Statistical guarantees for inverse problems

Hanne Kekkonen

Delft Institute of Applied Mathematics
Delft University of Technology

November 17, 2022

Outline

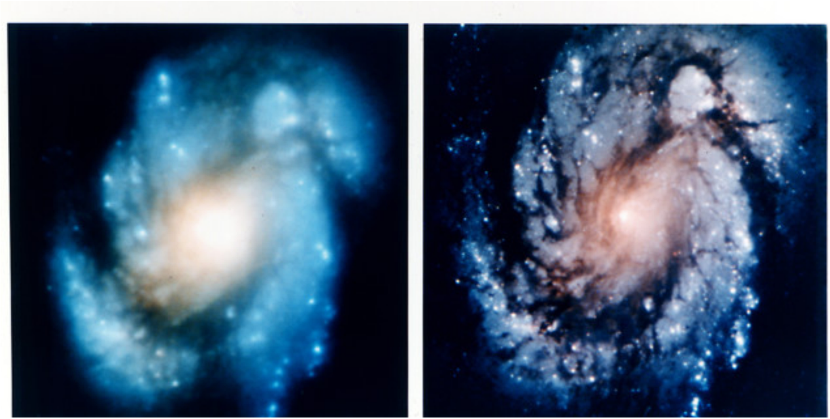
Outline

1. Introduction to inverse problems

2. Bayesian inverse problems

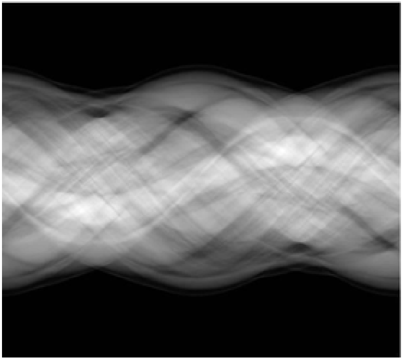
3. Consistency of nonparametric Bayesian methods

Deblurring (deconvolution)



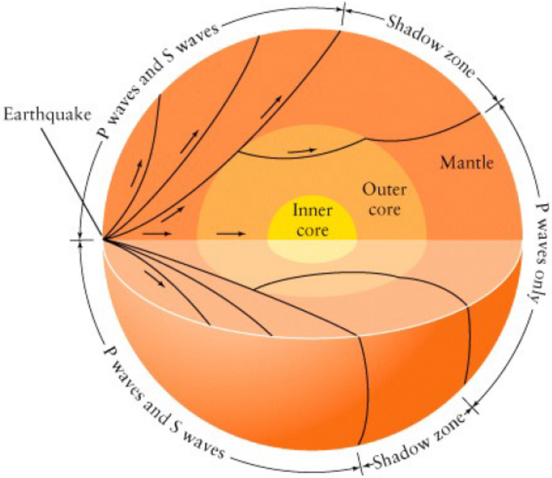
$$m(x) = (Af)(x) = \int_{\mathbb{R}^2} a(x-y) f(y) dy$$

Computerised tomography (CT)

$$M(\theta, s) = (\mathcal{G}u)(\theta, s) = \int_{x\cdot\theta=s} f(x)dx$$

Geodesic X-ray transform



$$m(\gamma) = (Af)(\gamma) = \int f(\gamma(t))dt$$

Many inverse problems arise from partial differential equations

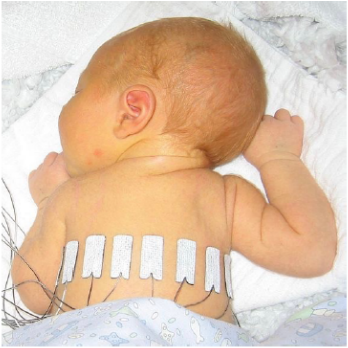
Elliptic PDEs: Given noisy measurements of $\mathcal{G}(f) = u_f$ recover $f > 0$ in the divergence form equation

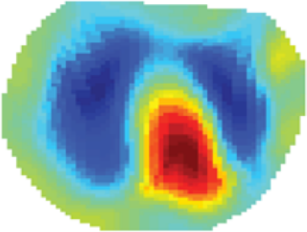
$$\nabla \cdot (f \nabla u) = g \text{ on } \mathcal{O}, \quad u = 0 \text{ on } \partial\mathcal{O}.$$

Time evolution equations: Given noisy measurements of $\mathcal{G}(f) = u_f$ recover $f > 0$ in the heat equation

$$\begin{cases} \frac{1}{2}\Delta_x u - \partial_t u - f u = 0 & \text{on } \mathcal{O} \times (0, \mathbf{T}) \\ u = g & \text{on } \partial\mathcal{O} \times (0, \mathbf{T}) \\ u(\cdot, 0) = u_0 & \text{on } \mathcal{O}. \end{cases}$$

Electrical Impedance Tomography (EIT)





Applying electric voltages f at the boundary leads to PDE

$$\nabla \cdot (\sigma \nabla v) = 0 \quad \text{in } \Omega \in \mathbb{R}^2$$
$$v|_{\partial\Omega} = f$$

Non-linear inverse problem:
Recover conductivity σ from boundary measurements
$\Lambda_\sigma(f) = \sigma \frac{\partial v}{\partial \vec{n}}|_{\partial\Omega}$

Inverse problems are ill-posed

We want to recover the unknown f from a noisy measurement m;

$$m = Af + noise,$$

where A is a forward operator that usually causes loss of information.

Well-posedness as defined by Jacques Hadamard:

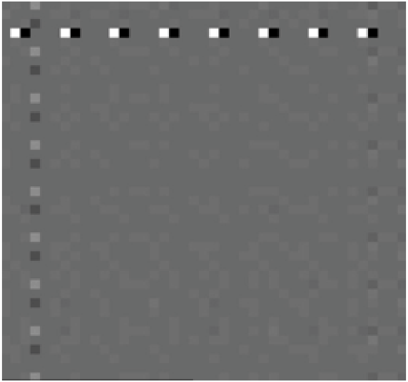
1. Existence: There exists at least one solution.
2. Uniqueness: There is at most one solution.
3. Stability: The solution depends continuously on data.

Inverse problems are ill-posed breaking at least one of the above conditions.

Naive reconstruction does not work for inverse problems

If A is invertible it is tempting to try $f^{naive} \approx A^{-1}m = f + A^{-1}noise$.

Blurry and noisy image



Naive inversion

The problem is ill-posedness: $\|A^{-1}noise\| \approx \|noise\|/\lambda_k \gg \|u\|$, where λ_k is the smallest eigenvalue of A.

Outline

Deterministic approach to inverse problems

Recover the f from $m = Af + \varepsilon$, where ε small and deterministic noise.

Tikhonov regularisation offers a stable solution to the problem

The classical way of solving inverse problems is minimising the penalised least squares criterion

$$\widetilde{f} = \arg\min_{f} \left\{ \|Af - m\|_2^2 + \alpha R(f) \right\}$$

The above can be understood as a balance between two requirements:

1. $\widetilde{f}$ should give a small residual $A\widetilde{f} - m$,
2. The penalty $R(\widetilde{f})$ should be small.

The regularisation parameter $\alpha > 0$ can be used to "tune" the balance.

Note that the minimisation problem is non-convex when A is non-linear.

Bayes formula combines data and a priori information

Reconstruct the most probable f from $m = Af + \varepsilon$, with ε random noise, in light of

- Measurement information: $m \,|\, f \sim P_f$ with density $\rho(m \,|\, f) = \rho_\varepsilon(m - Af)$.
- A priori information: $f \sim \Pi_{pr}$ with density $\pi_{pr}(f)$.

Bayes' formula

We can update the prior, given a measurement, to a posterior distribution using the Bayes' formula:

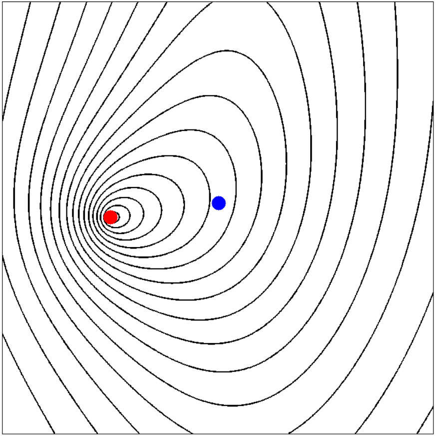
$$\pi(f \,|\, m) \propto \rho(m \,|\, f)\pi_{pr}(f)$$

The result of Bayesian inversion is the posterior distribution $\pi(f \mid m)$.

The result of Bayesian inversion is the posterior distribution, but typically one looks at point estimates

Maximum a posteriori (MAP) estimate:
$\arg\max_{u\in\mathbb{R}^n} \pi(u \,|\, m)$

Conditional mean (CM) estimate:
$\int_{\mathbb{R}^n} u\,\pi(u \,|\, m)\,du$



Uncertainty quantification has many applications

Studying the whole posterior distribution instead of just a point estimate offers us more information.

Uncertainty quantification

- Confidence and credible sets
- E.g. Weather and climate predictions

Using the whole posterior

- Geological sensing
- Bayesian search theory

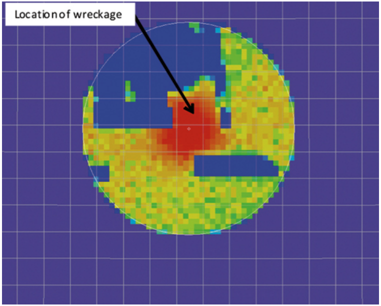
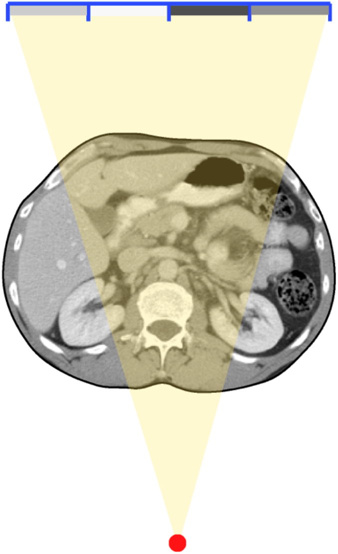


Figure: *Search for the wreckage of Air France flight AF 447, Stone et al.*

The measurement is always discrete but the unknown is usually a continuous function

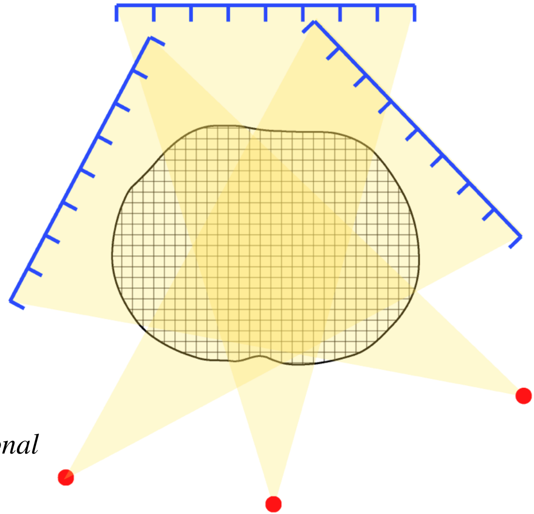
$m \in \mathbb{R}^4$

$f \in L^2$



Computational solutions require a finite approximate model for the unknown f

$m \in \mathbb{R}^{24}$
$f \in \mathbb{R}^{440}$

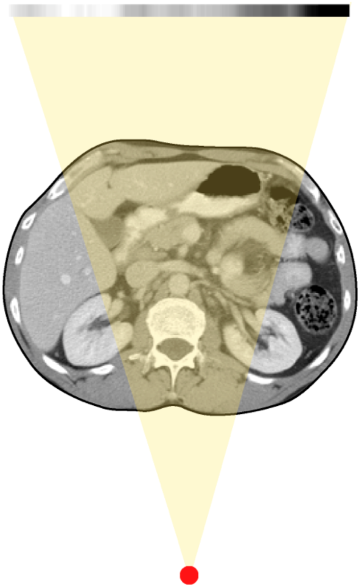


J. Kaipio & E. Somersalo, *Statistical and Computational Inverse Problems*, 2005

Avoid discretisation until the last possible moment

A.M. Stuart, *Inverse problems: A Bayesian perspective*, 2010.

- The first-order wave equation is not controllable to a given final state in arbitrarily small time (finite speed of propagation).
- Every finite difference spatial discretisation gives rise to a linear system of ordinary differential equations which is controllable, in any finite time, to a given final state.



White noise does not belong to L^2

Let ψ_j form an orthonormal basis for L^2. Then formally

$$\varepsilon = \sum_{k=0}^{\infty} \langle \varepsilon, \psi_k \rangle \psi_k.$$

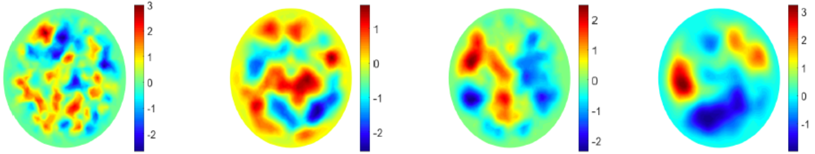
The Fourier coefficients of white noise satisfy $\langle \varepsilon, e_k \rangle \sim N(0,1)$, where $e_k(t) = e^{ikt}$. Hence

$$\mathbb{E}\|\varepsilon\|_2^2 = \sum_{k=0}^{\infty} \mathbb{E}|\langle \varepsilon, e_k \rangle|^2 = \sum_{k=0}^{\infty} 1 = \infty.$$

For the white noise we have

- $\varepsilon \in L^2$ with probability zero,
- $\varepsilon \in H^{-s}$, $s > d/2$, with probability one.

Gaussian priors



- Consider white noise $\varepsilon \sim \Pi = \mathcal{N}(0, I)$.
- We often write $\pi(\varepsilon) \underset{formally}{\propto} \exp(-\frac{1}{2}\|\varepsilon\|_{L^2}^2)$.
- Note that $\Pi(L^2) = 0$ and $\Pi(H^{-s}) = 1$, for $s > d/2$.
- L^2 characterises the directions in which the centred Gaussian measure Π can be shifted to obtain an equivalent Gaussian measure.
- L^2 is called the Cameron–Martin space for Π.

Bayesian approach to inverse problems

We want to recover the unknown f from a noisy measurement m;

$$m = Af + \varepsilon.$$

- Consider observing data m drawn at random from some unknown probability distribution $P^m_{f^\dagger}$, and sample size n.
- Specify a prior distribution Π for the unknown f and assume

$$m \,|\, f \sim P^m_f.$$

- Using Bayes' theorem the prior distribution can be updated to a posterior distribution

$$f \,|\, m \sim \Pi(\cdot \,|\, m).$$

Outline

Consistency of the Bayesian solution

The natural next step is to consider the consistency of a solution.

- Convergence of a point estimator to the 'true' $f^\dagger$.
- Contraction of the posterior distribution; Do we have, as the noise level $\varepsilon \to 0$,

$$\Pi(f : \|f - f^\dagger\| \geq \delta_\varepsilon \,|\, M_\varepsilon) \to^{P^M_{f^\dagger}} 0,$$

  for some posterior contraction rate $\delta_\varepsilon \to 0$.

  - Usually this also guarantees that the posterior mean converges to $f^\dagger$.
  - We can also study if the rate is optimal

$$\inf_{\hat{f}=\hat{f}(M)} \sup_{f \in \mathcal{F}} \mathbb{E}^M_f \|\hat{f} - f\| \simeq \delta_\varepsilon^{minimax}.$$

- Coverage of the credible sets.

Some contraction results for linear inverse problems

Singular value decomposition based

- Knapik, van der Vaart & van Zanten (2011); Mutually diagonalisable operators
- Ray (2013); Non-conjugate rate adaptive sequence setting
- Knapik, Szabó, van der Vaart, van Zanten (2016); Adaptive priors

General smoothness requirements

- Agapiou, Larsson & Stuart (2013); Mildly ill-posed problems
- Kekkonen, Lassas & Siltanen (2016); Pseudodifferential operators
- Knapik & Salomond (2018); Modulus of continuity
- Agapiou, Dashti & Helin (2021); p-exponential priors
- Agapiou & Mathé (2021); Truncated Gaussian series priors

Some contraction results for non-linear inverse problems

Consider measurements

$$m_i = A_f(X_i) + w_i, \quad i = 1, ..., N, \; w_i \sim \mathcal{N}(0, 1),$$

and

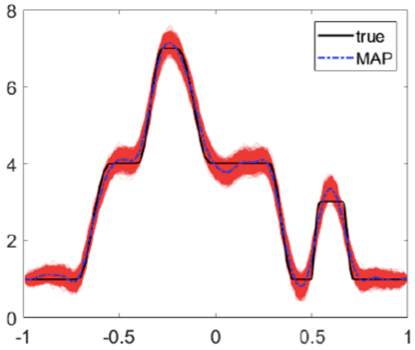
$$\Pi(f : \|f - f^\dagger\| \geq \delta_N \mid (m_i, X_i)_{i=1}^N) \to^{P^M_{f^\dagger}} 0.$$

Results using scaled Gaussian process priors

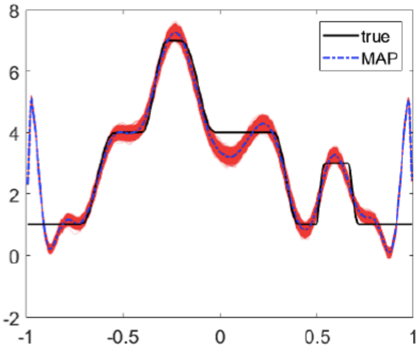
- Monard, Nickl, Paternain (2021); Non-linear X-rays, $\delta_N \approx N^{-\gamma}$.
- Abraham, Nickl (2019); Calderón problem, $\delta_N \approx (\log N)^{-\gamma}$.
- Giordano, Nickl (2020); Divergence form, $\delta_N \approx N^{-\gamma}$.
- Kekkonen (2021): Heat equation with absorption term, $\delta_N \approx N^{-\gamma}$.

R. Nickl, *Bayesian Non-linear Statistical Inverse Problems,* 2022

Do credible sets quantify frequentist uncertainty?



Correctly specified prior



Prior misspecified on the boundary

Monard, Nickl & Paternain, *The Annals of Statistics*, 2019

Optimal contraction does not guarantee correct coverage!

If $f \in \mathbb{R}^d$ credible sets have correct coverage

Do we have for $C = C(M)$

$$\Pi\Big(f \in C \,|\, M\Big) \approx 0.95 \quad \Leftrightarrow \quad P^M_{f^\dagger}\Big(f^\dagger \in C(M^\dagger)\Big) \approx 0.95?$$

Bernstein–von Mises Theorem (BvM)

For large sample size n, with $\hat{f}_{MLE}$ being the maximum likelihood estimator,

$$\Pi(\cdot \,|\, M) \approx N\Big(\hat{f}_{MLE}, \frac{1}{n} I(f^\dagger)^{-1}\Big), \quad \text{for } M \sim P^M_{f^\dagger},$$

whenever $f^\dagger \in \mathcal{F} \subset \mathbb{R}^d$ and the prior Π has positive density on $\mathcal{F}$, and the inverse Fisher information $I(f^\dagger)$ is invertible.

BvM guarantees confident credible sets

The contraction rate of the posterior distribution near $f^\dagger$ is

$$\Pi\Big(f : \|f - f^\dagger\|_{\mathbb{R}^d}^2 \geq \frac{L_n^2}{n} \,\Big|\, M\Big) \to^{P_{f^\dagger}^M} 0 \quad \text{as } L_n, n \to \infty$$

For a fixed d and large n computing posterior probabilities is roughly the same as computing them from $N\big(\hat{f}_{MLE}, \frac{1}{n} I(f^\dagger)^{-1}\big)$.

$$C_n \text{ s.t. } \Pi\big(f \in C_n \,|\, M\big) = 0.95 \quad \Longrightarrow \quad P_{f^\dagger}\big(f^\dagger \in C_n\big) \to 0.95$$

(Bayesian credible set) (Frequentist confident set)

$$|C_n|_{\mathbb{R}^d} = \mathcal{O}_{P_{f^\dagger}}\left(\frac{1}{\sqrt{n}}\right) \quad \text{(Optimal diameter)}$$

Consistency of nonparametric Bayesian methods

- If f is a function the BvM theorem does not hold in the L^2 sense: Cox (1993), and Diaconis & Freedman (1999).
- Castillo and Nickl (2013, 2014) showed for direct models that, while BvM results do not hold in L^2, they can hold in larger spaces, such as Sobolev spaces H^{-s}, with $s > d/2$.
- Coverage of the credible sets; Bernstein von Mises type theorems. Castillo & Nickl (2013, 2014), Ray (2014), Monard, Nickl & Paternain (2019), Nickl (2018), Nickl & Söhl (2019), Giordano & Kekkonen (2020).