Inverse problems:
From Regularization to Bayesian Inference

Daniela Calvetti

Case Western Reserve University

\(^1\)based on forthcoming WIRES overview article "Bayesian methods in inverse problems"
Inverse problems are everywhere

The solution of an inverse problem is needed when

- Recovering unknown causes from indirect, noisy measurement
- Determining poorly known - or unknown - parameters of mathematical model
- Inferring on the interior of a body from non invasive measurements
Going against causality

Forward models:
- Move in the causal direction from causes to consequences
- Are usually well-posed: small perturbations in the causes lead to little change in the consequences
- May be very difficult to solve

Inverse problems:
- Go against causality mapping consequences to causes
- Their instability is the price for the well-posedness of the associated forward model
- The more forgiving the forward problem, the more unstable the inverse one
Instability vs regularization

- Regularization trades original unstable problem for a nearby approximate one.

- For a linear inverse problem with forward model \( f(x) = Ax \), instability \( \iff \) wide range of singular values.

- Division by small singular values amplifies noise in the data.

- Lanczos (1958) replaces \( A \) with low rank approximation prior to inversion.

- Similar idea is formulated as Truncated SVD regularization (1987).
Tikhonov regularization

- Tikhonov (1960s) leave forward model intact
- The solution is penalized for traits due to amplified noise:

  $$x_{\text{Tik}} = \text{argmin} \left\{ \|b - f(x)\|^2 + \alpha G(x) \right\}$$

- Under suitable condition on $f$ and $G$ the solution exists and is unique
- $G(x)$ monitors growth of instability
- $\alpha > 0$ assesses relative weights of fidelity and penalty terms

Tikhonov penalty requires a priori belief about the solution!
Standard choices of the regularization operator are

- $G(x) = \|Lx\|^2$, with $L = I$ to penalize growth in the solution Euclidean norm
- $G(x) = \|Lx\|^2$, with $L$ equal to a discrete first order differencing to penalize growth in the gradient
- $G(x) = \|Lx\|^2$, with $L$ equal to the discrete Laplacian, to penalize growth in the curvature
- $G(x) = \|x\|_1$, to favor sparse solutions, related to compressed sensing
- Total variation (TV) of $x$, to favor blocky solutions.
Regularization parameter

- Tikhonov regularized solution depends heavily on value of $\alpha$
- In the limit as $\alpha \to 0$, solve an output least squares problem
- If $\alpha$ is very large the solution may not predict well the data
- Morozov discrepancy principle:
  If the norm of the noise, $\|e\| \approx \delta$, is known, $\alpha = \alpha(\delta)$ should be chosen so that
  \[
  \|b - f(x_{\alpha})\| = \delta.
  \]
- If norm of the noise is not known, selection of $\alpha$ can be based on heuristic criteria
Semi-convergence and stopping rules

Given the linear system

\[ Ax = b, \]

and \( x_0 = 0 \), a Krylov least squares solver generate a sequence of approximate solutions \( \{x_0, x_1, \ldots x_k, \ldots \} \), with

\[ x_k \in \mathcal{K}_k \left( A^T A, A^T b \right) = \text{span} \left\{ A^T b, \ldots (A^T A)^{k-1} A^T b \right\}. \]

- If \( Ax = b \) is a discrete inverse problem, iterates start to converge to the solution, then diverge (semi-convergence).
- If \( d_k = b - Ax_k \) form a decreasing sequence, iterate until

\[ \| b - Ax_k \| \leq \delta. \]
A priori info → Solution features

Noise level → Regularization parameter

Data → Fidelity term

Forward model → Solution smoothness

Fidelity term → Stopping rule

Regularization operator → Single solution

Penalty term → Single solution

Regularization parameter → Single solution

Fidelity term → Single solution
The Bayesian framework for inverse problems

- Model unknown as a random variable $X$, and describe it via a distribution, not a single value
- Model observed data as random variable $B$ describing its uncertainty
- Stochastic extension of $f(x) = b + e$

$$B = f(X) + E$$

- If $X$, $E$ are mutually independent, likelihood of the form

$$\pi_{B|X}(b \mid x) = \pi_E(b - f(x)).$$
Bayes’ formula

Update the belief about $X$ by the information through $B = b_{\text{observed}}$ using Bayes’ formula,

$$
\pi_{X|B}(x \mid b) = \frac{\pi_{B|X}(b \mid x)\pi_X(x)}{\pi_B(b)}, \quad b = b_{\text{observed}},
$$

where

$$
\pi_B(b) = \int_{\mathbb{R}^n} \pi_{B|X}(b \mid x)\pi_X(x)dx,
$$

normalization factor.

This is the complete solution of the inverse problem!
The art and science of designing priors

A prior is the initial description of the unknown $X$ before considering the data. There are different types of priors:

- Smoothness priors (0th, 1st and 2nd order)
- Correlation priors
- Structural priors
- Sparsity promoting priors
- Sample-based priors
Smoothness priors: 1st order

Let

\[ X_j - X_{j-1} = \gamma W_j, \quad W_j \sim \mathcal{N}(0, 1), \quad 1 \leq j \leq n \]

or, collectively

\[ L_1 X = \gamma W \sim \mathcal{N}(0, I_n), \quad L_1 = \begin{bmatrix} 1 & 1 \\ -1 & 1 \\ \vdots & \vdots \\ -1 & 1 \end{bmatrix}, \]

leading to a Gaussian prior model,

\[ \pi_X(x) \propto \exp\left( -\frac{1}{2\gamma^2} \|L_1x\|^2 \right), \quad x \in \mathbb{R}^n. \]
Smoothness priors: 2nd order

Let
\[ 2X_j - (X_{j-1} + X_{j+1}) \sim \gamma W_j, \quad 1 \leq j \leq n - 1, \]

Collectively

\[ L_2X = \gamma W \sim \mathcal{N}(0, I_{n-1}), \quad L_2 = \begin{bmatrix} 2 & -1 \\ -1 & 2 & \ddots \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix}. \]

This leads to the Gaussian prior
\[ \pi_X(x) \propto \exp \left( -\frac{1}{2\gamma^2} \|L_2x\|^2 \right), \quad x \in \mathbb{R}^{n-1}. \]
Whittle-Matérn priors,

- $D$ discrete approximation of Laplacian in $\Omega \subset \mathbb{R}^n$
- Prior for $X$ defined in $\Omega$ can be described by a stochastic model
  \[ (-D + \lambda^{-2}I)^\beta X = \gamma W, \]
- where $\gamma, \beta > 0$, $\lambda > 0$ is the correlation length, and $W$ is a white noise process.
Structural priors

- Solution to have a jump across structure boundaries
- Within the structures, the smoothness prior is a valid description

\[ X_j - X_{j-1} = \gamma_j W_j, \quad W_j \sim \mathcal{N}(0, 1), \quad 1 \leq j \leq n, \]

where \( \gamma_j > 0 \) is small if we expect \( x_k \) near \( x_{k-1} \), and large where \( x_k \) could differ significantly from \( x_{k-1} \).

\[ \pi_{\text{struct}}(x) \propto \exp \left( -\frac{1}{2} \| \Gamma^{-1} L_1 x \|^2 \right). \]

Examples:
- geophysical sounding problems with known sediment layers
- electrical impedance tomography with known anatomy
Sparsity promoting priors

These priors:

- assume very small values for most entries, while allowing isolated, sudden changes.
- Satisfy
  \[
  \pi_X(x) \propto \exp(-\alpha \|x\|_p).
  \]
- $\ell^p$-priors are non-Gaussian, introducing computational challenges.
Sample based priors

- Assume we can build a generative model to generate typical solutions
- Generate a sample of typical solutions
  \[
  \mathcal{X} = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}
  \]
- Regard solutions in \( \mathcal{X} \) as independent draws from approximate prior
- Approximate prior by a Gaussian
  \[
  \pi_X(x) \propto \exp \left( -\frac{1}{2} (x - \overline{x})^T \Gamma^{-1} (x - \overline{x}) \right),
  \]
  where
  \[
  \overline{x} = \frac{1}{N} \sum_{j=1}^{N} x^{(j)}, \quad \Gamma = \frac{1}{N} \sum_{j=1}^{N} (x^{(j)} - \overline{x})(x^{(j)} - \overline{x})^T + \delta^2 I,
  \]
Solving inverse problems the Bayesian way

The main steps in solving inverse problems are:

▶ constructing an informative and computationally feasible prior;
▶ constructing the likelihood using the forward model and information about the noise;
▶ forming the posterior density using Bayes’ formula;
▶ extracting information from the posterior.
The pure Gaussian case

Assume

\[ E \sim \mathcal{N}(0, C), \text{ or } \pi_E(e) = \frac{1}{(2\pi)^{m/2} \sqrt{\det(C)}} \exp \left( -\frac{1}{2} e^T C^{-1} e \right). \]

and

\[ \pi_X(x) = \exp \left( -\frac{1}{2} G(x) \right). \]

then

\[ \pi_{X|B}(x \mid b) \propto \exp \left( -\frac{1}{2} (x - \bar{x})^T B^{-1} (x - \bar{x}) \right), \]

where the posterior mean and covariance are given by

\[ \bar{x} = (G^{-1} + A^T C^{-1} A)^{-1} A^T C^{-1} b, \]

\[ B = (G^{-1} + A^T C^{-1} A)^{-1}. \]
Assume

- Scaled white noise model \( C = \sigma^2 I \)
- Zero-mean Gaussian prior with covariance matrix \( G \) with \( G^{-1} = L^T L \),

The posterior is of the form
\[
\pi_{X|B}(x | b) \propto \exp \left( -\frac{1}{2\sigma^2} \left( \| b - f(x) \|^2 + \sigma^2 \| Lx \|^2 \right) \right).
\]

Tikhonov regularized solution with \( G(x) = Lx \) is the Maximum A Posteriori estimate of the posterior
\[
x_\alpha = \arg\max \{ \pi_{X|B}(x | b) \}, \quad \alpha = \sigma^2.
\]
Hierarchical Bayes models and sparsity

- Conditionally Gaussian prior

\[
\pi_{X|\Theta}(x | \theta) = \frac{1}{(2\pi)^{n/2}} \exp \left( -\frac{1}{2} \sum_{j=1}^{n} \frac{x_j^2}{\theta_j} - \frac{1}{2} \sum_{j=1}^{n} \log \theta_j \right)
\]

- Solution is sparse, or close to sparse: most variances of \(x_j\) are small, with some outliers

- Gamma hyperpriors

\[
\pi_\Theta(\theta) \propto \prod_{j=1}^{n} \theta_j^{\beta-1} e^{-\theta/\theta_0},
\]

with \(\beta > 0\) and \(\theta_0\) are the shape and scaling parameters.

- Posterior distribution

\[
\pi_{X,\Theta|B}(x, \theta | b) \propto \pi_\Theta(\theta) \pi_{X|B,\Theta}(x | b, \theta)
\]
Efficient quasi-MAP estimate

To find a point estimate for the pair \((x, \theta) \in \mathbb{R}^n \times \mathbb{R}_+^n\)

- Given the current estimate \(\theta^{(k)} \in \mathbb{R}_+^n\), update \(x \in \mathbb{R}^n\) by minimizing

\[
E(x \mid \theta^{(k)}) = \frac{1}{2\sigma^2} \|b - f(x)\|^2 + \frac{1}{2} \sum_{j=1}^{n} \frac{x_j^2}{\theta_j^{(k)}},
\]

- Given the updated \(x^{(k)}\), update \(\theta^{(k)} \rightarrow \theta^{(k+1)}\) by minimizing

\[
E(\theta \mid x^{(k)}) = \frac{1}{2} \sum_{j=1}^{n} \frac{x_j^2}{\theta_j} + \eta \sum_{j=1}^{n} \log \theta_j + \sum_{j=1}^{n} \frac{\theta_j}{\theta_0}, \quad \eta = \beta - \frac{3}{2}.
\]
Bayes meets Krylov...

When the problems are large and the computing time at a premium:

- The least squares problems in step 1 can be solved by CGSL with suitable stopping rule
- A right preconditioner accounts for the information in the prior
- A left preconditioner accounts for the observational and modeling errors
- The quasi-MAP is computed very rapidly via an inner-outer iteration scheme
- A number of open questions about the success of the partnership are currently under investigation
The Bayesian horizon is wide open

When solving inverse problems in the Bayesian framework

- We compensate uncertainties in forward models (see Ville’s talk)
- We can explore the posterior (see John’s talk)
- We can generalize classical regularization techniques to account for all sort of uncertainties.