Support Inference in linear Statistical Inverse Problems

Frank Werner[1]

Statistical Inverse Problems in Biophysics Group
Max Planck Institute for Biophysical Chemistry, Göttingen

and

Felix Bernstein Institute for Mathematical Statistics in the Biosciences
University of Göttingen

Chemnitz Symposium on Inverse Problems 2016

[1]joint work with Katharina Proksch and Axel Munk

Outline

Outline

Setting and goal

- $\mathcal{X}$ and $\mathcal{Y}$ Hilbert spaces
- $\{\varphi_i\}_{i\in\mathbb{N}} \subset \mathcal{X}$ a dictionary
- $T : \mathcal{X} \to \mathcal{Y}$ bounded and linear

Can we identify the active components ($\hat{=}$'support') of an unknown f from noisy measurements of Tf?

More precisely:

Given noisy measurements $Y \approx Tf$ we want to generate a set $\mathcal{I}_{\mathrm{a}}$ such that at a controlled error level all $i \in \mathcal{I}_{\mathrm{a}}$ satisfy $\langle \varphi_i, f \rangle_{\mathcal{X}} > 0$

Support inference?

Special situation: Suppose that

- $\mathcal{X}$ is a space of functions (i.e. $\mathcal{X} = \mathbf{L}^2(\Omega)$),
- and the functions φ_i have compact support (e.g. wavelet dictionary)

Then:

$$\langle \varphi_i, f \rangle_{\mathcal{X}} > 0 \qquad \Rightarrow \qquad f_{|_{\mathrm{supp}(\varphi_i)}} \not\equiv 0$$

Consequently, we obtain information about the support of f!

Related problems and methods

Suppose f is sparse w.r.t. $\{\varphi_i\}_{i\in\mathbb{N}}$. Then for the recovery of f many methods are used:

- ℓ^1-penalized Tikhonov regularization / LASSO

$$\hat{f}_\alpha = \operatorname*{argmin}_{f\in\mathcal{X}} \left[\|Tf - Y\|_{\mathcal{Y}}^2 + \alpha \sum_{i=1}^{\infty} |\langle \varphi_i, f\rangle_{\mathcal{X}}|\right]$$

- Residual method / Danzig selector

$$\hat{f}_\alpha = \operatorname*{argmin}_{f\in\mathcal{X}} \sum_{i=1}^{\infty} |\langle \varphi_i, f\rangle_{\mathcal{X}}| \quad \text{subject to} \quad \|Tf - Y\|_{\mathcal{Y}} \leq \rho$$

- Orthorgonal matching pursuit
- ...

But none of these methods can identify the true support at a controlled error level!

Methodology

Inference for a single i:

- compute a function Φ_i such that $\langle \Phi_i, Tf \rangle_{\mathcal{Y}} = \langle \varphi_i, f \rangle_{\mathcal{X}}$, i.e. $\varphi_i = T^* \Phi_i$
- compute the (asymptotic) distribution of $\langle \Phi_i, Y \rangle_{\mathcal{Y}}$ (test statistic) under the hypothesis $\langle \varphi_i, f \rangle_{\mathcal{X}} = 0$
- with the $(1-\alpha)$-quantile $q_{1-\alpha}$ of this (asymptotic) distribution it holds under $\langle \varphi_i, f \rangle_{\mathcal{X}} = 0$ (asymptotically)

$$\mathbb{P}\left[\langle \Phi_i, Y \rangle_{\mathcal{Y}} \geq q_{1-\alpha}\right] \leq \alpha$$

- consequently if $\langle \Phi_i, Y \rangle_{\mathcal{Y}} \geq q_{1-\alpha}$ we have $\langle \varphi_i, f \rangle_{\mathcal{X}} > 0$ with probability $\geq 1 - \alpha$.

Methodology (cont')

Inference for a all i:

- if we infer for each i individually, the multiplicity adjustment makes the statements weak (if the statements are true for each single i with probability 90%, then they hold true for two i at the same time only with probability 81% etc.)

$\rightsquigarrow$ remedy: simultaneous testing

- consider the test statistic

$$M = \max_i \left[w_i \left(\frac{\langle \Phi_i, Y \rangle_{\mathcal{Y}}}{\sqrt{\mathbb{V}\left[\langle \Phi_i, Y \rangle_{\mathcal{Y}}\right]}} - w_i \right) \right].$$

- compute the asymptotic distribution of M under $f \equiv 0$ and its $(1-\alpha)$-quantile $q_{1-\alpha}$
- mark each i for which $\langle \Phi_i, Y \rangle_{\mathcal{Y}} > (q_{1-\alpha}/w_i + w_i) \sqrt{\mathbb{V}\left[\langle \Phi_i, Y \rangle_{\mathcal{Y}}\right]}$ as active

Methodology (cont')

- by taking the max in the statistic the statements become uniform in i:

$$\mathcal{I}_a := \left\{ i \;\middle|\; \langle \Phi_i, Y \rangle_{\mathcal{Y}} > \left(\frac{q_{1-\alpha}}{w_i} + w_i \right) \sqrt{\mathbb{V}\left[\langle \Phi_i, Y \rangle_{\mathcal{Y}} \right]} \right\}$$

satisfies (asymptotically)

$$\mathbb{P}\left[\langle \varphi_i, f \rangle_{\mathcal{X}} > 0 \text{ for all } i \in \mathcal{I}_a \right] \geq 1 - \alpha.$$

$\rightsquigarrow$ we can infer on all i at a controlled level!

Are the quantiles $q_{1-\alpha}$ well-defined? How to compute them?

Outline

Specific setting

- $\mathcal{Y} = L^2([0,1]^d)$, discrete measurements: $Y_j = (Tf)(x_j) + \xi_j$, $j \in \{1, ..., n\}^d$
- ξ_j are independent, centered ($\mathbb{E}[\xi_j] = 0$) and satisfy a moment condition (especially all moments need to exist)
- the dictionary has at most $N = N(n)$ elements φ_i which satisfy $\varphi_i = T^*\Phi_i$, and there is a transformed mother wavelet Φ such that

$$\{\Phi_i\} = \left\{\Phi\left(\frac{\cdot - t_i}{h_i}\right) \;\middle|\; 1 \leq i \leq N(n)\right\}$$

  with **scales** $h_i \in [0,1]^d$ and **positions** $t_i \in [0,1]^d$
- the function Φ has compact support in $[0,1]^d$

Test statistic

- $\langle \Phi_i, Y \rangle_{\mathcal{Y}}$ is not available, approximate it by

$$\langle \Phi_i, Y \rangle_n := n^{-d} \sum_{j \in \{1,\dots,n\}^d} Y_j \Phi_i(x_j)$$

- the variance $\sigma_i^2 := \mathbb{V}\left[\langle \Phi_i, Y \rangle_{\mathcal{Y}}\right]$ might also be unknown, consider a family of uniformly consistent estimators $\hat{\sigma}_i^2$
- we have to investigate the asymptotic distribution of

$$\mathcal{S}(Y) := \max_i \left[w_i \left(\frac{\langle \Phi_i, Y \rangle_n}{\hat{\sigma}_i} - w_i \right) \right]$$

- the calibration values are only scale-dependent and chosen as

$$w_i = \sqrt{2 \log \left(\frac{C_{\Phi_i}}{\prod h_i} \right)} + C_d \frac{\log \left(\sqrt{2 \log \left(\frac{C_{\Phi_i}}{\prod h_i} \right)} \right)}{\sqrt{2 \log \left(\frac{C_{\Phi_i}}{\prod h_i} \right)}}$$

Gaussian Approximation

Suppose that

- there are only polynomially many probe functions in the dictionary ($N(n) \leqslant n^{\kappa}$ for some $\kappa > 0$)
- the smallest scale tends to zero not too fast ($\min_i \min_{\text{entries}} h_i \geqslant \log(n)^p/n$ with some specific p)
- the largest scale tends to zero sufficiently fast ($\max_i \max_{\text{entries}} h_i \leqslant n^{-\delta}$ for some $\delta > 0$)

Gaussian Approximation

Then there are i.i.d. standard normal random variables ζ_j such that under $f \equiv 0$ it holds

$$\lim_{n\to\infty} \left|\mathbb{P}\left[\mathcal{S}\left(Y\right) > q\right] - \mathbb{P}\left[\mathcal{S}\left(\zeta\right) > q\right]\right| = 0 \qquad \text{for all } q$$

Gaussian Approximation (cont')

Continuous Gaussian Approximation

There exists a Brownian sheet W such that $\mathcal{S}(Y)$ can be approximated

$$\mathcal{S}(W) := \max_i \left[w_i \left(\frac{\int \Phi_i(x) \, \mathrm{d}W_x}{\|\Phi_i\|_{L^2}} - w_i \right) \right],$$

i.e. under $f \equiv 0$ it holds

$$\lim_{n \to \infty} \left| \mathbb{P}[\mathcal{S}(Y) > q] - \mathbb{P}[\mathcal{S}(W) > q] \right| = 0 \qquad \text{for all } q.$$

Moreover

- $\mathcal{S}(W)$ is a.s. bounded from below and above,
- $\mathcal{S}(W)$ is asymptotically non-degenerate, i.e. does not concentrate to any point,
- $\mathcal{S}(W)$ does not depend on any unknown quantities.

Gaussian Approximation - Implications

This means that we can use quantiles $q_{1-\alpha}$ from

$$\mathcal{S}(\zeta) = \max_i \left[w_i \left(\frac{\langle \Phi_i, \zeta \rangle_n}{\|\Phi_i\|_2} - w_i \right) \right]$$

to hold the asymptotic level.

- $\mathcal{S}(\zeta)$ is 'distribution free', i.e. it depends only on known quantities

$\rightsquigarrow$ quantiles can be simulated easily

- quantiles are meaningful as $n \to \infty$

If $\langle \Phi_i, Y \rangle_n > \left(\frac{q_{1-\alpha}}{w_i} + w_i \right) \hat{\sigma}_i$ mark i as active (i.e. $i \in \mathcal{I}_a$)!

Detection properties

So far: whenever $i \in \mathcal{I}_a$, then asymptotically $\mathbb{P}\left[\langle \varphi_i, f \rangle_{\mathcal{X}} > 0\right] \geq 1 - \alpha$.

But how large must $\langle \varphi_i, f \rangle_{\mathcal{X}}$ be to be detected?

Lower detection bound

If $\langle \varphi_i, f \rangle_{\mathcal{X}} \geq 2 \left(\frac{q_{1-\alpha}}{w_i} + w_i \right) \sigma_i$, then

$$\lim_{n \to \infty} \mathbb{P}\left[i \in \mathcal{I}_a\right] \geq 1 - \alpha$$

uniformly in i.

Special case: deconvolution

- Suppose $d = 2$ and T is a convolution operator, i.e.

$$(Tf)(y) = (k * f)(y) := \int_{[0,1]^2} k(x-y) f(y) \, \mathrm{d}y.$$

$\rightsquigarrow$ This implies that if we choose $\{\varphi_i\}$ of Wavelet-type we obtain

$$\{\Phi_i\} = \left\{ \tilde{\Phi}_{h_i} \left(\frac{\cdot - t_i}{h_i} \right) \;\middle|\; 1 \leq i \leq N(n) \right\}$$

where $\tilde{\Phi}_h$ depends on h.

- Suppose the Fourier transform of the kernel k has a polynomial decay, i.e.

$$\underline{c}\left(1 + \|\xi\|_2^2\right)^{-a} \leq |\mathcal{F}k(\xi)| \leq \overline{C}\left(1 + \|\xi\|_2^2\right)^{-a}.$$

$\rightsquigarrow$ This implies that the functions φ_i can be chosen such that they are non-negative and have compact support, and $\|\tilde{\Phi}_h\|_{L^2}$ behaves like $\max_{\text{entries}} h^{2a}$.

Special case: deconvolution (cont')

- $\mathcal{S}(Y)$ can be approximated by a Gaussian version which is a.s. bounded and non-degenerate

$\rightsquigarrow$ whenever $i \in \mathcal{I}_a$, then asymptotically $\mathbb{P}\left[\langle \varphi_i, f\rangle_{\mathcal{X}} > 0\right] \geq 1 - \alpha$.

- If $\langle \varphi_i, f\rangle_{\mathcal{X}}$ is sufficiently large, then i will be detected with probability $\geq 1 - \alpha$.

Optimality of the lower detection bound

In $d = 1$ this lower detection bound is optimal in the sense that no estimator for a β-Hölder-continuous function f can distinguish between $f_{|_{[t,t+h]}} = 0$ and $f_{|_{[t,t+h]}} \geq h^\beta$ at a faster rate in $h = h(n)$.

Outline

Considered problem

- T deconvolution problem, i.e.

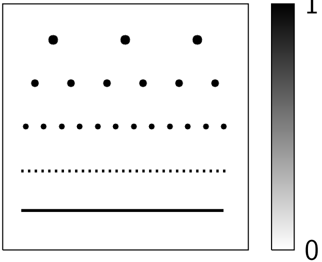
$$Y_j = (k * f)(x_j) + \xi_j, \qquad j \in \{1, ..., n\}^2$$

with an equidistant grid $\{x_j\}$ on $[0,1]^2$.

- The kernel k is chosen from the family

$$(\mathcal{F}k_{a,b})(\xi) = \left(1 + b^2 \|\xi\|_2^2\right)^{-a}, \qquad \xi \in \mathbb{R}^2.$$

- The variance is considered to be known.
- The mother wavelet φ is chosen to minimize the variance $\|\Phi\|_{L^2}^2$ ($\rightsquigarrow$ Tensor product of Beta-Kernels)

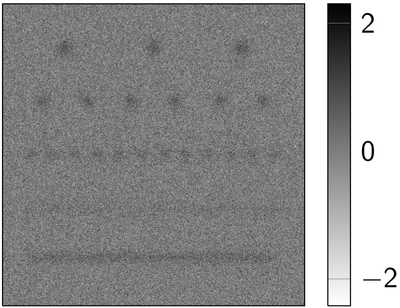


Testfunction f

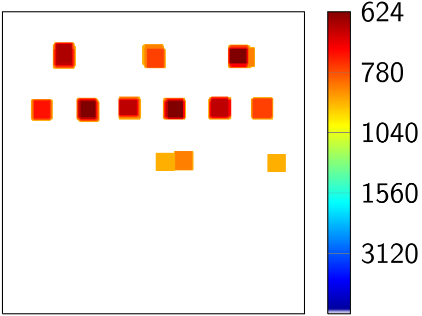
Some empirical levels for $\alpha = 0.1$

Noise scenario	Parameters	false positives %
Gaussian noise		8.8
Student's t noise	$\nu = 3$	100
	$\nu = 6$	94.7
	$\nu = 7$	72.3
	$\nu = 11$	21.8
	$\nu = 15$	15.7
	$\nu = 19$	13.0
	$\nu = 23$	13.3
CCD noise (Sneyder '93, '95):	$t = 100,\ b = 0.5,\ \sigma = 0.01$	9.8
obs. time t, background b,	$t = 1000, b = 0.005, \sigma = 0.01$	8.1
read-out errors $\mathcal{N}\left(0, \sigma^2\right)$	$t = 100,\ b = 0.005, \sigma = 0.01$	14.5

Support recovery - result

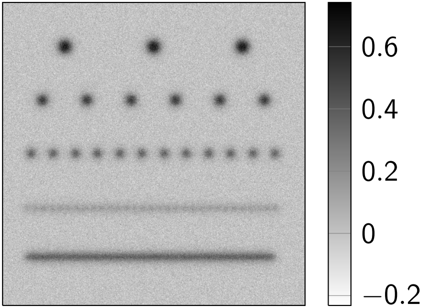


data ($\sigma = 0.5$)

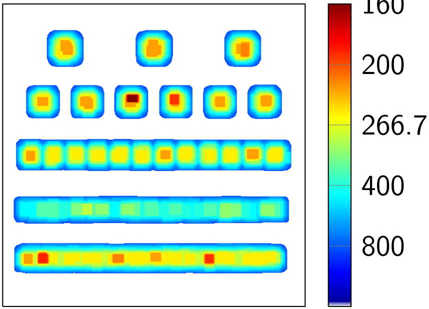


90% significance map

Support recovery - result

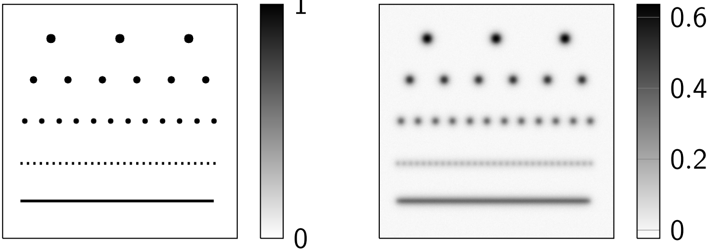


data ($\sigma = 0.05$)

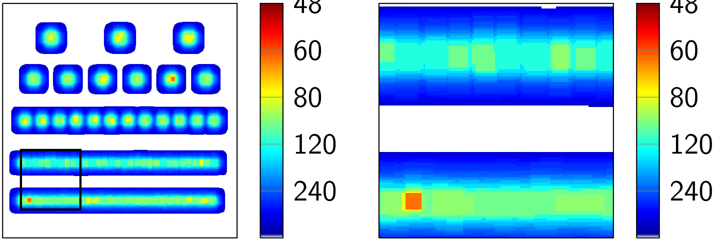


90% significance map

Support recovery - result



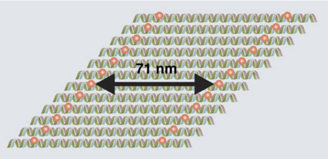
exact solution   data ($\sigma = 0.005$)

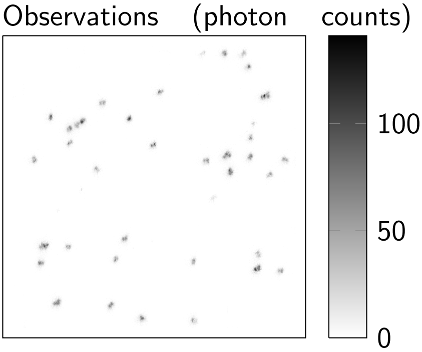


90% significance map   zoom

Real data example - Setup

- we analyze fluorescent dyes on single DNA Origami
- imaging is performed by a STED microscope
- each of the two strands can at most hold 12 markers



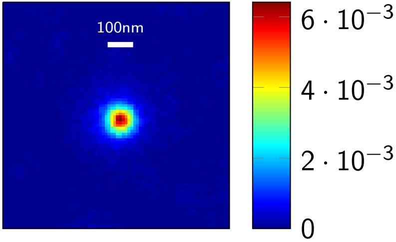


Modeling

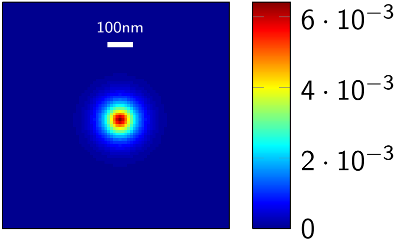
The observations can perfectly modeled by

$$Y_j \overset{\text{independent}}{\sim} \mathrm{Bin}\left(t, (k * f)(x_j)\right), \qquad j \in \{1, ..., n\}^2 .$$

- $\mathrm{Bin}(t, p)$: Binomial distribution with parameters $t \in \mathbb{N}$ and $p \in [0, 1]$
- $f(x)$: probability that a photon emitted at grid point x is recorded at the detector in a single excitation pulse

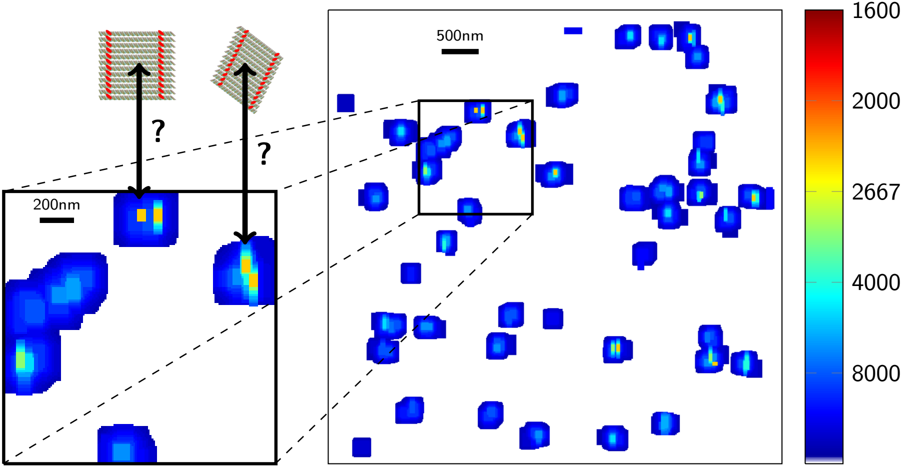


experimental kernel

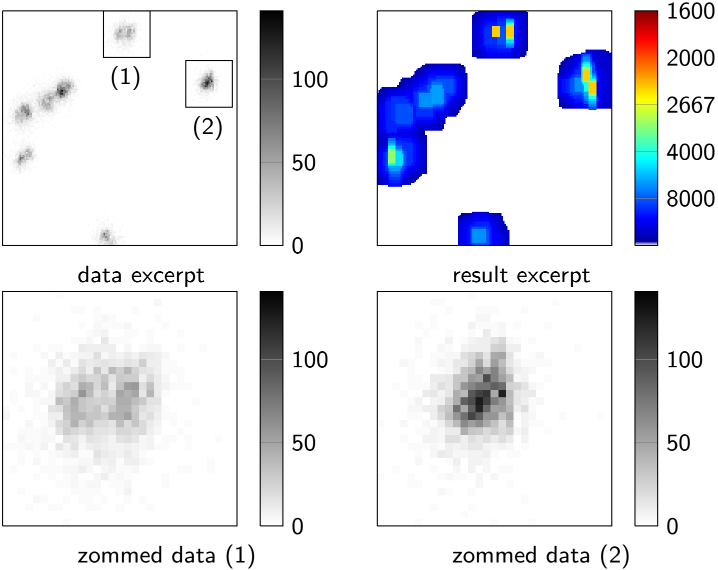


$k_{2,0.016}$

Result



Comparison of the result with the data



Outline

Outlook

- Methodology and theory:
  - inference on active coefficients $\langle \varphi_i, f \rangle_{\mathcal{X}}$ w.r.t. a dictionary $\{\varphi_i\}$
  - techniques from multiscale testing yield uniform inference at a controlled (asymptotic) error level
  - detection power is optimal in a suitable sense
- Application:
  - the method can be used to determine the support of a function observed in a convolution model
  - performs well in a real data example

Thank you for your attention!