\infty.$$

$\rightarrow$ In some sense $\sqrt{n}\|P_n * K_h - P_n\|_{\mathcal{G}}$ is small nonasymptotically for 'nice' P-Donsker $\mathcal{G}$.

$\rightarrow$ For pregaussian classes $\mathcal{G}$ that are NOT P-Donsker, such as the balls in $B^s_{1\infty}$ mentioned above, the situation changes.

$\rightarrow$ In this case we don't want $P_n * K_h$ close to P_n, and have to restrict the speed of convergence to zero of h_n.

$\rightarrow$ Moreover, the non-Donsker case will require deeper techniques.

IV. SMOOTHED EMPIRICAL PROCESSES INDEXED BY NON-DONSKER CLASSES

$\rightarrow$ We now wish to develop a theory for smoothed empirical processes

$$g \mapsto \sqrt{n}(P_n * K_h - P)g, \quad g \in \mathcal{G},$$

where we only assume that $\mathcal{G}$ is P-pregaussian, but not necessarily P-Donsker.

$\rightarrow$ We are interested in finding conditions such that

$$\sqrt{n}(P_n * K_h - P) \rightarrow^d \mathbb{G}_P \quad \text{in } \ell^\infty(\mathcal{G}).$$

Fundamentally it is at first not clear whether this is possible without assuming $\mathcal{G}$ to be P-Donsker.

$\rightarrow$ From a 'limit experiment' point of view we observe that the Gaussian shift limit experiment in $\ell^\infty(\mathcal{G})$ is well defined, but the empirical measure does not give a valid approximation.

$\rightarrow$ We can still hope to find efficient estimators, however.

$\rightarrow$ By Fubini (and for symmetric kernels) it is quite clear that studying $\sqrt{n}(P_n * K_h - P)$ on $\mathcal{G}$ is nothing else than studying the standard empirical process $\sqrt{n}(P_n - P)$ on the classes

$$\tilde{\mathcal{G}}_n = \{g * K_h : g \in \mathcal{G}\}.$$

$\rightarrow$ While for $h \rightarrow 0$ the class $\tilde{\mathcal{G}}_n$ is effectively $\mathcal{G}$, for fixed h the class $\tilde{\mathcal{G}}_n$ is potentially much more regular than $\mathcal{G}$. Sharp control of the increments of $\sqrt{n}(P_n - P)$ on $\tilde{\mathcal{G}}_n$ combined with control of the envelopes of that class allows to outperform non-smooth processes.

$\rightarrow$ Recall that to prove that a random process Z_n in $\ell^\infty(\mathcal{G})$ to converge weakly to some tight random element $Z \in \ell^\infty(\mathcal{G})$ if(f)

a) $(Z_n(g_1), \ldots, Z_n(g_k)) \rightarrow^d (Z(g_1), \ldots, Z(g_k))$ for every finite set $\{g_1, \ldots, g_k\} \subset \mathcal{G}$ and that

b) Z_n is uniformly tight in $\ell^\infty(\mathcal{G}) \iff$ asymptotic equicontinuity of $Z_n \iff Z_n$ concentrates in a compact subset of $\ell^\infty(\mathcal{G})$)

$\rightarrow$ Define increment classes

$$\mathcal{G}'_\delta = \{f - g : f, g \in \mathcal{G}, \|f - g\|_{2,P} \leq \delta\}.$$

Theorem 2 (Giné & N 08, PTRF) *Let $\mathcal{G}$ be any P-pregaussian class of functions on $\mathbb{R}^d$ and let K_h be a sequence of convolution kernels ($h = h_n \to 0$) defined on $\mathbb{R}^d$. Assume that $\mathcal{G} \subseteq L^1(|K_h|)$ $\forall h > 0$ and, in addition,*

1. *For each n, the class*

$$\tilde{\mathcal{G}}_n := \{g * K_h : g \in \mathcal{G}\}$$

has finite envelopes $M_n \geq \sup_{g \in \tilde{\mathcal{G}}_n} \|g\|_\infty$;

2. $\sup_{g \in \mathcal{G}'_\delta} E(g * K_h(X))^2 \leq 4\delta^2$ *for every $\delta >$ 0 and n large enough;*

3. *for i.i.d. Rademacher variables* $(\varepsilon_i)_i$*, independent of the* X_i*'s, we have*

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i g(X_i) \right\|_{(\tilde{\mathcal{G}}_n)'_{1/n^{1/4}}} \to 0$$

as $n \to \infty$ *in outer probability;*

4. $\cup_{n \geq 1} \tilde{\mathcal{G}}_n$ *is in the* $L^2(P)$*-closure of* $\sup_n \|K\|_1$*-times the symmetric convex hull of some fixed* P*-pregaussian class of functions* $\bar{\mathcal{F}}$*.*

5. *For all* $0 < \eta < 1$, *the* $L^2(P)$*-metric entropy of* $\tilde{\mathcal{G}}_n$ *satisfies*

$$H(\tilde{\mathcal{G}}_n, L^2(P), \eta) \leq \lambda_n(\eta)/\eta^2$$

for some $\lambda_n(\eta)$ *such that* $\lambda_n(\eta) \to 0$ *and* $\lambda_n(\eta)/\eta^2 \to \infty$ *as* $\eta \to 0$, *uniformly in* n, *and the bounds* M_n *of part (a) satisfy*

$$M_n \leq \left(5\sqrt{\lambda_n(1/n^{1/4})}\right)^{-1}.$$

Then the random processes

$$\{\sqrt{n}(P_n - P) * K_h(g) : g \in \mathcal{G}\}$$

are uniformly tight in the Banach space $\ell^\infty(\mathcal{G})$.

$\rightarrow$ This result uses fairly deep (but classical) machinery from Giné and Zinn (1984), adapted to the current situation.

$\rightarrow$ One randomises in the asymptotic equicontinuity condition with Rademachers, and uses Gaussian comparison inequalities for such processes. One idea is that a common dominating Gaussian process can be defined on

$$\bigcup_{h>0} \{g * K_h : g \in \mathcal{G}\},$$

since the latter class is contained in the $L^2(P)$-closure of the symmetric convex hull of $\mathcal{G}$.

$\rightarrow$ The bias has been ignored – it is dealt with as in the Donsker case.

$\rightarrow$ As an application, one can show (with some work!) that the conditions of this theorem can be verified for balls in the spaces $B^s_{1\infty}$ when $1/2 < s \leq 1$, thus giving concrete examples for non-Donsker classes for which the smoothed empirical CLT holds. (See Giné & N (2008) for details).

V. Efficient Inference for Inverse Problems I – Deconvolution

$\rightarrow$ A toy problem where the main ideas can be laid out is deconvolution

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

where $X \sim P$ with density f, ε has law f_ε with $\mathcal{F}[f_\varepsilon] \equiv \varphi \neq 0$, so

$$Y \sim P^Y = f * f_\varepsilon.$$

We use a kernel approximation and Plancherel

$$\begin{aligned} K_h * f(x) &= \int K_h(x-y) f(y) dy \\ &= \frac{1}{2\pi} \int \widehat{K_h(\cdot - y)} \overline{\widehat{f}(u)} du \\ &= \frac{1}{2\pi} \int \widehat{K_h(\cdot - y)} \frac{\overline{\widehat{P}^Y(u)}}{\varphi(u)} du \\ &= \bar{K}_h^{\varphi} * P^Y(x) \end{aligned}$$

where, for $\mathcal{F}^{-1}$ the inverse Fourier transform,

$$\bar{K}_h^{\varphi} \equiv \mathcal{F}^{-1} \left[\frac{1}{\varphi(-\cdot)} \right] * K_h.$$

Assume $supp\ (\widehat{K})$ is compact in what follows.

$\rightarrow$ The last expression can be estimated unbiasedly by the deconvolution estimator

$$\frac{1}{n}\sum_{i=1}^{n}\bar{K}_h^{\varphi}(X_i - x),$$

and if we want to estimate the cumulative distribution function F of P then the natural estimate becomes

$$F_n^{\varphi}(t) = \frac{1}{n}\int 1_{(-\infty,t]}(x)\sum_{i=1}^{n}\bar{K}_h^{\varphi}(X_i - x)dx$$

$$= \int 1_{(-\infty,t]} * \mathcal{F}^{-1}\left[\frac{1}{\varphi(-\cdot)}\right](x)dK_h * P_n(x)$$

$\rightarrow$ Conclude that

$$F_n^{\varphi}(t) - F(t) = \int g_t d(K_h * P_n - P), \ t \in \mathbb{R}$$

where, formally

$$g_t \in \mathcal{G} = \left\{ 1_{(-\infty,t]} * \mathcal{F}^{-1} \left[\frac{1}{\varphi(-\cdot)} \right] (x) : t \in \mathbb{R} \right\}$$

so that we are studying a smoothed empirical process on this class of functions.

$\rightarrow$ Now in particular

$$\|F_n^{\varphi} - F\|_{\infty} = \sup_{g \in \mathcal{G}} |P_n * K_h(g) - P(g)|.$$

$\rightarrow$ So proving a Donsker-type theorem for the distribution functions

$$\sqrt{n}(F_n^{\varphi} - F) \rightarrow^d \mathbb{G}^{\varphi} \quad \text{in } \ell^{\infty}(\mathbb{R})$$

is equivalent to proving a uniform central limit theorem for the smoothed process

$$\sqrt{n}(P_n * K_h - P) \rightarrow \mathbb{G} \quad \text{in } \ell^{\infty}(\mathcal{G})$$

on the class of functions

$$\mathcal{G} = \left\{ 1_{(-\infty,t]} * \mathcal{F}^{-1}\left[\frac{1}{\varphi(-\cdot)}\right] : t \in \mathbb{R} \right\}$$

where $\mathcal{F}^{-1}[1/\varphi(-\cdot)]$ is the deconvolution operator.

$\rightarrow$ To deal with these terms one proceeds as in the Lévy setting presented below, with some simplifications. The main message is that

$$\sqrt{n}(F_n^{\varphi} - F) \rightarrow^d \mathbb{G}^{\varphi} \text{ in } \ell^{\infty}(\mathbb{R})$$

as $n \rightarrow \infty$, where the limiting covariance is the usual P-Brownian bridge with

$$1_{(-\infty,t]} \text{ replaced by } 1_{(-\infty,t]} * \mathcal{F}^{-1}\left[\frac{1}{\varphi(-\cdot)}\right].$$

$\rightarrow$ This covariance is the semiparametric Cramér-Rao lower bound for this inverse estimation problem.

$\rightarrow$ See Söhl & Trabs, EJS, 2012, for the deconvolution theory. They consider even more general classes of functionals, not restricted to indicator functions. That $\sqrt{n}$-rates can be obtained pointwise was noted in Dattner, Goldenshluger and Iouditski (2011, AoS), by completely different means.

$\rightarrow$ We now turn to the more complicated Lévy inference case (which was actually treated first).

VI. Efficient Inference for Inverse Problems II – Lévy Processes

$\rightarrow$ A *Lévy process* $L = \{L_t : t \geq 0\}$ is a stochastic process satisfying $L_0 = 0$ a.s. and:
(L1) For any $t_0 < t_1 < \cdots < t_n$ the increments

$$L_{t_0}, L_{t_1} - L_{t_0}, L_{t_2} - L_{t_1}, \ldots, L_{t_n} - L_{t_{n-1}}$$

are independent.
(L2) For every $t \geq 0$ the distribution of $L_{s+t} - L_s$ is independent of s
(L3) L_t is stochastically continuous (and cadlag with prob. one)

$\rightarrow$ Lévy processes can be completely characterised (Lévy-Khintchine formula): Informally any Lévy process can be decomposed into independent components

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

where :

$\rightarrow$ $\sigma > 0, \gamma \in \mathbb{R}$ are parameters,

$\rightarrow$ B_t is a Brownian motion, and

$\rightarrow$ N_t is a pure jump process.

$\rightarrow$ More formally the celebrated Lévy-Khintchine theorem says that the Fourier transform of L_t equals

$$\widehat{\phi}_t(u) = e^{t\psi(u)}$$

with characteristic exponent

$$\psi(u) = -\frac{\sigma^2 u^2}{2} + i\gamma u + \int [e^{iux} - 1 - iux]\nu(dx)$$

and where ν is a Borel measure on $\mathbb{R}$ s.t.

$$\int (|x|^2 \wedge 1)\nu(dx) < \infty, \quad \nu(\{0\}) = 0.$$

$\rightarrow$ ν is called the *Lévy measure* of $(L_t)_{t\geq 0}$.

$\rightarrow$ The parameter $\sigma^2 > 0$ governs the variance of the Brownian component, and $\gamma \in \mathbb{R}$ is the drift of the process.

$\rightarrow$ The Lévy measure ν summarises all the information about the distribution of the jump process part (intensity/rate/size of the jumps occuring).

$\rightarrow$ We call (σ, γ, ν) the *Lévy triplet*.

$\rightarrow$ 'Extreme' Examples: a Poisson process $(0, 0, c\delta_1)$ with $c > 0$ its intensity, or a standard Brownian motion $(1, 0, 0)$.

$\rightarrow$ Another example is the *compound Poisson process* $(0, 0, c\mu)$ obtained from i.i.d. sums

$$S_t = \sum_{i=1}^{N_t} X_i, \quad X_i \sim^{i.i.d.} \mu,$$

with N_t an independent Poisson process with intensity c.

$\rightarrow$ 'Decompounding problem': Estimate μ from observing S_t. For example: Customers arrive according to a Poisson process N_t with
-) random claims worth X_i, $X_i \sim^{iid} \mu$ each.
-) The total claims then equal S_t.

$\rightarrow$ A more involved example: The family of *Gamma processes* with parameters α, λ has Lévy measure

$$\nu(dx) = \alpha x^{-1} e^{-\lambda x} 1_{(0,\infty)}(x) dx$$

which is *not* a finite measure, but it integrates $|x|$ at zero. This includes the exponential distributions with parameter a when $\lambda = 1$.

$\rightarrow$ The Gamma process is an example of a *subordinator process*: jumps are always positive (or negative), the parameters α, λ control the rate of jump arrivals and jump size.

$\rightarrow$ To consider positive and negative jumps simultaneously, introduce the class of self-decomposable processes with Lévy measures

$$\nu(dx) = \frac{k(x)}{|x|} 1_{\mathbb{R}\setminus\{0\}}(x)dx$$

for k unimodal with a jump at the origin, see, e.g., Sato (1999) for more details.

$\rightarrow$ The family of all possible Lévy measures forms a rich nonparametric (that is, *infinite-dimensional*) class of jump distributions.

$\rightarrow$ Can we reconstruct the jump distribution ν from a sampled trajectory of L_t??

Statistical Inference for Lévy Processes

$\rightarrow$ Consider observing a discrete realisation

$$X_{k\Delta},\ k = 1, \ldots, n,$$

of the trajectory of a Lévy process at equally spaced points (times)

$$t_k = k\Delta, \quad k \in \mathbb{N},$$

where $\Delta > 0$.

$\rightarrow$ Δ is the *sampling frequency*.

$\rightarrow$ If $\Delta \rightarrow 0$ we speak of a *high frequency regime*. As $\Delta \rightarrow 0$ we are 'zooming in' to see the fine details of L_t in a given time interval, say $[0, 1]$.

$\rightarrow$ If $\Delta > 0$ is a small but fixed constant we speak of a *low frequency regime*, and observe at times $\{t_k\}_{k=1}^n$. As $n \rightarrow \infty$ the time horizon increases and we see a large sample of increments of L_t.

$\rightarrow$ Both asymptotics can be combined by thinking of $\Delta = \Delta_n \rightarrow 0$ such that $n\Delta_n \rightarrow \infty$.

$\rightarrow$ Our goal is to estimate the jump distribution (Lévy measure) ν from n observed increments $X_1, \ldots, X_n$ of the process. More precisely, we focus on estimating the (generalised) distribution function

$$N(t) = \int_{-\infty}^{t} \nu(dx), \ t < 0,$$

and

$$N(t) = \int_{t}^{\infty} \nu(dx), \ t > 0,$$

of the Lévy measure.

$\rightarrow$ Note that $N(0)$ is not defined in general.

High Frequency Regimes

$\rightarrow$ As soon as $\Delta_n \rightarrow 0$, we notice that, for instance for ν with locally-Lipschitz Lévy density one can show

$$\left| \frac{1}{\Delta} P(X_\Delta \leq t) - N(t) \right| = O(\Delta)$$

as $\Delta \rightarrow 0$.

$\rightarrow$ The first quantity in the last display can be estimated unbiasedly, for $X_1^\Delta, \ldots, X_n^\Delta$ frequency $-$ Δ i.i.d. Lévy increments, by

$$\tilde{N}(t) = \frac{1}{n\Delta} \sum_{k=1}^{n} 1_{(-\infty,t]}(X_k^\Delta)$$

$\rightarrow$ By the above small-time asymptotics we also have the bound $1/\Delta n$ for the variances of $\tilde{N}$, so that, for any $t < 0$,

$$\left|\tilde{N}(t) - N(t)\right| = O_P((\Delta n)^{-1/2} + \Delta).$$

$\rightarrow$ For $\Delta = o(n^{-1/3})$ we get the pointwise rate

$$\left|\tilde{N}(t) - N(t)\right| = O(1/\sqrt{\Delta n}), t < 0,$$

under only a Lipschitz assumption on the Lévy density.

$\rightarrow$ This result can be made uniform in t as well (classical arguments).

$\rightarrow$ In particular the sampling frequency needs to converge to zero fast enough, and Δ fixed cannot be handled.

$\rightarrow$ Can we find more robust procedures for Δ_n not going to zero fast enough? Can we even handle Δ fixed, where intuitively the rate of convergence should be $\sqrt{n\Delta} \simeq \sqrt{n}$?

Low Frequency Asymptotics

$\rightarrow$ In the low frequency regime the observed increments

$$X_k = L_{t_k} - L_{t_{k-1}}$$

are *independent and identically distributed* random variables from a fixed infinitely divisible distribution P, with characteristic function

$$E[\exp(iuX_{t_1})] = E[\exp(iuL_\Delta)] = \phi(u).$$

$\rightarrow$ Small jumps cannot be distinguished from Brownian motion from a fixed sample, Δ fixed!

$\rightarrow$ If a Gaussian component is present in $(L_t)_{t\geq 0}$, the minimax rates of estimation in a discrete observation scheme are only logarithmic in n in any reasonable loss function on Lévy measures.

$\rightarrow$ We thus remove the Gaussian component and restrict to studying inference on pure-jump Lévy processes only.

A Donsker Theorem for Lévy Measures

$\rightarrow$ In fact it will be seen to be necessary to restrict to Lévy processes of finite variation. The Lévy-Khintchine formula then simplifies

$$\phi(u) = E[\exp(iuL_\Delta)] = e^{\Delta\psi(u)}$$

where

$$\frac{\log \phi(u)}{\Delta} = \psi(u) = i\gamma u + \int_{\mathbb{R}\setminus\{0\}} (e^{iux} - 1)\, \nu(dx)$$

$\rightarrow$ Differentiating this identity will identify ν in the Fourier domain.

$\rightarrow$ If ν has a finite moment:

$$\begin{aligned} \frac{1}{i\Delta}\frac{\phi'(u)}{\phi(u)} &= \gamma + \int e^{iux} x\nu(dx) \\ &= \gamma + \mathcal{F}[x\nu](u). \end{aligned}$$

so inverting the Fourier transform, and discarding the drift, the distribution function $N(t) = \int_{-\infty}^{t} d\nu(x)$ of ν formally equals

$$N(t) = \int_{-\infty}^{t} x^{-1}\mathcal{F}^{-1}\left[\frac{1}{i\Delta}\frac{\phi'}{\phi}\right](x)\,dx.$$

$\rightarrow$ In fact since $N(0) = \infty$ in general we need to restrict to $t < 0$ in the above derivation.

$\rightarrow$ Symmetrically we thus wish to estimate

$$N(t) = \int_{-\infty}^{t} \nu(dx),\ t < 0$$

$$N(t) = \int_{t}^{\infty} \nu(dx),\ t > 0.$$

$\rightarrow$ Write

$$P_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k},\ \ \phi_n(u) = \mathcal{F}P_n(u) = \frac{1}{n} \sum_{i=1}^{n} e^{-iX_k u}$$

for the empirical measure and empirical characteristic function, respectively.

Plugging in the empirical c.f. into the identification equation we obtain a natural estimator

$$\widehat{N}_n(t) := \int_{\mathbb{R}} g_t(x) \mathcal{F}^{-1} \left[\frac{1}{i\Delta} \frac{\phi_n'}{\phi_n} \mathcal{F} K_h \right](x)\, dx$$

where K is a band-limited kernel function of polynomial decay,

$$K_h(x) := h^{-1} K(x/h), \quad supp[\mathcal{F}K] \subset [-1, 1],$$

$h > 0$ a bandwidth to be chosen and where

$$g_t(x) := \begin{cases} x^{-1} \mathbf{1}_{(-\infty, t]}(x), & t < 0, \\ x^{-1} \mathbf{1}_{[t, \infty)}(x), & t > 0, \end{cases}.$$

Condition 1 *We require for some* $\epsilon > 0$*:*

1. $\int \max(|x|, |x|^{2+\epsilon})\, \nu(dx) < \infty$;

2. $x\nu$ *has a bounded Lebesgue density and* $|\mathcal{F}[x\nu](u)| \lesssim (1 + |u|)^{-1}$;

3. $(1 + |u|)^{-1+\epsilon}\phi^{-1}(u) \in L^2(\mathbb{R})$.

Proposition 2 *Condition 1 is satisfied for any Lévy process which is an independent sum of processes of the following types:*

1) a compound Poisson process whenever the jump law has a density ν such that $x\nu$ is of bounded variation,

2) a Gamma process with parameters $\alpha \in (0, 1/(2\Delta))$ and $\lambda > 0$,

3) a pure-jump self-decomposable process with k function satisfying $k(0-) + k(0+) < 1/(2\Delta)$,

$\rightarrow$ The parameter constraints in 2) and 3) can be shown to be necessary for $n^{-1/2}$ rates.

$\rightarrow$ *Any polynomial decay* of $|\phi(u)|$ at $\pm\infty$ becomes admissible if Δ is chosen small enough (more below).

$\rightarrow$ For $\zeta > 0$, let

$$\ell_{\zeta}^{\infty} = \ell^{\infty}((-\infty, -\zeta] \cup [\zeta, \infty))$$

be the Banach space of bounded real-valued functions on

$$(-\infty, -\zeta] \cup [\zeta, \infty)$$

equipped with the supremum norm.

$\rightarrow$ Convergence in law in this space, denoted by $\rightarrow^{\mathcal{L}}$, is defined as in Dudley (1999).

Theorem 3 (N and Reiß, JFA, 2012) . *Grant Assumption 1, and let*

$$h_n \sim n^{-1/2}(\log n)^{-\rho}$$

for some $\rho > 1$. *Then, as* $n \to \infty$,

$$\sqrt{n}(\widehat{N}_n - N) \to^{\mathcal{L}} \mathbb{G}^{\varphi} \quad \text{in } \ell^{\infty}_{\zeta},$$

where $\mathbb{G}^{\varphi}$ *is a centered Gaussian Borel random variable in* ℓ^{∞}_{ζ} *with covariance* $\Sigma_{t,s}$ *given by* Δ^{-2} *times*

$$\int_{\mathbb{R}} \left(\mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * [xg_s]\right)\left(\mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * [xg_t]\right) dP.$$

$\rightarrow$ For $s, t < 0$ the above covariance is just

$$\int_{\mathbb{R}} \left(\mathcal{F}^{-1} \left[\frac{1}{\phi(-\cdot)} \right] * 1_{(-\infty, t]} \right) \left(\mathcal{F}^{-1} \left[\frac{1}{\phi(-\cdot)} \right] * 1_{(-\infty, s]} \right) dP$$

$\rightarrow$ This is intuitively appealing when compared to the classical Donsker theorem, where the uncentered covariance is

$$\int_{\mathbb{R}} 1_{(-\infty, t]} 1_{(-\infty, s]} dP.$$

The inverse problem induces the presence of a pseudo-differential operator acting on the standard covariance of the P-Brownian bridge.

$\rightarrow$ The limiting variable can be viewed as the image of the standard P-Brownian bridge under the operator $\mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * [\cdot]$.

$\rightarrow$ For statistical applications: The covariance can easily be estimated by plugging in estimates for ϕ, P, so that Theorem 1 can readily be used for inference.

$\rightarrow$ One can show that this covariance is the Cramér-Rao lower bound in this estimation problem. This is slightly more involved than in the deconvolution case in Söhl and Trabs (2012).

About the Proof:

$\rightarrow$ After suitable approximation steps (bias, linearisation, drift) one reduces the problem to a smoothed empirical process indexed by the class ($t < 0$ for simplicity)

$$\mathcal{G}_\phi := \left\{ \mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * 1_{(-\infty,t]} : t \leq -\zeta \right\}.$$

$\rightarrow$ We apply the general smoothed empirical process theorem from above, and need to first show that $\mathcal{F}_\phi$ is P-pregaussian, and then verify the conditions of the Theorem.

$\rightarrow$ For pregaussianity, note that P is NOT bounded at 0 in general. If

$$\mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * 1_{(-\infty,t]}$$

is supported away from zero, this should not pose a problem, and one should be able to proceed as in the Besov case.

$\rightarrow$ To establish this support property, one needs to establish pseudolocality of the deconvolution operator, meaning that after its application, the singularity at t remains at t.

$\rightarrow$ Some careful analysis of the Lévy structure shows that indeed

$$\mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * 1_{(-\infty,t]}$$

is a function in a Besov space $B_{1\infty}^{1/2+\varepsilon}(\mathbb{R})$ with a singularity only at t.

$\rightarrow$ *Pseudolocality of the 'deconvolution' operator follows from probabilistic properties of the Lévy process only, using Fourier analytical techniques*.

$\rightarrow$ Consider as a representative example the Gamma process case with parameters $(\alpha, 1), \alpha < 1$ such that $\phi(u) = (1 - iu)^{-\alpha}$. Then

$$\begin{aligned}
\mathcal{F}^{-1}\left[\frac{1}{\varphi(-\cdot)}\right] * (\cdot) &= \mathcal{F}^{-1}\left[(1+iu)^{\alpha}\right] * (\cdot) \\
&= \mathcal{F}^{-1}\left[(1+iu)^{\alpha-1}(1+iu)\right] * (\cdot) \\
&= \mathcal{F}^{-1}\left[(1+iu)^{\alpha-1}\right] * (Id - D)(\cdot) \\
&= \gamma_{1-\alpha,1} * (Id - D)(\cdot)
\end{aligned}$$

So the action of this operator on $1_{(-\infty,t]}$ is

$$\mathcal{F}^{-1}\left[\frac{1}{\varphi(-\cdot)}\right] * 1_{(-\infty,t]} = \gamma_{1-\alpha} * 1_{(-\infty,t]} - \gamma_{1-\alpha}(\cdot - t)$$

$\rightarrow$ It remains to check the conditions of Theorem 2 for the pregaussian class

$$\mathcal{G}_\phi := \left\{ \mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * 1_{(-\infty,t]} : t \leq -\zeta \right\}.$$

$\rightarrow$ For this we to split $1_{(-\infty,t]}(x),\ t<0$, into

$1_{(-\infty,t]} = g_t^c + g_t^s,$ where $g_t^c \in W_2^1,\ g_t^s \in L^1 \cap BV,$

e.g., by taking

$$g_t^s = 1_{(-\infty,t]} e^{x-t},$$

and concentrate on the critical part g_t^s.

$$\left\| K_h * \mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * g_t^s \right\|_{BV} \lesssim h_n^{-\alpha}, \quad \text{some } \alpha < \frac{1}{2},$$

which follows from $\mathcal{F}^{-1}\left[1/\phi(-\cdot)\right] * g_t^s \in B_{1\infty}^{1/2+\varepsilon}$, interpolation, and regularity of the kernel K. This controls the envelopes $\|\cdot\|_\infty \lesssim \|\cdot\|_{BV}$ and implies that

$$\left\{ \mathcal{F}^{-1}\left[\frac{1}{\phi(-\cdot)}\right] * g_t^s : t \leq -\zeta \right\}$$

consists, after rewriting, of translates of a fixed BV-function, hence is a VC-class of functions with polynomial covering numbers.

$\rightarrow$ We can thus use bracketing or uniform metric entropy bounds to show that the $n^{-1/4}$-increments of the Rademacher process converge to zero if $h_n^{-\alpha} \lesssim n^{1/4}$. The envelope conditions also follows with

$$M_n \simeq h_n^{-\alpha}.$$

$\rightarrow$ Of the remaining conditions, in Theorem 2, only the last one trading of the enveolopes with covering numbers of $K_h * \mathcal{G}_\varphi$, is difficult.

$\rightarrow$ For this, a sharp polynomial estimate on the $L^2(\mathbb{P})$-covering numbers of $\mathcal{G}_\phi$ is the key.

$\rightarrow$ For Gamma processes, direct arguments can be given. In the general Lévy setting, the tool is the following Fourier multiplier inequality: For $f \in L^2$ supported away from the origin,

$$\|\mathcal{F}^{-1}[\phi^{-1}(-u)] * f\|_{2,P} \lesssim$$

$$\|(1+|u|)^{1-\varepsilon}\mathcal{F}f(u)\|_{L^{2+4/\varepsilon}(\mathbb{R})} + \left(\int \frac{f(y)^2}{1+y^2}\,dy\right)^{1/2},$$

under the condition

$$(1+|u|)^{-1+\epsilon}\phi^{-1}(u) \in L^2(\mathbb{R})$$

from Theorem 3.

$\rightarrow$ How does 'our' estimator perform when $\Delta \rightarrow 0$ with n?

$\rightarrow$ We can (?) prove that the alternative estimator $\widehat{N}$, designed for the low frequency case, also works when $\Delta_n \rightarrow 0$ (arbitrarily slowly), with pointwise rate $\sqrt{n\Delta}$, and under the only assumption of at most polynomial decay of ϕ. [Uniformity probably also..]

$\rightarrow$ The asymptotic covariance then changes to the standard Brownian bridge with $N(t)$ replacing $F(t)$.

$\rightarrow$ Interesting Extensions:

$\rightarrow$ What happens with estimation of $N(t)$ when $t \rightarrow 0$?

$\rightarrow$ What happens for $\Delta_n \rightarrow 0$ very fast in the presence of a Gaussian component?

$\rightarrow$ Abstract limit theorems: instead of cdf $N(t) = \int_{-\infty}^{t} d\nu$ of ν, we can consider

$$\nu(g) = \int g d\nu; \quad f \in \mathcal{G}$$

for abstract classes $\mathcal{G}$.

VII. Weak Limit Theory and Confidence Sets for Likelihood Based Procedures

Nonparametric Maximum Likelihood Estimation (MLE)

$\rightarrow$ Consider data $X_1, \ldots, X_n$ from law P_0 with density p_0 on $[0, 1]$.

$\rightarrow$ The likelihood and log-likelihood for probability densities p are

$$L_n(p) = \prod_{i=1}^{n} p(X_i), \quad \ell_n(p) = \sum_{i=1}^{n} \log p(X_i).$$

$\rightarrow$ One needs to regularise the model for p to define a nonparametric MLE.

$\rightarrow$ Suppose P_0 has a probability density p_0 contained in a Sobolev ball

$$\mathcal{P} = \mathcal{P}(t, \zeta, B) = \left\{ p \geq \zeta > 0, \int_0^1 p = 1, \|p\|_{W_2^t} \leq B \right\}$$

$\rightarrow$ A nonparametric maximum likelihood estimator can be defined as the function $\widehat{p}_n$ satisfying

$$\ell_n(\widehat{p}_n) = \sup_{p \in \mathcal{P}} \ell_n(p),$$

which can be shown to exist as a (random) element in $\mathcal{P}$.

$\rightarrow$ This is the dual problem to a penalised MLE, where one maximises over the whole Sobolev space with a penalty $\lambda \|p\|_{W_2^t}$, λ a fixed penalisation parameter.

$\rightarrow$ The probability density $\widehat{p}_n$ gives rise to a random probability measure $\widehat{P}_n$, and we want to study, as before

$$\sqrt{n}(\widehat{P}_n - P_0) \text{ in the space } \ell^\infty(\mathcal{G}).$$

$\rightarrow$ We shall consider $\mathcal{G}$ equal to a s-Sobolev ball, $s > 1/2$, so a P-Donsker class. It thus makes sense to compare $\widehat{P}_n$ to $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$.

Theorem 4 (N 2007, PTRF.) *Let* p_0 *be an internal point of* $\mathcal{P}$*, and let* $\mathcal{G}_s$ *be a ball in* $W_2^s, s > 1/2$*. We have*

$$\|\widehat{p}_n - p_0\|_2 = O_P(n^{-t/(2t+1)})$$

and

$$\|\hat{P}_n - P_n\|_{\mathcal{G}_s} = o_P(n^{-1/2})$$

so in particular

$$\sqrt{n}(\hat{P}_n - P_0) \to^d \mathbb{G}_P \text{ in } \ell^\infty(\mathcal{G}).$$

$\to$ The same result holds for $\mathcal{G} \equiv \mathcal{G}_{1/2,\delta}$, the 'sharp' 1/2-Sobolev ball with $\log^\delta$-correction.

$\rightarrow$ From this theorem it is natural to construct a confidence set

$$C_n = \left\{ p \in \mathcal{P}_t : \|P - \widehat{P}_n\|_{\mathcal{G}} \leq \frac{z_\alpha}{\sqrt{n}} \right\}$$

with z_α suitable quantiles of $\|\mathbb{G}_P\|_{\mathcal{G}}$.

$\rightarrow$ From the above limit theorem and since $p_0 \in \mathcal{P}$ we have immediately

$$P_0(p_0 \in C_n) = P_0\left(\|P_0 - \widehat{P}_n\|_{\mathcal{G}_s} \leq \frac{z_\alpha}{\sqrt{n}} \right) \rightarrow 1-\alpha$$

as $n \rightarrow \infty$.

$\rightarrow$ We now show that this confidence set is optimal in the sense that its L^2-diameter satisfies

$$|C_n|_2 = O_P\left(n^{-t/(2t+1)} \log^{\delta/2} n\right).$$

To see this, take arbitrary $f, g \in C_n, h = f - g$,

$$\|h\|_2^2 = \sum_{k \in \mathbb{Z}} |\langle h, e_k \rangle|^2$$

where e_k is some orthonormal basis of $L^2([0,1])$ that also generates the Sobolev spaces.

$\rightarrow$ Since $C_n \subset \mathcal{P}$ we know that the high frequencies will be constrained by the Sobolev-condition, for $H_n \sim n^{1/(2t+1)}$,

$$\begin{aligned}\sum_{|k|>H} |\langle h, e_k\rangle|^2 &= \sum_{|k|>H} \frac{(1+k^2)^t}{(1+k^2)^t} |\langle h, e_k\rangle|^2 \\ &\leq \|h\|_{W_2^t} H^{-2t} \\ &= O_P\left(n^{-2t/(2t+1)}\right).\end{aligned}$$

$\rightarrow$ For the low frequencies we can use an interpolation argument, as follows:

$$\begin{aligned}
&\sum_{|k|\leq H} |\langle h, e_k\rangle|^2 \\
&= \sum_{|k|\leq H} \frac{(1+k^2)^{1/2}}{(1+k^2)^{1/2}} \frac{\log(e+k)^\delta}{\log(e+k)^\delta} |\langle h, e_k\rangle|^2 \\
&\lesssim H \log^\delta H \sum_{|k|\leq H} \frac{(1+k^2)^{-1/2}}{\log(e+k)^\delta} |\langle h, e_k\rangle|^2 \\
&\lesssim \|h\|^2_{\mathcal{G}_s} H \log^\delta H \lesssim \frac{H}{n} \log^\delta n \\
&= O_P\left(n^{-2t/(2t+1)} \log^\delta n\right)
\end{aligned}$$

$\rightarrow$ Heuristically these confidence sets can be thought of as including those coefficients

$$|\langle p - \widehat{p}_n, e_k \rangle| \leq \min(k^{-t}B, z_\alpha/\sqrt{n}),$$

so that the bias-variance tradeoff can be seen without the usual undersmoothing problem.

$\rightarrow$ In some sense one constructs the confidence sets from the simultaneous asymptotic distribution of many fixed linear functionals, and intersects this with the smoothness constraint.

Nonparametric Bayes Procedures

$\rightarrow$ We now consider Bayesian procedures from a nonparametric point of view. For simplicity we shall consider the white noise model

$$dX^{(n)} = f(t)dt + \frac{1}{\sqrt{n}}dW(t)$$

which we interpret, as in the first lecture, as a tight Gaussian shift experiment

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

in $\ell^\infty(\mathcal{G})$, where again $\mathcal{G}$ is the critical Sobolev ball ($s = 1/2, \delta > 1$).

$\rightarrow$ We consider a *prior* Π for $f \in L^2$, that is we assume that $\mathbb{X}^{(n)}$ has law f conditional on f being drawn from Π. The posterior distribution is

$$\Pi_n \sim \Pi(\cdot|X^{(n)}) = \Pi(\cdot|\mathbb{X}^{(n)}),$$

a random probability measure in $L^2 \subset \ell^\infty(\mathcal{G})$.

$\rightarrow$ A level $1-\alpha$ credible region of the posterior is any set C_n such that

$$\Pi(C_n|\mathbb{X}^{(n)}) = 1-\alpha.$$

$\rightarrow$ A frequentist question, related to Aad's question of 'uncertainty quantification', is whether such credible regions are frequentist confidence sets, that is, if we assume $\mathbb{X}^{(n)}$ is actually generated by a fixed f_0, whether

$$P_{f_0}(f_0 \in C_n) \rightarrow 1 - \alpha.$$

$\rightarrow$ In finite-dimensional parametric models this is usually true, in view of the Bernstein-von Mises (BvM) theorem: as $n \rightarrow \infty$, the posterior is close to a normal distribution centered at an efficient estimator

$$\|\Pi_n - N(\widehat{\theta}_n, I^{-1}(\theta_0)/n)\|_{TV} \rightarrow^{P^n_{\theta_0}} 0$$

$\rightarrow$ A nonparametric version of the BvM theorem is a delicate subject, see the negative results Freedmann (1999), in particular it cannot happen in L^2-spaces. Likewise, the connection between credible sets and confidence sets is not obvious in these situations, see Cox (1993). Semiparametric BvMs exist, however, see Castillo (2012) for instance.

$\rightarrow$ We show now that it can be obtained in the larger space $\ell^\infty(\mathcal{G})$, and that this can be used to show that *some* posterior credible regions are indeed exact frequentist confidence sets asymptotically.

$\rightarrow$ If we define the pushforward

$$\tau : \theta \mapsto \sqrt{n}(\theta - \widehat{\theta}_n)$$

then the parametric BvM is the same as saying

$$\|\Pi_n \circ \tau^{-1} - N(0, I^{-1}(\theta_0))\|_{TV} \rightarrow^{P^n_{\theta_0}} 0,$$

so in other words the posterior asymptotically looks, on $1/\sqrt{n}$-neighborhoods of $\widehat{\theta}$, like a normal distribution with inverse Fisher information covariance.

$\rightarrow$ As soon as we have $1/\sqrt{n}$-rates in a statistical model such a result can at least be formulated.

$\rightarrow$ The analogue of $N(0, I^{-1}(\theta_0))$ in the infinite-dimensional Gaussian shift experiment

$$\mathbb{X}^{(n)} = f + \frac{1}{\sqrt{n}}\mathbb{W}$$

is the law $\mathcal{N}$ of $\mathbb{W}$. Thus an analogue of the Bernstein-von Mises theorem in this model for, $\tau : f \mapsto \sqrt{n}(f - \mathbb{X}^{(n)})$ (noting that $\mathbb{X}^{(n)}$ is efficient), would be to ask, for the shifted posterior, that

$$\|\Pi(\cdot|\mathbb{X}^{(n)}) \circ \tau^{-1} - \mathcal{N}\|_{TV} \rightarrow^{P_{f_0}} 0$$

as $n \rightarrow \infty$ where $\|\cdot\|_{TV}$ is the total variation of finite signed measures on $\ell^\infty(\mathcal{G})$.

$\rightarrow$ This is a too strong requirement in general. In the conjugate situation where the prior is also Gaussian one can show that such a strong result only holds true if effectively

$$\Pi \prec\succ \mathcal{N}$$

are absolutely continuous, which means that the typical nonparametric priors are ruled out.

$\rightarrow$ The problem comes from the fact that we insist on the total variation distance, so that $\Pi(\cdot|\mathbb{X}^{(n)})$ has to be absolutely continuous to the limiting distribution $\mathcal{N}$.

$\rightarrow$ Instead, we could weaken total variation convergence to weak convergence. Indeed, let β be a (e.g., the bounded Lipschitz) metric for weak convergence of p.m.'s in $\ell^\infty(\mathcal{G})$.

Definition 1 *We then say by definition that* Π_n *satisfies a weak Bernstein von Mises phenomenon in* $\ell^\infty(\mathcal{G})$ *if*

$$\beta(\Pi(\cdot|\mathbb{X}^{(n)}) \circ \tau^{-1}, \mathcal{N}) \rightarrow^{P_{f_0}} 0$$

as $n \rightarrow \infty$.

$\rightarrow$ Unlike in total variation convergence one does not have uniformity in all Borel sets B in

$$|\Pi(\cdot|\mathbb{X}^{(n)}) \circ \tau^{-1}(B) - \mathcal{N}(B)| \rightarrow^{P^n_{f_0}} 0$$

as $n \rightarrow \infty$.

$\rightarrow$ One *does have* uniformity in all sets that have a uniformly smooth boundary for the probability measure $\mathcal{N}$. This includes in particular all norm balls

$$\{B(0,t) : 0 < t \leq M\} \text{ in } \ell^\infty(\mathcal{G}).$$

$\rightarrow$ Consider priors of the form

$$\Pi = \otimes_{lk} \pi_{lk}$$

defined on the coordinates of the orthonormal basis $\{\psi_{lk}\}$ of $L^2([0,1])$, where π_{lk} are probability distributions with Lebesgue density φ_{lk} on the real line.

$\rightarrow$ Further assume, for some fixed density φ

$$\varphi_{lk}(\cdot) = \frac{1}{\sigma_l} \varphi \left(\frac{\cdot}{\sigma_l} \right) \quad \forall k \in \mathcal{Z}_l, \qquad \text{with } \sigma_l > 0.$$

$\rightarrow$ To model γ-smooth functions in a wavelet basis, take $\sigma_l = 2^{-l(\gamma+1/2)}$ and $g_{lk} \sim^{iid} \varphi$,

$$G_\gamma = \sum_{l=J_0}^{\infty} \sum_{k=0}^{2^l-1} 2^{-l(\gamma+1/2)} g_{lk} \psi_{lk}, \quad \gamma > 0.$$

$\rightarrow$ We also allow for standard expansions

$$G_\gamma = \sum_{k \in \mathbb{Z}} (1+k^2)^{\gamma/2} g_k e_k, \quad \gamma > 0,$$

where $\{e_k\}$ is a basis of L^2.

$\rightarrow$ Denote by $\theta_{0,lk} = \langle f_0, \psi_{lk} \rangle$ the 'true coefficients' on the basis.

Condition 2 *Suppose that there exists a finite constant* $M > 0$ *s.t.*

$$\textbf{(P1)} \qquad \sup_{l,k} \frac{|\theta_{0,lk}|}{\sigma_l} \leq M.$$

Suppose also that φ *is bounded and s.t. there exists* $\tau > M$ *and* $0 < c_\varphi$ *with*

$$\textbf{(P2)} \quad \varphi(x) \geq c_\varphi \ \forall x \in (-\tau, \tau), \ \int_{\mathbb{R}} x^2 \varphi(x) dx < \infty.$$

$\rightarrow$ For the wavelet prior above this requires

$$|\theta_{0,lk}| \leq M 2^{-l(\gamma+1/2)} \ \ \forall k, l \iff f_0 \in C^\gamma.$$

Theorem 5 (Castillo and N, 12.) *Any product prior* Π *and* f_0 *satisfying Condition 2 satisfy the weak Bernstein-von Mises phenomenon in* $\ell^\infty(\mathcal{G})$ *for* $\mathcal{G}$ *a critical* $1/2$*-Sobolev ball, i.e.,*

$$\beta(\Pi(\cdot|\mathbb{X}^{(n)}) \circ \tau^{-1}, \mathcal{N}) \to^{P_{f_0}} 0.$$

Moreover the posterior mean $\bar{f}_n$ *is efficient in* $\ell^\infty(\mathcal{G})$ *in the sense that*

$$\|\bar{f}_n - \mathbb{X}^{(n)}\|_{\mathcal{G}} = o_P(1/\sqrt{n}).$$

$\rightarrow$ The proof is not obvious. It uses some duality theory for Sobolev spaces, a BvM for finite-dimensional projections with control of all terms simultaneously in all dimensions, and a contraction result

$$\Pi\left(f : \|f - f_0\|_{\mathcal{G}} > L/\sqrt{n} | \mathbb{X}^{(n)}\right) \rightarrow^{P^n_{f_0}} 0,$$

which can be refined to imply uniform tightness of the posterior measures. This is based on a sharp analysis of the posterior distribution in each coordinate, with dimension-free constants.

Bayesian Credible Sets

$\rightarrow$ A natural $1-\alpha$ credible set is then obtained by solving for $R_n = R_n(\alpha, X^{(n)})$ such that

$$C_n = \left\{ f : \|f - \bar{f}_n\|_{\mathcal{G}} \leq R_n/\sqrt{n} \right\} = 1-\alpha,$$

corresponding to the smallest ball centered at the posterior mean such that the posterior charges it with probability $1-\alpha$.

$\rightarrow$ As above, this credible set should be further intersected with the support of the posterior, using its regularity properties.

$\rightarrow$ Consider first the special case of a uniform wavelet prior Π on L^2 arising from the law of the random wavelet series

$$U_{\gamma,M} = \sum_{l=J_0}^{\infty} \sum_{k=0}^{2^l-1} 2^{-l(\gamma+1/2)} u_{lk} \psi_{lk}(\cdot), \quad \gamma > 0,$$

where the u_{lk} are i.i.d. uniform on $[-M, M]$.

$\rightarrow$ Such priors model functions that lie in a fixed Hölder ball of radius M, with posteriors $\Pi(\cdot|\mathbb{X}^{(n)})$ contracting about f_0 at the L^2-minimax rate $n^{-\gamma/(2\gamma+1)}$ within log factors if $\|f_0\|_{\gamma,\infty} \leq M$. *These posteriors have the plug-in property.*

$\rightarrow$ In this situation it is natural to intersect the credible set C_n with the Hölderian support of the prior (or posterior),

$$C_n' = \left\{ f : \|f\|_{\gamma,\infty} \leq M, \|f - \bar{f}_n\|_{\mathcal{G}} \leq R_n/\sqrt{n} \right\}$$

Obviously $\Pi_n(C_n') = 1 - \alpha$.

Corollary 1 *We have*

$$P_{f_0}^n(f_0 \in C_n') \rightarrow 1 - \alpha$$

as $n \rightarrow \infty$ *and the* L^2*-diameter* $|C_n'|_2$ *of* C_n' *satisfies, for some* $\kappa > 0$,

$$|C_n'|_2 = O_P(n^{-\gamma/(2\gamma+1)}(\log n)^{\kappa}).$$

$\rightarrow$ We consider next the situation of a general series prior Π modeling γ-regular functions, including the important case of Gaussian priors. Let, as above

$$G_\gamma = \sum_{l=J_0}^{\infty} \sum_{k=0}^{2^l-1} 2^{-l(\gamma+1/2)} g_{lk} \psi_{lk}(\cdot), \quad \gamma > 0,$$

where $g_{lk} \sim^{iid} \varphi$. Define

$$\tilde{C}'_n = \left\{ f : \|f\|_{\gamma,2} \leq M_n, \ \|f - \bar{f}_n\|_H \leq R_n/\sqrt{n} \right\},$$

where $M_n \to \infty, \ M_n = O(\log n)$. This parallels the frequentist practice of 'undersmoothing', and can be shown to work.

$\rightarrow$ Alternatively, we intersect with a logarithmically weakened Sobolev ball:

$$C_n'' = \{f : \|f\|_{\gamma,2,1} \leq M_n + 4\delta,\ \|\bar{f}_n - f\|_{\mathcal{G}} \leq R_n/\sqrt{n}\},$$

where M_n is defined as follows: For any n and $\delta_n = (\log n)^{-1/4}$, let M_n equal

$$\inf\left\{M > 0 : \quad \Pi_n(f : |\|f\|_{\gamma,2,1} - M| \leq \delta) \geq 1 - \delta_n\right\},.$$

Corollary 2 *We have*

$$P_{f_0}^n(f_0 \in C_n'') \rightarrow 1-\alpha, \qquad \Pi_n(C_n'') = 1-\alpha+o_P(1)$$

as $n \rightarrow \infty$ *and the* L^2*-diameter* $|C_n''|_2$ *of* C_n'' *satisfies, for some* $\kappa > 0$,

$$|C_n''|_2 = O_P(n^{-\gamma/(2\gamma+1)}(\log n)^{\kappa}).$$

$\rightarrow$ The proofs require the weak Bernstein-von Mises theorem only, by exploiting the uniformity classes for weak convergence towards $\mathcal{N}$ in $\ell^\infty(\mathcal{G})$, and the strong contraction results in L^2 for these priors.

$\rightarrow$ The abstract framework allows for similar BvM results for semiparametric examples, such as all linear functions

$$f \mapsto \int fg, \quad g \in W_2^s,$$

including for instance all moment functionals, but also covers smooth nonlinear functionals such as $\int f^2$.

$\rightarrow$ The above credible sets are the intersection of two Sobolev-type ellipsoids, so do not perhaps reflect precisely the L^2-geometry of a confidence ball or even of a confidence band. No other exact coverage results seem to be known at the moment (!?).

$\rightarrow$ One can also give results in sampling models, but here the situation is substantially more difficult (Castillo & N13, to come..).

$\rightarrow$ Adaptation...? currently open

CONCLUSION

$\rightarrow$ Smoothed empirical processes are at least useful for two purposes:

a) To provide a large enough space in which the exact probabilistic treatment of nonparametric procedures is tractable, so that one can obtain exact confidence sets for these procedures. Combined with interpolation arguments this can give optimal procedures in the more common L^p-type loss functions of nonparametrics.

b) To prove efficient semiparametric results for nonparametric estimators in inverse problems where the plug in based on the empirical measure cannot be used.

$\rightarrow$ The techniques are at the intersection of statistics, probability and functional analysis.

$\rightarrow$ VIII. REFERENCES: Most results and many further references can be found in

I.Castillo, R.Nickl. Nonparametric Bernstein-von Mises theorems. preprint (arxiv.org).

E. Giné, R. Nickl. Uniform central limit theorems for kernel density estimators. *Probability Theory and Related Fields (2008).*

R. Nickl. Donsker-type theorems for nonparametric maximum likelihood estimators. *Probability Theory and Related Fields (2007)*

R. Nickl, M. Reiß. A Donsker theorem for Lévy measures. *Journal of Functional Analysis (2012).*

Bad Belzig Football Game 2013 Goalscorer Table

[1:0 Nickl (Assist: Nickl)]

2:0 own goal (Assist: Reiß)

3:0 Nickl (Assist: van der Vaart)