Linear Inversion via Variational Method

Bangti JIN

Universität Bremen, Zentrum für Technomathematik

Seminar, January 15, 2009

Outline

Outline

Linear inverse problem

$$\mathbf{Hm} = \mathbf{d},$$

with $\mathbf{H} \in \mathbb{R}^{n\times m}$ ill-conditioned and the noisy data $\mathbf{d}$

$$\mathbf{d} = \bar{\mathbf{d}} + \omega \rightsquigarrow \text{noise vector}$$

Least-squares method (Gauss, 1794)

$$\hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \left\{ \|\mathbf{Hm} - \mathbf{d}\|_2^2 \right\}$$

- statistically unbiased
- over-whelming oscillations: huge variance

Linear inverse problem

$$\mathbf{Hm} = \mathbf{d},$$

with $\mathbf{H} \in \mathbb{R}^{n \times m}$ ill-conditioned and the noisy data $\mathbf{d}$

$$\mathbf{d} = \bar{\mathbf{d}} + \omega \rightsquigarrow \text{noise vector}$$

Tikhonov regularization (Tikhonov, 1963)

$$\mathbf{m}_\eta = \arg\min_{\mathbf{m}} \left\{ \|\mathbf{Hm} - \mathbf{d}\|_2^2 + \eta \|\mathbf{Lm}\|_2^2 \right\}$$

- $\mathbf{L}$: (often) discretized diff. oper.
- mathematically rigorous & well-established
- choice of regularization parameter η

Linear inverse problem

$$\mathbf{Hm} = \mathbf{d},$$

with $\mathbf{H} \in \mathbb{R}^{n \times m}$ ill-conditioned and the noisy data $\mathbf{d}$

$$\mathbf{d} = \bar{\mathbf{d}} + \omega \rightsquigarrow \text{noise vector}$$

Bayesian inference (Bayes, 1764)

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m}, \tau)p(\mathbf{m}|\lambda)$$

- flexible and systematic framework
- few theoretical results
- *a priori* hyperparameters λ and τ

Hierarchical Bayesian formulations

posterior probability density function (PPDF) $p(\mathbf{m}, \lambda, \tau|\mathbf{d})$

$$p(\mathbf{m}, \lambda, \tau|\mathbf{d}) \quad \propto \quad \tau^{\frac{n}{2}} \exp\left(-\frac{\tau}{2}\|\mathbf{Hm} - \mathbf{d}\|_2^2\right) \cdot \lambda^{\frac{m}{2}} \exp\left(-\frac{\lambda}{2}\|\mathbf{Lm}\|_2^2\right)$$
$$\cdot \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \cdot \tau^{\alpha_1 - 1} e^{-\beta_1 \tau}.$$

underlying assumptions

- i.i.d. additive Gaussian noise
- Markov random field for prior model
- conjugate priors for λ and τ (Gamma distr.)

PPDF contains all info but is analytically intractable!
popular sampling methods, e.g. MCMC, are expensive

Augmented Tikhonov regularization

maximum *a posteriori* of $p(\mathbf{m}, \lambda, \tau | \mathbf{d})$: $\mathcal{J}(\mathbf{m}, \lambda, \tau)$

$$\mathcal{J}(\mathbf{m}, \lambda, \tau) = \frac{\tau}{2}\|\mathbf{Hm} - \mathbf{d}\|_2^2 + \frac{\lambda}{2}\|\mathbf{Lm}\|_2^2 + \alpha_0'\lambda - \beta_0 \ln \lambda + \alpha_1'\tau - \beta_1\tau$$

with $\alpha_0' = \alpha_0 + \frac{m}{2} - 1$ and $\alpha_1' = \alpha_1 + \frac{n}{2} - 1$.

- fcnl mimics L-curve criterion
- variance estimate similar to GCV
- consistency conditions
- point estimates only, no uncertainty quantification v.s. complete probabilistic description of PPDF

Outline

Example: variational method for diff. eq.

- 1D Poisson problem

$$-u'' = f \quad \text{on} \quad (0, 1)$$

  and $u(0) = u(1) = 0$
- numerical methods: FEM, FDM, BEM, ...
- approximate solution to optimization reformulation
- let u^* be exact solution, and define metric d as

$$d(u, u^*) = \int_0^1 (u'(x) - u^{*\prime}(x))^2 dx$$

  d is a distance.
  impractical optim.: minimizing d is no use for unknown u^*

Example: variational method for diff. eq. (cont.)

- practical optim. problem

$$
\begin{aligned}
d(u, u^*) &= \int_0^1 (u^{*\prime}(x))^2 dx - 2\int_0^1 u'(x)u^{*\prime}(x)dx + \int_0^1 (u'(x))^2 dx \\
&= const - u'(x)u^*(x)|_0^1 + \int_0^1 u^{*\prime\prime}(x)u(x)dx + \int_0^1 (u'(x))^2 dx \\
&= const - \int_0^1 f(x)u(x)dx + \int_0^1 (u'(x))^2 dx
\end{aligned}
$$

up to an unknown const, equivalent to

$$
J(u) = -\int_0^1 f(x)u(x)dx + \int_0^1 (u'(x))^2 dx
$$

Example: variational method for diff. eq. (cont.)

- approximate $u(x)$ by

$$u(x) \approx \sum_{i=1}^{n} \alpha_i \phi_i(x)$$

- finite-dim. optim. problem

$$\boldsymbol{\alpha}^* = \arg\min \left\{ 2\mathbf{b}^{\mathrm{T}}\boldsymbol{\alpha} + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{A}\boldsymbol{\alpha} \right\}$$

  with $b_i = \int_0^1 f(x)\phi_i(x)dx$ and $a_{ij} = \int_0^1 \phi_i'(x)\phi_j'(x)dx$

- key ingredients: practical optim. problem + approximation

Variational Bayesian: fundamental idea

approximate intractable PPDF $p(\mathbf{m}, \lambda, \tau|\mathbf{d})$ by simpler distr., while hopefully capturing its salient features.

Kullback-Leibler divergence $D_{KL}(q(\mathbf{m}, \lambda, \tau)|p(\mathbf{m}, \lambda, \tau|\mathbf{d}))$

$$
\begin{aligned}
D_{KL} &= \int\int\int q(\mathbf{m}, \lambda, \tau) \log \frac{q(\mathbf{m}, \lambda, \tau)}{p(\mathbf{m}, \lambda, \tau|\mathbf{d})} d\mathbf{m} d\lambda d\tau \\
&= \int\int\int q(\mathbf{m}, \lambda, \tau) \log \frac{q(\mathbf{m}, \lambda, \tau)}{p(\mathbf{m}, \lambda, \tau, \mathbf{d})} d\mathbf{m} d\lambda d\tau + \log p(\mathbf{d}),
\end{aligned}
$$

- D_{KL} unsymmetric in p and q
- Jensen inequality: $D_{KL} = 0$ iff $q(\mathbf{m}, \lambda, \tau) = p(\mathbf{m}, \lambda, \tau|\mathbf{d})$
- minimizing D_{KL} directly reprod. $p(\mathbf{m}, \lambda, \tau|\mathbf{d})$ (intractable)

Key observations

- difficulty: strong interactions between $\mathbf{m}$ and (λ, τ)
- conditional independence emerges as the key ingredient in developing approx. in probability world
- simpler distr.: separable approx. for posterior distr.

$$q(\mathbf{m}, \lambda, \tau) = q(\mathbf{m})q(\lambda, \tau) \text{ or } \delta(\mathbf{m} - \tilde{\mathbf{m}})q(\lambda, \tau)$$

- similar to 'mean-field' theoretic in stat. mechanics
- variational Bayesian = D_{KL} + separable approx!

Theorem

There exists at least one minimizer to the optimization problem.

optimality system

$$
\begin{aligned}
q^*(\mathbf{m}) &= \mathcal{N}\left(\mathbf{m}^*, (\tau^*\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\right), \\
q^*(\lambda) &= G\left(\lambda; \alpha_0'', \frac{1}{2}E_{q^*(\mathbf{m})}[\|\mathbf{L}\mathbf{m}\|_2^2] + \beta_0\right), \\
q^*(\tau) &= G\left(\tau; \alpha_1'', \frac{1}{2}E_{q^*(\mathbf{m})}[\|\mathbf{H}\mathbf{m} - \mathbf{d}\|_2^2] + \beta_1\right),
\end{aligned}
$$

$\mathcal{N} \sim$ normal distr., $G \sim$ Gamma distr.
$\tau^* = E_{q^*(\tau)}[\tau]$, $\lambda^* = E_{q^*(\lambda)}[\lambda]$ and $\eta^* = \frac{\lambda^*}{\tau^*}$
$\alpha_0'' = \frac{m}{2} + \alpha_0$ and $\alpha_1'' = \frac{n}{2} + \alpha_1$

Observations

inverse sol $\mathbf{m} \sim$ normal distri. with mean $\mathbf{m}^*$ and covariance
$\mathrm{cov}_{q^*(\mathbf{m})} = (\tau^*\mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}$
$\lambda, \tau \sim$ Gamma distr. $\Leftarrow$ conjugate prior

variance estimate

bias-variance decomposition and $\mathrm{var}(\mathbf{H}\omega) = \mathbf{H}\mathrm{var}(\omega)\mathbf{H}^{\mathrm{T}}$

$$\tau^* = \frac{\alpha_1''}{\frac{1}{2}\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 + \frac{1}{2}\mathrm{tr}((\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})\frac{1}{\tau^*} + \beta_1}.$$

rearranging the terms

$$\sigma^2(\eta^*) = \frac{\frac{1}{2}\|\mathbf{H}\mathbf{m}_{\eta^*} - \mathbf{d}\|_2^2 + \beta_1}{\alpha_1'' - \frac{1}{2}\mathrm{tr}((\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})}.$$

GCV estimate (let $\mathcal{T}(\eta) \equiv \mathrm{tr}\left(\mathbf{I}_n - \mathbf{H}(\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\right)$)

$$\mathcal{V}(\eta) = \frac{\|\mathbf{H}\mathbf{m}_\eta - \mathbf{d}\|_2^2}{\mathcal{T}(\eta)}$$

variance estimate

identity

$$\mathcal{T}(\eta) = n - \mathrm{tr}((\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{H})$$

noninformative prior for τ: $\alpha_1 \approx 1$ and $\beta_1 \approx 0$

$$\sigma^2(\eta) \approx \mathcal{V}(\eta),$$

consistency

fix τ at σ_0^{-2}

- Bakushinskii's negative result
- GCV is good for variance estimation

$$\eta^* \left[\|\mathbf{Lm}_{\eta^*}\|_2^2 + \mathrm{tr}((\mathbf{H}^{\mathrm{T}}\mathbf{H} + \eta^*\mathbf{L}^{\mathrm{T}}\mathbf{L})^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{L})\sigma_0^2 + 2\beta_0 \right] = 2\alpha_0''\sigma_0^2.$$

Lemma

There exists at least one solution and at most $2p+1$ solutions on $(0, +\infty)$.

Lemma

Assume that ω satisfies $\|\omega\|_2^2 \leq c\sigma_0^2$. Then there exist two const. $c_{r,0}$ and $c_{r,1}$ dependent on α_0'' s.t.

$$c_{r,0}\sigma_0^2 \leq \eta^* \leq c_{r,1}\sigma_0^2.$$

Theorem

Assume that ω satisfies $\|\omega\|_2^2 \leq c\sigma_0^2$. Then for fixed β_0 and $\alpha_0'' \sim \mathcal{O}(\sigma_0^{-d})$ with $0 < d < 2$, the mean $\mathbf{m}_{\eta^*}$ converges to $\mathbf{m}^+$ as σ_0 tends to zero.

Remark

The convergence of $q^*(\mathbf{m})$ to $p^+(\mathbf{m}) = \delta(\mathbf{m} - \mathbf{m}^+)$ in some probabilistic metrics, e.g. Prokhorov metric and Ky Fan metric, might also be established.

Implications

- hierarchical formulations with fixed α_0 and β_0 might fail for arbitrarily varying noise (not regularizing).
- strategies to adapt α_0 are necessary.

Choice of parameters

- variance estimate: $\alpha_1 \approx 1$, $\beta_1 \approx 0$
- convergence analysis: $\alpha_0 \sim \mathcal{O}(\sigma_0^{-d})(0 < d < 2)$, $\beta_0 \approx \mathcal{O}(\|\mathbf{Lm}^{+}\|_2^2)$

Outline

alternating iterative algorithm

strict biconvexity of $D_{KL}(q(\mathbf{m},\lambda,\tau)|p(\mathbf{m},\lambda,\tau))$:

(i) Give an initial guess $q^0(\lambda,\tau)$, and set $k=0$.

(ii) Find $q^k(\mathbf{m})$ by

$$q^k(\mathbf{m}) = \arg\min_{q(\mathbf{m})} D_{KL}(q(\mathbf{m})q^k(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)).$$

(iii) Find $q^{k+1}(\lambda,\tau)$ by

$$q^{k+1}(\lambda,\tau) = \arg\min_{q(\lambda,\tau)} D_{KL}(q^k(\mathbf{m})q(\lambda,\tau)|p(\mathbf{m},\lambda,\tau)).$$

(iv) Check the stopping criterion. If not met, set $k=k+1$, and repeat from Step (ii).

alternating iterative algorithm

optimality condition:

$$
\begin{aligned}
q^k(\mathbf{m}) &\propto \exp\left(E_{q^k(\lambda,\tau)}[\log p(\mathbf{m},\lambda,\tau)]\right) \\
&= \mathcal{N}(\mathbf{m}_{\eta_k}, \left[\tau_k \mathbf{H}^{\mathrm{T}}\mathbf{H} + \lambda_k \mathbf{L}^{\mathrm{T}}\mathbf{L}\right]^{-1}),
\end{aligned}
$$

with $\tau_k = E_{q^k(\tau)}[\tau]$, $\lambda_k = E_{q^k(\lambda)}[\lambda]$ and $\eta_k = \lambda_k \tau_k^{-1}$

$$
q^{k+1}(\lambda,\tau) \propto \exp\left(E_{q^k(\mathbf{m})}[\log p(\mathbf{m},\lambda,\tau)]\right).
$$

$$
\begin{aligned}
q^{k+1}(\lambda) &= G\left(\lambda; \alpha_0'', \frac{1}{2}E_{q^k(\mathbf{m})}[\|\mathbf{Lm}\|_2^2] + \beta_0\right), \\
q^{k+1}(\tau) &= G\left(\tau; \alpha_1'', \frac{1}{2}E_{q^k(\mathbf{m})}[\|\mathbf{Hm} - \mathbf{d}\|_2^2] + \beta_1\right).
\end{aligned}
$$

Theorem

The sequence $\left\{D_{KL}(q^k(\mathbf{m})q^k(\lambda,\tau)|p(\mathbf{m},\lambda,\tau))\right\}_k$ *decreases monotonically.*

Lemma

The sequence $\{\eta_k\}_k$ *is uniformly bounded.*

Theorem

The sequence $\{(q^k(\mathbf{m})q^k(\lambda,\tau))\}_k$ *has a subsequence converging to a stationary point of the functional* D_{KL}*.*

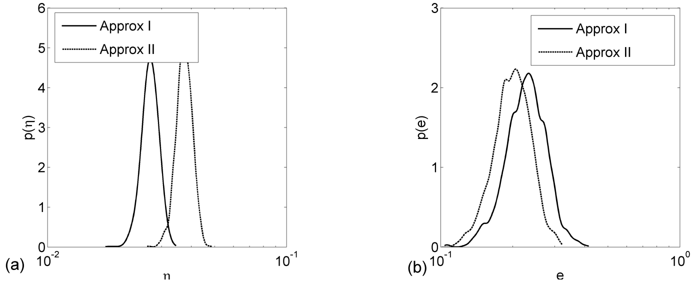
Lemma

For fixed τ *and any* η_0*, the sequence* $\{\eta_k\}_k$ *converges monotonically.*

Outline

Cauchy problem for Laplace equation

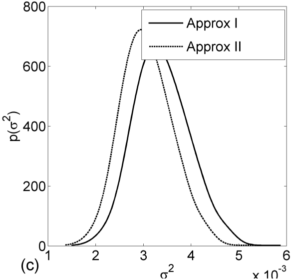
- $\Omega \subset \mathbb{R}^2$ open bdd. domain with disjoint bdry. Γ_i and Γ_c
- governing eq: $\Delta u(x) = 0$
- b.c.: Dirichlet and Neumann data on Γ_c
- inverse problem: estimate Dirichlet b.c. on Γ_i
- applications: thermal imaging, NDE and electrocardiogr.
- analysis: uniqueness, stability, ill-posedness, existence
- numerical algorithms: BGM, meshfree methods, QRM, alternating iterative algorithm ...



comparison of variational method with AT in terms of η^* and e^* (density estimated from 1000 simulations).

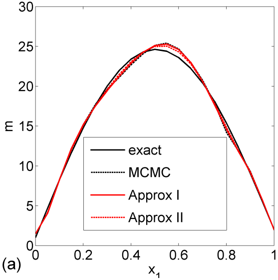
Observations

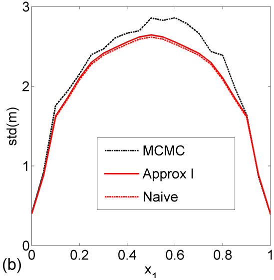
- difference in η^* is due to $E_{q^*(\mathbf{m})}[\|\mathbf{Lm}\|_2^2] \gg \|\mathbf{Lm}^*\|_2^2$
- difference in e is insignificant



variance estimates

- variational method agrees well with AT, and slightly larger

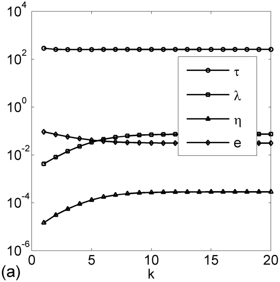




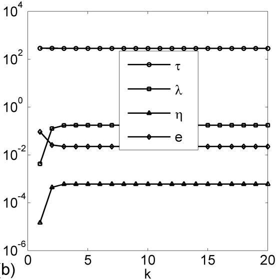
Numerical results in case of 3% noise.

Observations

- mean estimate agrees well
- both variational method and AT slightly under-est. uncertainties of the inverse solution



(a)



(b)

Convergence of the algorithm.

Observations

- AT converges faster than variation method
- variance converges within one-step

Numerical results for Example 1.

ε	σ_0	$\sigma_{\rm mc}$	$\sigma_{\rm ai}$	$\sigma_{\rm aii}$	$\eta_{\rm mc}$	$\eta_{\rm ai}$	$\eta_{\rm aii}$	$e_{\rm mc}$	$e_{\rm ai}$	$e_{\rm aii}$
1%	1.97e-2	2.10e-2	2.07e-2	1.96e-2	3.70e-5	3.58e-5	6.64e-5	2.54e-2	2.31e-2	1.85e-
3%	5.91e-2	6.28e-2	6.20e-2	5.90e-2	3.01e-4	2.87e-4	6.04e-4	3.44e-2	3.13e-2	2.27e-
5%	9.84e-2	1.05e-1	1.03e-1	9.86e-2	7.83e-4	7.52e-4	1.71e-3	3.29e-2	3.34e-2	2.11e-

Observation

The results by variational method represent better true PPDF.

Summary

Summary

- brief introduction to variational Bayes
- prelim. results about the formulation
- convergence analysis of the algorithm

Further reading list

1. Eggermont PPB. Maximum entropy regularization for Fredholm integral equations of the first kind. SIAM J. Math. Anal. 1993;24: 1557–1576.
2. Engl HW, Hanke M, Neubauer A. Regularization of Inverse Problems. Kluwer, 1996.
3. Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Machine Learning 1999;37: 183–233.
4. Smidl V, Quinn A. The Variational Bayes Method in Signal Processing. Springer: New York; 2006.
5. Jin B, Zou J. Linear inversion via variational method, submitted.