Chengdu, September 14th, 2018

An approximate empirical Bayesian method for large-scale linear-Gaussian inverse problems

Qingping Zhou

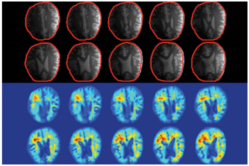
School of Mathematical Sciences

SHANGHAI JIAOTONG UNIVERSITY

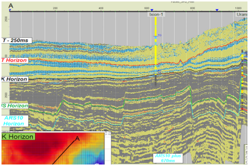
joint work with Wenqing Liu (SJTU), Jinglai Li (University of Liverpool & SJTU), and Youssef M. Marzouk (MIT)

Inverse problems

medical imaging



remote sensing



deblurring

- Various sources of uncertainty in such problems: observation noise, model error, numerical error...
  The estimation results have uncertainty!

- Data are inevitably incomplete and corrupted by noise, rendering the problem ill-posed.

Linear Gaussian inverse problems

Many real-world inverse problems, like X-ray CT, remote sensing, deblurring, can be modeled as Gaussian-linear problems:

$$\mathbf{y} = G\mathbf{x} + \boldsymbol{\eta}$$

where $\mathbf{x} \in \mathbb{R}^n$: unknown, $\mathbf{y} \in \mathbb{R}^m$: data , $\mathbf{G} \in \mathbb{R}^{m\times n}$: forward operator, $\boldsymbol{\eta}$ is observation error

Giving observations $\boldsymbol{\eta} = N(\mathbf{0}, \Gamma_{\text{obs}})$ and prior $\mathbf{x} = N(\mathbf{0}, \Gamma_{\text{pr}})$, it turns out that the posterior distribution is also Gaussian: $\mathbf{x}|\mathbf{y} \sim N(\boldsymbol{\mu}_{\text{pos}}, \Gamma_{\text{pos}})$, with

$$\boldsymbol{\mu}_{\text{pos}} = \Gamma_{\text{pos}}\, G^\top \Gamma_{\text{obs}}^{-1}\mathbf{y} \quad \text{and} \quad \Gamma_{\text{pos}} = \left(H + \Gamma_{\text{pr}}^{-1}\right)^{-1},$$

where $H = G^\top \Gamma_{\text{obs}}^{-1} G$ is the Hessian of the log-likelihood.

Empirical Bayes

- In practice, the prior and/or the likelihood function may contain unspecified hyperparameters.
- An commonly used strategy is the empirical Bayes (EB) approach, which first estimates the hyperparameters by maximizing their marginal likelihood function:

$$\theta^* = \arg\max_{\theta \in \Theta} \pi(\mathbf{y}|\theta) := \int \pi(\mathbf{y}|\mathbf{x}, \theta)\pi(\mathbf{x}|\theta) d\mathbf{x},$$

  and then plugs in the estimated values to compute the posterior of the inversion parameters:

$$\pi^*(\mathbf{x}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x}, \theta^*)\pi(\mathbf{x}|\theta^*).$$

Maximum likelihood estimaton

- It is easy to see that maximizing $\pi(\mathbf{y}|\theta)$ is equivalent to minimizing the negative log marginal likelihood:

$$\min_{\theta\in\Theta} L(\theta,\mathbf{z}) := \min_{\theta\in\Theta} -\log\pi(\mathbf{y}|\theta)$$

  and for the linear-Gaussian problems, it can be derived,

$$L(\theta,\mathbf{z}) = \frac{1}{2}\mathbf{y}^T\Gamma_{\mathrm{obs}}^{-1}\mathbf{y} + \frac{1}{2}\log|\Gamma_{\mathrm{obs}}| - \frac{1}{2}\mathbf{z}^T\Gamma_{\mathrm{pos}}\mathbf{z} + \frac{1}{2}\log\frac{|\Gamma_{\mathrm{pr}}|}{|\Gamma_{\mathrm{pos}}|},$$

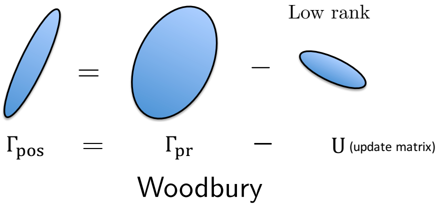
  where G, Γ_{pr} and Γ_{obs} depend on θ and $\mathbf{z} = G^T\Gamma_{\mathrm{obs}}^{-1}\mathbf{y}$.

- Direct evaluation of $L(\theta,\mathbf{z})$ is not desirable for large scale problems, as it requires several operations with $O(n^3)$ complexity.

- In what follows, we present an accurate and efficient—with $O(n^2r)$ complexity for some $r \ll n$—method to approximate $L(\theta)$, based on a rank-r update approximation of Γ_{pos}.

Low-rank approximation

- Dimension reduction: data only informative in a low-dimensional subspace of the prior state space.
  Γ_{pos} differs from Γ_{pr} in a relatively small number of directions.



Low rank

$\Gamma_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - \mathrm{U}$ (update matrix)

Woodbury

- Note that the update can be expressed as,

$$U = \Gamma_{\mathrm{pr}}\, G^{\top} \Gamma_{\mathrm{y}}^{-1} G\, \Gamma_{\mathrm{pr}}, \quad \Gamma_{\mathrm{y}} = \Gamma_{\mathrm{obs}} + G\, \Gamma_{\mathrm{pr}}\, G^{\top}.$$

- This update of Γ_{pr} is negative semidefinite (namely, $U \succ 0$), because the data add information; they cannot increase the prior variance in any direction.

Spantini et al., Optimal low-rank approximations of Bayesian linear inverse problems, *SIAM Journal on Scientific Computing* (2015).

Low-rank approximation

- Idea: to find a low-rank approximation of U, say $\hat{\mathrm{U}}$, and approximate the posterior covariance as $\widehat{\Gamma}_{\mathrm{pos}} = \Gamma_{\mathrm{pr}} - \hat{\mathrm{U}}$.
- The low-rank update:

$$\widehat{U} = \sum_{i=1}^{r} \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{\mathbf{w}}_i \widehat{\mathbf{w}}_i^\top,$$

where $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}}\, S_{\mathrm{pr}}^\top$ and $\widehat{H} = S_{\mathrm{pr}}^\top H\, S_{\mathrm{pr}}$, and $\widehat{\mathbf{w}}_i = S_{\mathrm{pr}} \mathbf{w}_i$ and $(\delta_i^2, \mathbf{w}_i)$ are the eigenvalue-eigenvector pairs of $\widehat{H}$ with the ordering $\delta_i^2 \geq \delta_{i+1}^2$.

- As a result, we obtain an approximate objective function:

$$\hat{L}(\theta, \mathbf{z}) = \frac{1}{2}\mathbf{y}^T \Gamma_{\mathrm{obs}}^{-1} \mathbf{y} + \frac{1}{2}\log|\Gamma_{\mathrm{obs}}| - \frac{1}{2}\mathbf{z}^T \widehat{\Gamma}_{\mathrm{pos}} \mathbf{z} + \frac{1}{2}\log\frac{|\Gamma_{\mathrm{pr}}|}{|\widehat{\Gamma}_{\mathrm{pos}}|},$$

Minimax optimality

Theorem

Suppose that we approximate $L(\theta, \mathbf{z})$ *with* $L'(\theta, \mathbf{z})$ *for some matrix* $\widehat{\Gamma}_{\text{pos}} \in M_r'$. *The matrix* $\widehat{\Gamma}_{\text{pos}}$ *given by*

$$\widehat{\Gamma}_{\text{pos}} = \Gamma_{\text{pr}} - \widehat{U}, \quad BB^\top = \textstyle\sum_{i=1}^r \delta_i^2 \left(1 + \delta_i^2\right)^{-1} \widehat{\mathbf{w}}_i \widehat{\mathbf{w}}_i^\top,$$

achieves the minimax approximation error, i.e., it solves

$$\min_{\widehat{\Gamma}_{\text{pos}} \in M_r'} \max_{\mathbf{z} \in Z_c} |\Delta L(\theta, \mathbf{z})|,$$

where $M_r' = \left\{ \widehat{\Gamma}_{\text{pos}} = (\Gamma_{\text{pr}} - \widehat{U}) : \widehat{\Gamma}_{\text{pos}} - \Gamma_{\text{pos}} \succeq 0, \operatorname{rank}(\widehat{U}) \leq r \right\}$, $\Delta_L = L(\theta, \mathbf{z}) - L'(\theta, \mathbf{z})$ *and* $Z_c = \{\|z\|_2 \leq C\}$, *i.e. the transformed data* $\mathbf{z}$ *is bounded.*

That is, the low-rank approximation is **optimal in the minimax sense**: the largest possible error $\max_{\mathbf{y} \in Y} |L(\theta, \mathbf{y}) - \hat{L}(\theta, \mathbf{y})|$ is minimized for all $\hat{\mathrm{U}}$ with a given rank r.

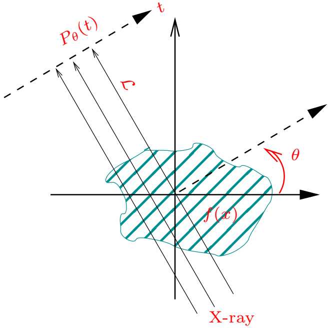
Numerical implementation

- Two intensive steps ($O(n^3)$ complexity) in computing the low rank update:
  - Compute the square root of the prior covariance: $\Gamma_{\mathrm{pr}} = S_{\mathrm{pr}}\, S_{\mathrm{pr}}^{\top}$.
  - Eigenvalue decomposition of $\widehat{H} = S_{\mathrm{pr}}^{\top} H\, S_{\mathrm{pr}}$.
- Randomized SVD requires to compute $\widehat{H}\Omega = S_{\mathrm{pr}}^{\top} H\, S_{\mathrm{pr}}\Omega$, and so we only need to compute the the matrix product $S_{\mathrm{pr}}\Omega$.
- Chebyshev spectral method for computing $S_{\mathrm{pr}}\Omega$: suppose that D is a real symmetric positive definite matrix. There exists a polynomial $p(\cdot)$ such that $\sqrt{D} = p(D)$.

Halko *et. al*, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review* (2011).

Jiang *et. al*, A fast algorithm for Brownian dynamics simulation with hydrodynamic interactions, *Mathematics of Computation* (2013).

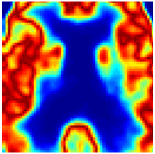
X-ray computerized tomography



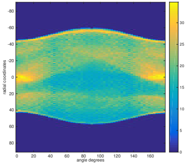
- $u(\mathbf{x})$: unknown
- $\mathcal{L}(\theta, t)$: X-ray beam
- Radon Transform:

$$\psi(t,\theta) = Af = \int_{\mathcal{L}(\theta,t)} f(\mathbf{x})|d\mathbf{x}|.$$

- the observed data y depends on the (noise-free) sinogram $\psi(t,\theta)$
- Goal: reconstruct $f(\mathbf{x})$ and **quantify the uncertainty in the reconstruction**.



$$\xRightarrow[Transform]{Radon}$$



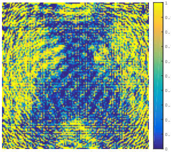
Inference results

- Use a Gaussian prior with mean μ and Matérn covariance:

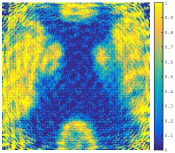
$$K(\mathbf{t},\mathbf{t}') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\, d(\mathbf{t},\mathbf{t}')\right)^{\nu} B_\nu \left(\sqrt{2\nu}\, d(\mathbf{t},\mathbf{t}')\right),$$

  where $d(\mathbf{t},\mathbf{t}') = \sqrt{(t_1 - t_1')^2/\rho_1^2 + (t_2 - t_2')^2/\rho_2^2}$. We also assume that the variance ϵ of the measurement noise is not available.

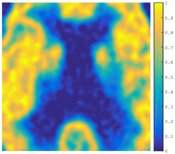
- We recover the image using the hyperparameters estimated with different ranks:



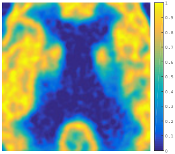
$r = 500$



$r = 2000$

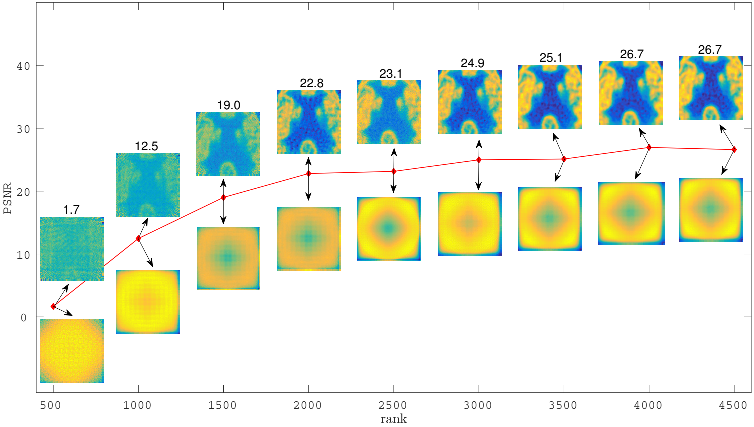


$r = 4000$



full-rank

PSNR vs Rank



1.7
12.5
19.0
22.8
23.1
24.9
25.1
26.7
26.7
PSNR
0
10
20
30
40
500
1000
1500
2000
2500
3000
3500
4000
4500
rank

Conclusions

- For large scale problems, evaluation hyperparameters by maximizing its likelihood function can be highly computationally intensive.
- we present an accurate and efficient—with $O(rn^2)$ complexity for some $r \ll n$—method to approximate marginal likelihood, based on a rank-r update approximation of posterior covariance.
- We develop a low-rank approximation method that allows us to efficiently compute the likelihood function, and we are able to show that it is the optimal approximation in the minimax sense.

Thank you for your attention