Regularized inversion methods and error bounds for statistical inverse problems

Nicolai Bissantz

Institut für Mathematische Stochastik
Georg-August Universität Göttingen

www.stochastik.math.uni-goettingen.de

and P. Anders, L. Boysen, T. Hohage, A. Munk & F. Ruymgaart
Two problems are inverses of each other, if the formulation of each involves all or part of the solution of the other. Historically, one direction of the problem is well-studied since long time ago (usually called the ”direct” problem), the other only recently studied and less well understood (”inverse problem”).

(J.B. Keller ’76, Am. Math. Mon.)
Introduction

Definition 1 (Hadamard) A problem is well-posed if

1. There exists a solution to the problem (existence).

2. There is at most one solution to the problem (uniqueness).

3. The solution depends continuously on the data (stability).

We consider problems with deterministic and with stochastic (random) noise in the data for which 3. does not hold.
So, what is an inverse problem?
Example 2 (Num. differentiation / integration)

Direct problem:

\[ F(x) = (Kf)(x) = \int_0^x f(t)dt, \quad f \in C^{(0)}[0, 1] \]

Assume a deterministic noise model

\[ f_{n,\delta}(t) = f(t) + \delta \sin \left( \frac{nt}{\delta} \right), \quad \delta \text{ noise level} \]

Then

\[ \|f_{n,\delta} - f\|_\infty = \delta \]
\[ \|F_{n,\delta} - F\|_\infty = \left\| \frac{\delta^2}{n} \int_0^{nx/\delta} \sin(t)dt \right\|_\infty = O \left( \frac{\delta^2}{n} \right) \]
**Introduction**

**Inverse problem:** Assume the same (deterministic) noise model

\[ F_{n,\delta}(x) = F(x) + \delta \sin \left( \frac{nx}{\delta} \right) \]

\[ ||F_{n,\delta} - F||_{\infty} = \delta \]

but

\[ F'_{n,\delta}(x) = F'(x) + n \cos \left( \frac{nx}{\delta} \right) \]

and thus

\[ ||F'_{n,\delta} - F'||_{\infty} = n, \quad \delta \to 0 \]
Globular clusters in the Antennae galaxies
Globular clusters in the Antennae galaxies

Model for observed luminosity $X$:

$$X = F + W, \quad F \sim \text{YMCLF}, \quad \text{noise } W \sim \mathcal{N}(0, \sigma^2(f)): $$

YMCLF = ”Young massive cluster luminosity function”,
Note: Variance $\sigma^2(f)$ depends on (true) cluster luminosity

$F \equiv ”\text{The fainter the cluster, the more difficult it is to determine its brightness.”}$

Density $g$ of $X$:

$$g = \int_{\mathbb{R}} \frac{e^{-(-y)^2/\sigma^2(y)}}{\sqrt{2\pi\sigma^2(y)}} \text{YMCLF}(y) dy. $$

$\Rightarrow$ If $\sigma^2(y) \neq \text{const}$ this is not convolution & spectral information of the operator is unknown!
Linear statistical inverse problems

\( \mathbb{H}_1, \mathbb{H}_2 \) separable, \( K : \mathbb{H}_1 \to \mathbb{H}_2 \), linear, injective, bounded

\[
Y = Kf + \delta_{det} \xi_{det} + \delta_{ran} \xi_{ran}
\]

**Deterministic:** \( \|\xi_{det}\| = 1, \quad \delta_{det} \) noise level

**Stochastic:** \( \xi_{ran} \) \( H \)-valued rv (in general a \( H \)-space process), centered, \( \|\text{Cov}\xi_{ran}\| = 1 \), with variance \( \delta_{ran}^2 \)
Example 3 (*Error-in-variables, deconvolution*)

\[ X_1, \ldots, X_n \sim X = F + W \]

\( F, W \) stoch. indep. with densities \( f, w \in L^2, w \) known.

Wanted: \( f \)

\[ g = Kf = w * f \]

\[ \hat{q}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} w(X_j - \cdot) \]

unbiased, \( \sqrt{n}\)-consistent for \( q = K^*Kf \).

\[ \delta_{ran}K^*\xi_{ran} = \hat{q}_n - q \]
Example 4 *(Inverse regression, Fredholm equation)*

\[ Y_i = K f(X_i) + \epsilon_i, \quad (X_i, \epsilon_i), \ i = 1, \cdots, n, \ i.i.d. \]

\[ \mathbb{E}[Y|X] = K f(X), \quad K \text{ lin. Int. Op.} \]

*Generalized empirical process*

\[ \hat{q}_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} Y_i K(X_i, \cdot) \]

unbiased for \( q = (K^*K)f \), and \( \sqrt{n} \)-consistent. \( \xi_{ran} \) is defined as

\[ \delta_{ran} K^* \xi_{ran} = \hat{q}_n - q \]

Other models: error-in-variables, deconvolution (non-compact operators!), quasi-deconvolution, ...
Estimators

\[ K : \mathbb{H}_1 \rightarrow \mathbb{H}_2 \text{ bounded, injective} \]

\[ K^\dagger = (K^* K)^{-1} K^* : R(K) \oplus R(K)^\perp \rightarrow \mathbb{H}_1 \text{ unbounded.} \]

Regularized inverse:
Approximate \( K^\dagger \) by a sequence of bounded operators
Example 5 (Spectral Cut-off, Orth. series est.) Let $K$ compact.

\[ g = Kf = \sum_{n=0}^{\infty} \sigma_n < f, \phi_n > g_n, \]

Singular value decomp. $(\sigma_n, \phi_n, g_n)_{n \in \mathbb{N}}$. Under Picard condition:

\[ f = \sum_{n=1}^{\infty} \sigma_n^{-1} < g, g_n > \phi_n \]

Spectral Cut-off-estimator (Diggle, Hall’93, Mair, Ruymgaart’96, ...):

\[ \hat{f}_\alpha = \sum_{n=1}^{\infty} 1\{\sigma_n \geq \alpha\} \sigma_n^{-1} < Y, g_n > \phi_n \]

Computationally simple, but spectral information for the estimator has to be known explicitly. However this is often not the case, e.g. in the globular cluster application!
Unified way of presentation and analysis: **Functional calculus** (→ spectral theorem)

\[ A = K^*K \text{ selfadj. } \Phi : \sigma(A) \rightarrow \mathbb{R}, \text{ bounded, measurable,} \]

\[ \Phi(A) = U^* M_{\Phi(\rho)} U \]

**Regularized inverse:** Approximate \( K^+ \) by \( \Phi_\alpha(K^*K)K^* \)

**Generalized regularization estimator:**

\[ \hat{f}_\alpha = \Phi_\alpha(K^*K)K^*Y \]
### Estimators

<table>
<thead>
<tr>
<th>name</th>
<th>( \Phi_\alpha(t) )</th>
<th>( \nu_0 )</th>
</tr>
</thead>
</table>
| spectral cut-off \( [\Phi^{SC}_\alpha] \) | \[
\begin{cases}
  t^{-1}, & t \geq \alpha \\
  0, & t < \alpha
\end{cases}
\] | \( \infty \) |
| Tikhonov regularization \( [\Phi^{Tikh}_\alpha] \) | \( \frac{1}{t+\alpha} \)                                      | \( 1 \)   |
| iterated Tikhonov, \( k \geq 1 \)        | \( \frac{(t+\alpha)^k - \alpha^k}{t(t+\alpha)^k} \)          | \( k \)   |
| Landweber iteration                      | \( \sum_{j=0}^{k-1} (1 - t)^j \)                              | \( \infty \) |
| \( \nu \)-method \( (\nu > 0) \)        | \( \in P_{k-1} \)                                             | \( \nu \) |

Table 1: Regularization methods (cf. Engl, Hanke & Neubauer, 96)
How good is such an estimator?

**Smoothness classes:** \( \Lambda : [0, \infty) \rightarrow [0, \infty) \) continuous, \( \Lambda(0) = 0 \)

\[
\mathcal{F}_\Lambda = \{ f : f = \Lambda(K^*K)w : w \in H_1, ||w|| \leq 1 \}
\]

\( \Lambda(t) = t^\nu \): of Hölder-type, often correspond to Sobolev scales, etc.

**Approximation error** for noise-free data \( g = Kf \):

\[
||\Phi_\alpha(K^*K)K^*g - f|| \leq \sup_{t \in \sigma(A)} |(\Phi_\alpha(t)t - 1)\Lambda(t)|
\]

**Qualification:**

\[
\sup_{t \in \sigma(A)} |t^\nu(1 - t\Phi_\alpha(t))| \leq \gamma_\nu \alpha^\nu, \quad \text{für alle } \alpha \text{ und } 0 \leq \nu \leq \nu_0.
\]

Measures the largest amount of smoothness for which the method converges of optimal order.

Similarly: Generalized assumptions for exp. ill-posed problems.
Mean Integrated Square Error:

\[
\mathbb{E} \| \hat{f}_\alpha - f \|^2 = \underbrace{\mathbb{B}^2(\hat{f}_\alpha, \delta)}_{\text{bias}} + \underbrace{\mathbb{E} \| \hat{f}_\alpha - \mathbb{E} \hat{f}_\alpha \|^2}_{\text{variance}},
\]

Approximation error and deterministic noise:

\[
\mathbb{B}(\hat{f}_\alpha) := \| \mathbb{E} \hat{f}_\alpha - f \| \leq \| \Phi_\alpha(A)A f - f \| + \delta_{\text{det}} \underbrace{\| \Phi_\alpha(A)K^* \xi_{\text{det}} \|}_{\text{propag. det. noise}} \leq \sqrt{C/\alpha}
\]

Approximation error \( \leq \gamma_\Lambda \Lambda(\alpha), \alpha \to 0. \)
**Statistical noise: Variance**

\[
E \| \hat{f}_\alpha - E \hat{f}_\alpha \|^2 = \delta^2_{ran} E \| \Phi_\alpha(\rho) UK^* \xi_{ran} \|^2 \\
\leq \delta^2_{ran} C \int_{S} \Phi_\alpha^2(\rho) \rho \, d\Sigma.
\]

if

\[
C_1 \rho(s) \leq \text{Var} (UK^* \xi_{ran}(s)) \leq C_2 \rho(s).
\]
Theorem 6 Assume $\Sigma\{\rho \geq \alpha\} \sim S(\alpha), \alpha \downarrow 0$ for $S$ decreasing, smooth.

$$E \|\hat{f}_{\alpha, \delta} - f\|^2 \lesssim \left( \gamma \Lambda(\alpha) + \sqrt{\frac{C_1}{\alpha}} \delta_{det} \right)^2 + \frac{C_2 \delta_{ran}^2}{\alpha^2} \int_0^\alpha S(\beta) \, d\beta, \quad \alpha \downarrow 0$$

uniformly in $f \in F_\Lambda$.

Theorem 7 Let $\delta_{det} = 0$. Then there exists a constant $C > 0$ s.t.

$$\sup_{f \in F_\Lambda} E \|\hat{f}_{\alpha} - f\|^2 \leq C \sup_{f \in F_\Lambda} E \|\hat{f}_{SC, \alpha} - f\|^2$$

for all $\delta_{ran} > 0$ and all $\alpha > 0$ sufficiently small.

Corollary. Spectral Cut-off achieves statistical minimax rates (Mair, Ruymgaart’96) and thus this holds for all other methods (within their qualification)!
Young massive cluster luminosity function

Figure 1: Left panel: $\hat{\text{YMCLF}}(\cdot)$ after 4 steps of a $\nu$-method, middle panel: $\hat{q}(\cdot)$ (with preconditioning), right panel: estimated std function $\hat{\sigma}(\cdot)$. 
"Mode hunting" in the Antennae YMCLF

Incompleteness of the data becomes a severe problem at the faint tail of the distribution, and a Maximum Likelihood approach was used.
"Mode hunting" in the Antennae YMCLF

Figure 2: The fit of the Gauss and the power law model for the V-Band bands. The grey line is a kernel density estimate of the observed data and the black line shows the Maximum Likelihood fit times the completeness function. The dashed line corresponds to the estimated luminosity function without considering the completeness.
Confidence bands for deconvolution estimators

\[ X_j = Z_j + \varepsilon_j, \quad j = 1, \ldots, n, \quad \text{i.i.d.} \]
\[ \varepsilon_j \text{ ordinary smooth with known density } \Psi, \ f \text{ the (unknown) density of } Z_j. \]

Let \( \hat{f} \) a spectral cut-off estimator for \( f \). How sure can we be about any conclusion from \( \hat{f} \)?
⇒ we need confidence bands for \( \hat{f} \! \rightleftharpoons \)
⇒ study the maximum of the process

\[ Y_n(t) = \frac{n^{1/2} h^{\beta+1/2}}{g^{1/2}(t)} \left( \hat{f}_n(t) - E \hat{f}_n(t) \right), \quad t \in [0, 1], \]

\( g(t) \) an estimator of \( f \ast \Psi \) (i.e. the density of \( X \)), \( \beta \) depends on \( \Psi \) (i.e. \( \beta = 2 \) for the Laplace distribution), \( h \) the bandwidth of the spectral cut-off estimator \( \hat{f} \).

Asymptotic distribution can be computed. For finite sample size we suggest a bootstrap:
Confidence bands for deconvolution estimators

1. Compute \( \hat{f}_n \).

2. Resample with replacement \( B \) samples of size \( n \) from \( X_1, \ldots, X_n \).

3. Compute the spectral cut-off estimator \( \hat{f}_n^* \) for each of these samples and the distance

\[
T_n^* = \sup_{t \in [0,1]} Y_n^*(t) = \sup_{t \in [0,1]} \frac{n^{1/2}h^{\beta+1/2}}{\tilde{g}_n^{1/2}(t)} \left( \hat{f}_n^*(t) - \hat{f}_n(t) \right)
\]

for each of these samples.

4. Determine the \( 1 - \alpha \) quantile \( q_{1-\alpha} \) from the distribution of \( T^* \).

\[
\implies \text{level-} \alpha \text{ confidence bands:}
\]

\[
\hat{f}_n(t) - \frac{q_{1-\alpha} \tilde{g}_n(t)}{\sqrt{n}h^{\beta+1/2}} \leq f(t) \leq \hat{f}_n(t) + \frac{q_{1-\alpha} \tilde{g}_n(t)}{\sqrt{n}h^{\beta+1/2}}, \quad t \in [0,1].
\]

- confidence bands give a uniform measure of the estimation error, whereas pointwise confidence intervals are less strict!
- similar results available for derivatives of \( f \).
Application: the local metallicity distribution

Figure 3: Estimates of the metallicity distribution of F and G dwarfs in the solar neighbourhood for the full sample (left panel) and the volume complete subsample (right panel). The full curves show a kernel estimate of the observations $X_i$, and the dashed curves (inverse) estimators and the associated 90% nominal coverage confidence bands of the density of [Fe/H].
Summary [Linear statistical inverse problems]

**Linear problems**

- **Spectral Cut-off** ist often difficult to implement (spectral properties of operator have to known explicitly)

- **Tikhonov reg.** underqualified from a deterministic point of view (⇒ often has suboptimal convergence rate!)

- **Iterative methods** achieve optimal rates of convergence and are computationally (more) simple (but are hardly known to statistical and (many) applied communities!)

- **Confidence bands** for deconvolution (etc...) can be obtained by asymptotic and bootstrap methods.