Inverse Problems: Minimization, Bayesian and Robust Bayesian Perspectives

Tim Sullivan

University of Warwick

SIXNS Workshop
Getting from Data to Science
Brookhaven National Laboratory, NY, U.S.A.
14–15 August 2013

Overview

Credits

Bayesian Inverse Problems in Function Spaces

A. M. Stuart & al.

Robust Bayesian / Optimal UQ Perspective

M. McKerns

M. Ortiz

H. Owhadi

C. Scovel

Scattering and Nanostructure Problems

S. J. L. Billinge & al.

Inverse Problems

- Often in science we are faced with an inverse problem of the form

$$y = \mathcal{G}(u),$$

  where
  - $u \in \mathcal{U}$ is the unknown;
  - $y \in \mathcal{Y}$ is some observed data;
  - $\mathcal{G} \colon \mathcal{U} \to \mathcal{Y}$ is the observation operator that maps u to y.
- E.g. $u =$ nanostructure, $y =$ pair distribution function.
- E.g. $u =$ solution to some PDE, $y =$ point/Eulerian/Lagrangian data.

Inverse Problems

- Often in science we are faced with an inverse problem of the form

$$y = \mathcal{G}(u),$$

where

  - $u \in \mathcal{U}$ is the unknown;
  - $y \in \mathcal{Y}$ is some observed data;
  - $\mathcal{G} \colon \mathcal{U} \to \mathcal{Y}$ is the observation operator that maps u to y.

- E.g. $u =$ nanostructure, $y =$ pair distribution function.
- E.g. $u =$ solution to some PDE, $y =$ point/Eulerian/Lagrangian data.
- Often, the observation map $\mathcal{G}$ is difficult to invert in practice.
- Even when $\mathcal{G}$ is a straightforward linear operator, for a given y, there may be no solution, or there may be multiple solutions, or a unique solution with sensitive dependence upon y.
- Often, we observe not $\mathcal{G}(u)$ but some noisy version of it:

$$y = \mathcal{G}(u) + \eta.$$

Least Squares

- What to do when $y = \mathcal{G}(u)$ has no solution?
- Least squares approach: find u that minimizes

$$\|y - \mathcal{G}(u)\|_{\mathcal{Y}}^2$$

with respect to some norm $\|\cdot\|_{\mathcal{Y}}$ on $\mathcal{Y}$. But there is a modelling question: what norm to use?

Least Squares

- What to do when $y = \mathcal{G}(u)$ has no solution?
- Least squares approach: find u that minimizes

$$\|y - \mathcal{G}(u)\|_{\mathcal{Y}}^2$$

  with respect to some norm $\|\cdot\|_{\mathcal{Y}}$ on $\mathcal{Y}$. But there is a modelling question: what norm to use?
- Classical setting, linear observation operator $\mathcal{G} = A \in \mathbb{R}^{n \times m}$ maps solution $u \in \mathbb{R}^m$ to observed data $y \in \mathbb{R}^n$:
  - without observation noise, $y = Au$, minimizer of $\|y - Au\|_2$ is given by pseudo-inversion:

$$\hat{u} = (A^* A)^{-1} A^* y$$

  - with observation noise $\eta \sim \mathcal{N}(0, Q)$, $Q \in \mathbb{R}^{m \times m}$, $y = Au + \eta$, the Gauss–Markov theorem says that minimizing $\|y - Au\|_{Q^{-1}}^2$ gives

$$\hat{u} = (A^* Q^{-1} A)^{-1} A^* Q^{-1} y,$$

    which has mean u and covariance $(A^* Q^{-1} A)^{-1}$. This estimate also has minimal mean-square error and minimal covariance matrix.

Regularization

- If $\mathcal{G}$ is non-linear, and when $\mathcal{U}$ is infinite-dimensional, then there also difficulties with the least squares approach: there may be multiple minimizers, there may be sequences of approximate minimizers that do not converge in $\mathcal{U}$, or there may be sensitive dependence upon the observed data $y \in \mathcal{Y}$.
- One way around this is to regularize the problem by giving preference to solutions u that are close to a preferred candidate $\bar{u}$: seek u that minimizes
$$\|y - \mathcal{G}(u)\|_{\mathcal{Y}}^2 + \|u - \bar{u}\|_{\mathcal{U}}^2.$$
- The choice of regularization — i.e. what norm $\|\cdot\|_{\mathcal{U}}$ to put on $\mathcal{U}$ — seems to be somewhat ad hoc.
- A Bayesian perspective can illuminate the role of LSQ and regularization.

Overview

Bayesian Interpretation of LSQ

- The Bayesian interpretation of these minimization problems is that the equation

$$y = \mathcal{G}(u) + \eta$$

defines the conditional distribution $y|u$. Suppose that $\mathcal{Y}$ is finite-dimensional and that η has probability density function $\rho(\,\cdot\,)$.

- Let the potential $\Phi(u; y)$ be any function that differs from $\rho(y - \mathcal{G}(u))$ by an additive function of y alone, so that

$$\frac{\rho(y - \mathcal{G}(u))}{\rho(y)} \propto \exp(-\Phi(u; y)).$$

- E.g. when $\eta \sim \mathcal{N}(0, Q)$, take

$$\Phi(u; y) = \tfrac{1}{2} \| y - \mathcal{G}(u) \|_{Q^{-1}}^2 .$$

- The LSQ minimizers are the minimizers of Φ and the maximizers of the likelihood $L(y|u) \propto \exp(-\Phi(u; y))$.

Bayesian Interpretation of Regularization

- Introduce a prior measure μ on $\mathcal{U}$ encoding a priori beliefs about u.
- Bayes' rule for the posterior distribution $u|y \sim \mu^y$:

$$\mathbb{P}_\mu(u|y) \propto L(y|u)\mathbb{P}_\mu(u)$$

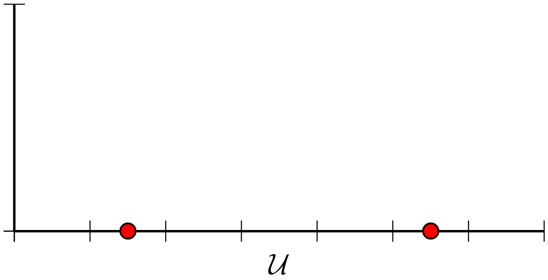
or, in language that makes more sense in functional settings,

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu} \propto L(y|u) \propto \exp(-\Phi(u;y))$$

- Gaussian prior $\longleftrightarrow$ quadratic regularization term.
- When μ is a Gaussian measure on $\mathcal{U}$ and the potential Φ is quadratic in u, the posterior μ^y is a well-defined Gaussian measure on $\mathcal{U}$.
- For Gaussian μ and general "nice" Φ, μ^y is a well-defined but non-Gaussian measure on $\mathcal{U}$ — lots of structure and information content to interrogate.

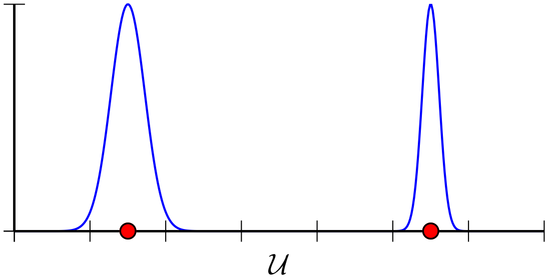
Richer Information Content

- The potential Φ or the regularized potential may have multiple minimizers, i.e. there may be multiple max. likelihood or max. a posteriori estimates of u.
- The posterior μ^y encodes more information than the location of the minimizers, e.g. how much probability mass lies near them.



Richer Information Content

- The potential Φ or the regularized potential may have multiple minimizers, i.e. there may be multiple max. likelihood or max. a posteriori estimates of u.
- The posterior μ^y encodes more information than the location of the minimizers, e.g. how much probability mass lies near them.



The Optimization–Bayesian Connection Reiterated

cost function $\Phi \longleftrightarrow$ likelihood $\exp(-\Phi)$

This concerns the observation process, the forward model (physics) $\mathcal{G}$, and the observation error/noise η.

---

regularization $\longleftrightarrow$ prior

This concerns beliefs about which u are more or less likely... or even ruled out entirely.

Overview

Well-Posedness Results

- The posterior conclusions can be shown to be robust (Lipschitz continuous in the Hellinger distance) with respect to perturbations of the data y and to N-dimensional approximation of the forward model $\mathcal{G}$, e.g.

$$d_{\mathrm{Hell}}(\mu^{y}, \mu^{y'}) \leq C\|y - y'\|_{\mathcal{Y}}.$$

  - Gaussian priors: Stuart (2010)
  - Besov priors: Dashti & al. (2012)
- It is important to *first* establish the validity of the Bayesian problem on the full spaces $\mathcal{U}$ and $\mathcal{Y}$ and *then* discretize — as is standard practice in numerical analysis.
  - Cf. the continuum wave equation has finite speed of wave propagation and so is not controllable to any desired state in finite time, whereas finite-difference schemes are controllable in thie way.

Robust Bayesian Interpretation

- What if one does not have sufficient information to commit to a single prior?
- For underdetermined inverse problems, the prior strongly influences the Bayesian solution even when the observation noise is small.
- Even with lots of high-quality data, the posterior cannot put probability where the prior does not — and this is particularly a concern in functional contexts.
- Your assumptions may be — are! — wrong to some degree. Ergo, your computational framework must be one in which the basic assumptions can be easily perturbed and turned on/off.
  - Robust Bayesian paradigm (Berger (1994))
  - Optimal Uncertainty Quantification framework (Owhadi & al. (2013))
- Optimization $\to$ Bayesian perspective $\to$ optimization again: now optimizing with respect to priors to find ranges of posterior predictions.

Robust Bayesian Interpretation

- For example, there is a huge difference between specifying the prior $\mu = \mathcal{N}(0, 1)$ and saying that any prior with mean 0 and variance 1 would be OK. The difference in the probability of a 6σ deviation can be a factor of

Robust Bayesian Interpretation

- For example, there is a huge difference between specifying the prior $\mu = \mathcal{N}(0,1)$ and saying that any prior with mean 0 and variance 1 would be OK. The difference in the probability of a 6σ deviation can be a factor of $\approx 1.4 \times 10^7$.

Robust Bayesian Interpretation

- For example, there is a huge difference between specifying the prior $\mu = \mathcal{N}(0, 1)$ and saying that any prior with mean 0 and variance 1 would be OK. The difference in the probability of a 6σ deviation can be a factor of $\approx 1.4 \times 10^7$.
- In practice, we have more numerous and complicated pieces of information than the two above. Nevertheless, it's desirable to
  - be able to calculate the posterior range of a quantity of interest with respect to all priors satisfying the known prior information, not just one simple representative;
  - to know which pieces of information (i.e. constraints) determine the typical and extreme posterior behaviour — in optimization-theoretic terms, this is the identification of binding constraints;
  - to be able to play with the constraints and identify the maximally informative next experiment.

Overview

Handling of Constraints and Other Needs

Whether we're doing an optimization, interrogating a Bayesian posterior, or doing a full RB/OUQ calculation, we need software in which

- information/constraints are enforced exactly without corrupting the physics;
- information/constraints and their enforcement methods are easily adjustable/swappable;
- calculations can be paused, adjusted, restarted;
- calculations can be distributed across heterogeneous resources, with interacting addressable and asynchronous components;

and a persistent database of results to ease the (re-)computing burden.

Handling of Constraints and Other Needs

Whether we're doing an optimization, interrogating a Bayesian posterior, or doing a full RB/OUQ calculation, we need software in which

- information/constraints are enforced exactly without corrupting the physics;
- information/constraints and their enforcement methods are easily adjustable/swappable;
- calculations can be paused, adjusted, restarted;
- calculations can be distributed across heterogeneous resources, with interacting addressable and asynchronous components;

and a persistent database of results to ease the (re-)computing burden.

From the OUQ team (McKerns & al.):

- the mystic optimization framework;
- the pathos distribution framework.

Thank You

- J. O. Berger, *Test*, 3(1):5–124, 1994.
- M. Dashti & al., *Inverse Probl. Imaging*, 6(2):183–200, 2012.
- H. Owhadi, TJS & al., *SIAM Review*, 55(2):271–345, 2013.
- A. M. Stuart, *Acta Numerica* 19:451–559, 2010.