Minimax theory for a class of non-linear statistical inverse problems

Kolyan Ray
(joint work with Johannes Schmidt-Hieber)

Leiden University

Van Dantzig Seminar
26 February 2016

We consider the following non-linear inverse problem:

$$dY_t = (h \circ Kf)(t)dt + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0,1],$$

where

- h is a known strictly monotone link function,
- K is a known (possibly ill-posed) linear operator,
- W is a standard Brownian motion.

Note that the non-linearity comes from h, which acts pointwise. If h is the identity, we recover the classical linear inverse problem with Gaussian noise.

We will look at several specific choices of h (and K) motivated by statistical applications.

Asymptotic equivalence between two experiments roughly means that there is a model transformation that does not lead to an asymptotic loss of information about the parameter. It can be useful to examine such models since they are often easier to analyse.

Many non-Gaussian statistical inverse problems can be rewritten as

$$dY_t = (h \circ Kf)(t)dt + \frac{1}{\sqrt{n}} dW_t, \quad t \in [0,1],$$

using the notion of asymptotic equivalence.

We study pointwise estimation in such models, which has been studied by numerous authors. We are particularly interested in the case where f takes small (or zero) function values.

Let us first assume that K is the identity for simplicity. We consider the following examples (under certain constraints):

- *Density estimation*: we observe i.i.d. data $X_1, ..., X_n \sim f$.
  Poisson intensity estimation: we observe a Poisson process on $[0, 1]$ with intensity function nf.

  These can both be rewritten with $h(x) = 2\sqrt{x}$ to give

$$dY_t = 2\sqrt{f(t)}dt + n^{-1/2}dW_t.$$

Let us first assume that K is the identity for simplicity. We consider the following examples (under certain constraints):

- *Density estimation*: we observe i.i.d. data $X_1, ..., X_n \sim f$.
  Poisson intensity estimation: we observe a Poisson process on $[0, 1]$ with intensity function nf.

  These can both be rewritten with $h(x) = 2\sqrt{x}$ to give

  $$dY_t = 2\sqrt{f(t)}dt + n^{-1/2}dW_t.$$

- *Binary regression*: we observe n independent Bernoulli random variables with success probability $P(X_i = 1) = f(i/n)$, where $f : [0, 1] \to [0, 1]$ is an unknown regression function.

  This can be rewritten with $h(x) = 2 \arcsin \sqrt{x}$ to give

  $$dY_t = 2 \arcsin \sqrt{f(t)}dt + n^{-1/2}dW_t.$$

- *Spectral density estimation*: we observe a random vector of length n coming from a stationary Gaussian distribution with spectral density f.
  Gaussian variance estimation: We observe $X_1, ..., X_n$ independent with $X_i \sim N(0, f(i/n)^2)$, where $f \geq 0$ is unknown.

  This can be rewritten with $h(x) = 2^{-1/2} \log x$ to give

$$dY_t = \frac{1}{\sqrt{2}} \log f(t) dt + n^{-1/2} dW_t.$$

The choice of h is linked to the *variance stabilizing transformation* of the model.

The linear operator K is typically an ill-posed operator (not continuously invertible). Perhaps the two most common examples for $h(x) = 2\sqrt{x}$ are:

- *Density deconvolution*: we observe data $X_1 + \epsilon_1, ..., X_n + \epsilon_n$, where $X_i \sim f$ and $\epsilon_i \sim g$ for g a known density.
- *Poisson intensity estimation*: K is typically a convolution operator modelling the blurring of images by a so-called point spread function. The 2-dimensional version of this problem has applications in photonic imaging.

In both cases we have $Kf(t) = f * g(t)$ for some known g, giving

$$dY_t = 2\sqrt{f * g(t)}dt + n^{-1/2}dW_t.$$

We will discuss the case $h(x) = 2\sqrt{x}$ (density esitmation, Poisson intensity estimation). The other cases are similar.

What happens if we assign f classical Hölder smoothness C^β?

If $f \in C^\beta$ then $\sqrt{f} \in C^{\beta/2}$ for $\beta \leq 2$.

We will discuss the case $h(x) = 2\sqrt{x}$ (density esitmation, Poisson intensity estimation). The other cases are similar.

What happens if we assign f classical Hölder smoothness C^β?

If $f \in C^\beta$ then $\sqrt{f} \in C^{\beta/2}$ for $\beta \leq 2$.

Theorem (Bony et al. (2006))

There exists a function $f \in C^\infty$ such that $\sqrt{f} \notin C^\beta$ for any $\beta > 1$.

So we cannot exploit higher order Hölder regularity beyond $\beta = 2$. The problem arises due to very small non-zero function values, where the derivatives of $\sqrt{f}$ can fluctuate greatly.

We propose an alternative restricted space:

$$\mathcal{H}^\beta = \{f \in C^\beta \ : \ f \geq 0, \ \|f\|_{\mathcal{H}^\beta} := \|f\|_{C^\beta} + |f|_{\mathcal{H}^\beta} < \infty\},$$

where $\|\cdot\|_{C^\beta}$ is the usual Hölder norm and

$$|f|_{\mathcal{H}^\beta} = \max_{1 \leq j < \beta} \left(\sup_{x \in [0,1]} \frac{|f^{(j)}(x)|^\beta}{|f(x)|^{\beta - j}} \right)^{1/j} = \max_{1 \leq j < \beta} \left\| |f^{(j)}|^\beta / |f|^{\beta - j} \right\|_\infty^{1/j}$$

is a seminorm ($|f|_{\mathcal{H}^\beta} = 0$ for $\beta \leq 1$).

The quantity $|f|_{\mathcal{H}^\beta}$ measures the flatness of a function near 0 in the sense that if $f(x)$ is small then the derivatives of f must also be small in a neighbourhood of x. This can be thought of as a shape constraint.

$\mathcal{H}^\beta$ contains

- all C^β functions uniformly bounded away from 0 (the typical assumption for such problems),
- functions that take small values in a 'controlled' way, e.g. $(x - x_0)^\beta g(x)$ for $g \geq \varepsilon > 0$ in C^∞.

Theorem

If $f \in \mathcal{H}^\beta$ then $\sqrt{f} \in \mathcal{H}^{\beta/2}$ for **all** $\beta \geq 0$.

In fact, it turns out that $\mathcal{H}^\beta = C^\beta$ for $0 < \beta \leq 2$ (hence why the relation holds for C^β).

We propose a two-stage procedure:

1. Let $[h(Kf)]_{HT}$ denote the hard wavelet thresholding estimator of $h(Kf)$. Estimate Kf by the estimator

$$\widehat{Kf} = h^{-1}([h(Kf)]_{HT})$$

(recall that h is injective). Using this we have access to

$$\widehat{Kf}(t) = Kf(t) + \delta(t),$$

where $\delta(t)$ is the noise level (which is the minimax rate with high probability).

2. Treat the above as a deterministic inverse problem with noise level δ. Solve this for f using classical methods (e.g. Tikhonov regularization, Bayesian methods, etc.)

For the noise level δ (step 1), without loss of generality set $K = id$. We consider a pointwise function-dependent rate for $f \in \mathcal{H}^\beta$:

$$r_{n,\beta}(f(x)) = \left(\frac{\log n}{n}\right)^{\frac{\beta}{\beta+1}} \vee \left(f(x)\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}.$$

Theorem

The estimator $\widehat{f} = h^{-1}([h(f)]_{HT})$ *satisfies*

$$\mathbb{P}_f\left(\sup_{x\in[0,1]} \frac{|\widehat{f}(x) - f(x)|}{r_{n,\beta}(f(x))} \leq C\right) \geq 1 - n^{-C'},$$

uniformly over $\cup_{\beta,R}\{f : \|f\|_{\mathcal{H}^\beta} \leq R\}$ *(β, R in compact sets).*

The estimator adapts to $\mathcal{H}^\beta$-smoothness and local function size uniformly over $x \in [0,1]$.

$$r_{n,\beta}(f(x)) = \left(\frac{\log n}{n}\right)^{\frac{\beta}{\beta+1}} \vee \left(f(x)\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$$

- The $\log n$-factors are needed for adaptive estimation in pointwise estimation (as usual).
- For $f(x) \gtrsim (\log n/n)^{\frac{\beta}{\beta+1}}$ we recover the usual nonparametric rate, albeit with pointwise dependence on the radius.
- For $f(x) \lesssim (\log n/n)^{\frac{\beta}{\beta+1}}$, we have faster than $n^{-1/2}$ rates for $\beta > 1$, i.e. superefficiency.
- For small function values: variance $\gg$ bias.
- Related to irregular models: similar to nonparametric regression with one-sided errors, e.g. Jirak et al. (2014).
- The smaller regime is caused by the non-linearity of $h(x) = \sqrt{x}$ near 0.

The same rate has recently (and independently) been proved directly in the case of density estimation by Patschkowski and Rohde (2016) for $0 < \beta \leq 2$. They consider classical Hölder smoothness C^β, which is why they get stuck at $\beta = 2$.

Suppose we take $f(x) = (x - 1/2)^2$. Then $f \in C^\infty([0,1]) \cap \mathcal{H}^2$, but $f \notin \mathcal{H}^\beta$ for any $\beta > 2$. Intuitively, we see that

$$h(f(x)) = \sqrt{f(x)} = |x - 1/2|$$

is C^1, but no more regular. We recover rate based on this smoothness, which corresponds to $\beta/2 = 1$, but not faster. This corresponds to the correct flatness condition.

We have more precise examples of such lower bounds, but they are not as intuitive.

Derivation of upper bound relies on careful analysis of (local) smoothness of $h \circ f$. Use resulting wavelet bounds and usual wavelet thresholding proof to obtain the result.

We have the corresponding lower bound (without $\log n$ factors):

Theorem

For any $\beta > 0$, $R > 0$, $x_0 \in [0,1]$ and any sequence $(f_n^*)_n$ with $\limsup_{n\to\infty} \|f_n^*\|_{\mathcal{H}^\beta} < R$,

$$\liminf_{n\to\infty} \inf_{\hat{f}_n(x_0)} \sup_{\substack{f:\|f\|_{\mathcal{H}^\beta}\le R \\ KL(f,f_n^*)\le 1}} \mathbb{P}_f \left(\frac{|\hat{f}_n(x_0) - f(x_0)|}{r_{n,\beta}(f(x_0))} \ge C \right) > 0,$$

where the infimum is taken over all measurable estimators of $f(x_0)$.

We have replaced the whole parameter space $\{f : \|f\|_{\mathcal{H}^\beta} \leq R\}$ with local parameter spaces

$$\{f : \|f\|_{\mathcal{H}^\beta} \leq R, \quad \mathrm{KL}(f, f_n^*) \leq 1\}$$

about every *interior* point $f_n^* \in \mathcal{H}^\beta$. This allows us to obtain local (function-dependent) rates.

Global rate: *somewhere* on the parameter space, the estimation rate can not be improved.

Local rate: the estimation rate can not be improved on a local neighbourhood of *any* point in the parameter space.

For example consider (f_n^*) with $f_n^*(x_0) \to 0$. The minimax lower bound over an $\mathcal{H}^\beta$-ball is $n^{-\frac{\beta}{2\beta+1}}$, while our upper bound gives faster rates (e.g. $n^{-\frac{\beta}{\beta+1}}$). The matching lower bound works since we restrict to the smaller spaces: the local parameter space $\{f \in \mathcal{H}^\beta : \mathrm{KL}(f, f_n^*) \leq 1\}$ also contains only functions vanishing at x_0 for large n.

For example consider (f_n^*) with $f_n^*(x_0) \to 0$. The minimax lower bound over an $\mathcal{H}^\beta$-ball is $n^{-\frac{\beta}{2\beta+1}}$, while our upper bound gives faster rates (e.g. $n^{-\frac{\beta}{\beta+1}}$). The matching lower bound works since we restrict to the smaller spaces: the local parameter space $\{f \in \mathcal{H}^\beta : \mathrm{KL}(f, f_n^*) \leq 1\}$ also contains only functions vanishing at x_0 for large n.

The lower bounds also give insight into the form of the rates. For $h(x) = \sqrt{x}$, the Kullback-Leibler divergence equals

$$\mathrm{KL}(f, g) = \frac{n}{2} \int (\sqrt{f} - \sqrt{g})^2.$$

- If functions are uniformly bounded away from 0 this behaves like the L^2 distance $\implies$ classic nonparametric rate.
- If functions are near 0, behaves likes L^1 distance $\implies$ rate for irregular models.

We have now access to $\widehat{Kf}$ which satisfies

$$\widehat{Kf}(t) = Kf(t) + \delta(t)$$

with high probability, where $|\delta(t)| = r_{n,\beta}(Kf(t))$. We solve this deterministic inverse problem using classical methods (e.g. Tikhonov regularization).

The rate depends on the noise level δ, which we need to know to obtain rate-optimal procedures. However, we can use a plug-in estimate to estimate the noise level.

Theorem

$$C^{-1} r_{n,\beta}(\widehat{Kf}(t)) \leq r_{n,\beta}(Kf(t)) \leq \overline{C} r_{n,\beta}(\widehat{Kf}(t))$$

with high probability and uniformly over $t \in [0,1]$.

Similar results hold in the other cases.

- *Binary regression*:

$$dY_t = 2 \arcsin \sqrt{f(t)} dt + n^{-1/2} dW_t,$$

$$r_{n,\beta}(f(x)) = \left(\frac{\log n}{n} \right)^{\frac{\beta}{\beta+1}} \vee \left(f(x)(1-f(x)) \frac{\log n}{n} \right)^{\frac{\beta}{2\beta+1}}.$$

- *Spectral density estimation*:

$$dY_t = \frac{1}{\sqrt{2}} \log f(t) dt + n^{-1/2} dW_t,$$

$$r_{n,\beta}(f(x)) = f(x) \wedge (f(x)^2/n)^{\frac{\beta}{2\beta+1}}$$

(the last rate up to some subpolynomial factors).