Hierarchical Bayesian Sparsity: ℓ_2 Magic.

Daniela Calvetti

Joint work with Erkki Somersalo and Alexander G Strang

Case Western Reserve University
Department of Mathematics, Applied Mathematics and Statistics

SIAM UQ, April 16-19, 2018

Linear inverse problem

Estimate $x \in \mathbb{R}^n$ with an underdetermined observation model

$$b = Ax + e, \quad A \in \mathbb{R}^{m \times n},$$

where

- $m \ll n$
- additive Gaussian noise e is a realization of random variable $E \sim \mathcal{N}(0, I_m)$
- *a priori*, x is believed to be sparse, that is

$$\|x\|_0 \ll n.$$

Sparsity promotion: hierarchical model

- Hierarchical conditionally Gaussian prior hypermodel

$$X \sim \mathcal{N}(0, \mathsf{D}_\theta), \quad \mathsf{D}_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n),$$

- Assume the prior variances $\theta_j > 0$ are mutually independent random variables following a Gamma distribution,

$$\Theta_j \sim \mathrm{Gamma}(\beta, \theta_j^*) \propto \theta_j^{\beta-1} \exp\left(-\frac{\theta_j}{\theta_j^*}\right), \quad 1 \leq j \leq n.$$

- Posterior density

$$\pi_{X,\Theta|B}(x, \theta) \propto \exp\left(-\frac{1}{2}\|b - \mathsf{A}x\|^2 - \frac{1}{2}\sum_{j=1}^{n}\frac{x_j^2}{\theta_j} + \eta\sum_{j=1}^{n}\log\theta_j - \sum_{j=1}^{n}\frac{\theta_j}{\theta_j^*}\right)$$

where $\eta = \beta - 3/2 > 0$.

Iterared Alternating Sequential (IAS) algorithm

To compute x_{MAP} we minimize the Gibbs energy

$$\mathscr{E}(x;\theta) = \overbrace{\frac{1}{2}\|b - \mathsf{A}x\|^2 + \underbrace{\sum_{j=1}^{n} \frac{x_j^2}{2\theta_j}}_{}}^{(a)} \underbrace{- \sum_{j=1}^{n} \left(\eta \log \theta_j - \frac{\theta_j}{\theta_j^*} \right)}_{(b)} \tag{1}$$

Given the initial value $\theta^0 = \theta^*$, $x^0 = 0$, and $k = 0$, iterate until convergence:

(a) Update $x^k \to x^{k+1}$ by minimizing $\mathscr{E}(x;\theta^k)$;

(b) Update $\theta^k \to \theta^{k+1}$ by minimizing $\mathscr{E}(x^{k+1};\theta)$;

(c) Increase $k \to k+1$.

Exact IAS algorithm

Initialize: $k = 0$, $\theta_0 = \theta^*$;

While $\|\theta_k - \theta_{k-1}\| > \mathrm{tol}$

1. Update x; $x_{k+1} = \mathrm{argmin}\left\{\|b - \mathsf{A}x\|^2 + \|\mathsf{D}_\theta^{-1/2}x\|^2\right\}$ by solving

$$\begin{bmatrix} \mathsf{A} \\ \mathsf{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

   in the least squares sense.

2. Setting $x = x_{k+1}$, update the components of θ_{k+1} according to the formula

$$\theta_j = \theta_j^* \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{x_j^2}{2\theta_j^*}} \right)$$

Convexity and Convergence of exact IAS

The exact IAS algorithms with the Gamma hyperprior is such that:

- The Gibbs energy functional is strictly convex
- The Gibbs functional has a unique minimizer
- If Step 1 and Step 2 are solved exactly the algorithm converges to the global minimizer
- For $\eta > 0$ small, the Gibbs energy (1) is approximately equal to the penalized least squares functional with a weighted ℓ_1-penalty [1].

.

[1] Calvetti D, Pascarella A, Pitolli F, Somersalo E, Vantaggi B (2015) A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning. Inverse Problems 31:125005

A convergence result

Let the function for updating the variance $f : \mathbb{R}^n \to \mathbb{R}^n$ have components

$$f_j(x_j) = \theta_j^* \left(\frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} + \frac{x_j^2}{2\theta_j^*}} \right)$$

Theorem

For a Gamma hyperprior, the exact IAS algorithm converges to the unique minimizer $(\widehat{x}, \widehat{\theta})$ of the Gibbs energy functional. Moreover, the minimizer $(\widehat{x}, \widehat{\theta})$ satisfies the fixed point condition

$$\widehat{x} = \mathrm{argmin}\left\{ \mathscr{E}\big(x \mid F(x)\big) \right\}, \quad \widehat{\theta} = F(\widehat{x}),$$

where F is the map with jth component f_j.[2].

[2] Calvetti D, Pascarella A, Pitolli F, Somersalo E, Vantaggi B (2015) A hierarchical Krylov–Bayes iterative inverse solver for MEG with physiological preconditioning. Inverse Problems 31:125005

Scale parameter and sparsity

Under the assumptions of our hierarchical Bayesian model we have shown that

- The exact IAS iteration converges to the global minimizer of the functional

$$\mathscr{L}_\eta(x) = \mathscr{E}(x, f(x))$$

and, for small $\eta > 0$

- $$\mathscr{L}_\eta(x) = \mathscr{L}_0(x) + \underbrace{\eta g(x, \eta)}_{\to 0 \text{ as } \eta \to 0},$$

where

$$\mathscr{L}_0(x) = \frac{1}{2}\|b - \mathsf{A}x\|^2 + \sqrt{2}\sum_{j=1}^{n} \frac{|x_j|}{\sqrt{\theta_j^*}}.$$

Gamma hyperprior parameters

From the result above it follows that

1. For small η, the IAS minimization problem is a small perturbation of the weighted ℓ_1 penalized least squares functional
2. The parameter η controls the sparsity of the solution.
3. The scale parameters θ_j^* play the role of sensitivity weights in inverse problems
4. Data components may have different sensitivity to different components x_j.

ℓ_2 Stable Signal Recovery

Two remarks

$$\underbrace{x_\eta = \text{argmin}\left\{\mathscr{L}_\eta(x)\right\}}_{=IAS\ \ solution} \qquad \underbrace{x_1 = \text{argmin}\left\{\mathscr{L}_0(x)\right\}}_{=\ell_1 penalized\ \ solution}.$$

1. The difference $x_\eta - x_1$ is a vector whose size depends continuously on η.
2. If A has the Restricted Isometry Property (RIP) and the data comes from a sparse vector[3], then x_η is close to the underlying sparse solution.

[3]Candes E, Romberg JK and Tao T(2006): Stable Signal Recovery from Incomplete and Inaccurate Measurements, Comm Pure Appl Math LIX: 1207–1223

Sparse signal and exchangeability

Assume the underlying signal x is sparse $\mathrm{supp}(x) = \mathsf{I} \subset \{1, 2, \ldots, n\}$ and b_0 is the noiseless measurement. Define

$$\mathrm{SNR}_{|_\mathsf{I}} = \frac{E\left\{\|b_0\|^2 \mid \mathrm{supp}(x) = \mathsf{I}\right\}}{E\left\{\|e\|^2\right\}}.$$

Lemma

With our assumptions about X and E

$$\mathrm{SNR}_{|_\mathsf{I}} = \frac{\sum_{j\in\mathsf{I}} \beta\theta_j^* \|\mathsf{A}e_j\|^2}{\mathrm{tr}\left(\Sigma\right)}.$$

Proof.

$$E\left\{\|b_0\|^2\right\} = \mathrm{Tr} E\left\{b_0 b_0^{\mathrm{T}}\right\} = \mathrm{Tr} E\left\{\mathsf{A}xx^{\mathrm{T}}\mathsf{A}^{\mathrm{T}}\right\} = \mathrm{Tr}\big(\mathsf{A}E\left\{xx^{\mathrm{T}}\right\}\mathsf{A}^{\mathrm{T}}\big),$$

and from the Gamma hyperprior

$$E\left\{xx^{\mathrm{T}}\right\} = E_\theta\left\{E\left\{xx^{\mathrm{T}} \mid \theta\right\}\right\} = E\big(\mathrm{diag}(\theta)\big) = \mathrm{diag}\big(\beta\theta^*\big).$$

Choice of scale parameter

How should θ^* be chosen?

Theorem

Given an estimate $\overline{\mathrm{SNR}}$ *of SNR, if*

$$P\big(\|x\|_0 = k\big) = p_k, \quad p_0 = 0, \quad \sum_{k=1}^{n} p_k = 1$$

and if

$$\mathrm{SNR}\,|_{\mathsf{I}} = \mathrm{SNR}\,|_{\mathsf{I}'}, \quad \forall\, \mathsf{I}, \mathsf{I}' : \mathrm{card}(\mathsf{I}) = \mathrm{card}(\mathsf{I}'),$$

then

$$\theta_j^* = \frac{\mathsf{C}}{\|\mathsf{A}e_j\|^2}, \quad \mathsf{C} = \overline{\mathrm{SNR}}\mathrm{Tr}(\Sigma) \sum_{j=1}^{n} \frac{p_k}{k}$$

In the literature $\|\mathsf{A}e_j\|^2$ is the *sensitivity* of the data to jth component of x.

Sparsity and quadratic convergence

Theorem

If

- $f_j(x_j)$ *is monotonically increasing*
- $f_j(0) \geq \tilde{\theta} > 0$
- x_* *is sparse with* $\mathrm{supp}(x_*) = T, \ |T| = S$

then exact IAS converges quadratically in θ_{T^c}, *where* T^C *is the complement of* T.

Theorem

Under the conditions of the previous theorem, if

- x_* *is nearly sparse (compressible)*
- $\|x_{T^c}\|_\infty \leq \zeta$

then

$$J_{T^c} \leq \sqrt{S}\frac{\zeta}{\tilde{\theta}}.$$

Sparsity and quadratic convergence

The theorems implies that

- If ζ is small enough, the convergence is effectively quadratic in θ_{T^c}
- The ℓ_1 solution from noisy data approximates well the underlying sparse signal[4],
- As $\beta \to (3/2)^+$, the exact IAS solution approaches the ℓ_1 penalized solution hence is close to the underlying sparse signal.

[4] Candes E, Romberg JK and Tao T(2006): Stable Signal Recovery from Incomplete and Inaccurate Measurements, Comm Pure Appl Math LIX: 1207–1223

Inexact IAS and quasi-MAP estimate

1. For large scale problems and few observations ($\mathsf{A} \in \mathbb{R}^{m\times n},\ m < n$), the least squares step of IAS

$$\begin{bmatrix} \mathsf{A} \\ \mathsf{D}_\theta^{-1/2} \end{bmatrix} x = \begin{bmatrix} b \\ 0 \end{bmatrix}$$

can be solved approximately by priorconditioned CGLS[5] [6]

2. If $\Sigma^{-1} = \mathsf{S}^{\mathrm{T}}\mathsf{S}$, apply the CGLS method to

$$\mathsf{SAD}_\theta^{1/2} w = \mathsf{S}b$$

stopping on Morozov discrepancy principle at jth step, where

$$d_j > \sqrt{m} > d_{j+1}, \quad d_j = \|\mathsf{S}b - \mathsf{SAD}_\theta^{1/2} w_j\|.$$

3. Retrieve original variable $x_k = \mathsf{D}_\theta^{1/2} w_j$

[5] Calvetti et al. Priorconditioned CGLS-Based Quasi-MAP Estimate, Statistical Stopping Rule, and Ranking of Priors. SIAM J. Sci. Comput. 39-5 (2017)

[6] Calvetti et al. Bayes meets Krylov: preconditioning CGLS for underdetermined systems. D Calvetti, F Pitolli, E Somersalo, B Vantaggi. SIAM Review, to appear.

Pros of inexact IAS

1. Suitable for large problems
2. Computationally efficient
3. Follows classical scheme of inner/outer iterations
4. Can be interpreted as flexible right preconditioning
5. Number of CGLS steps decreases with outer iterations
6. When the underlying signal is sparse, the algorithm automatically reduced the effective dimensionality of the problem to solve

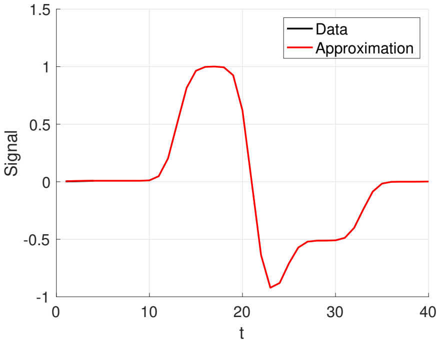
Open questions for inexact IAS

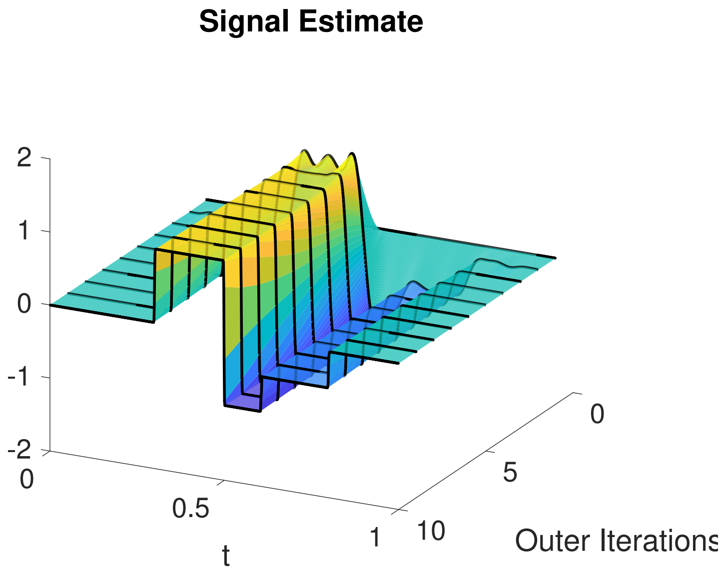
At the present, the inexact IAS

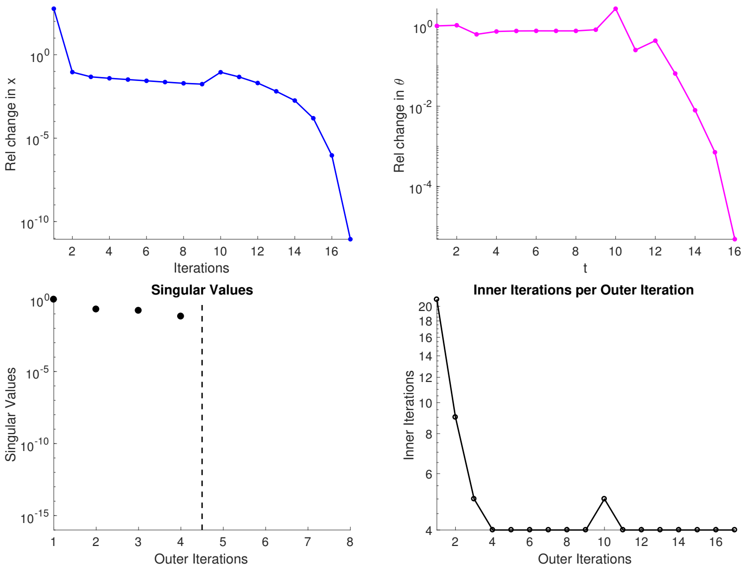
1. Does not have a proof of convergence, only numerical evidence
2. The solution produced is an approximation of the MAP estimate, hence we call it *quasi* MAP estimate (qMAP)
3. A quadratic rate of convergence of signal and prior variances has been observed numerically: a rigorous proof is in progress.

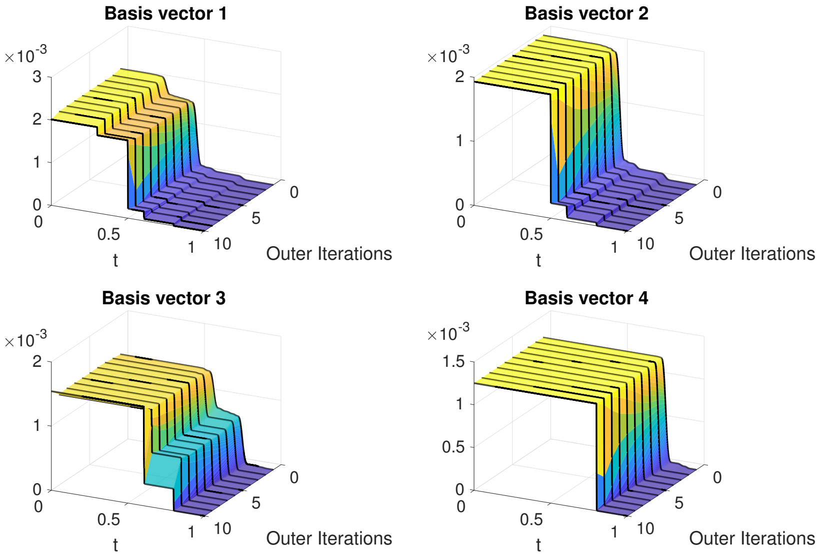
Computed examples

- $A \in 1000 \times 1000$ is a Gaussian blurring kernel over 15 pixels
- Data come from a piecewise constant signal
- Additive scaled white noise 0.01% of max of noise-free data
- Conditionally Gaussian first order smoothness prior
- Gamma hyperprior


1.5
1
0.5
0
-0.5
-1
0
10
20
30
40
t
Signal
Data
Approximation
1.5
1
0.5
0
-0.5
-1
0
0.2
0.4
0.6
0.8
1
t
Signal
Original Signal
Last Iterate


Signal Estimate
2
1
0
-1
-2
0
0.5
1
t
0
5
10
Outer Iterations

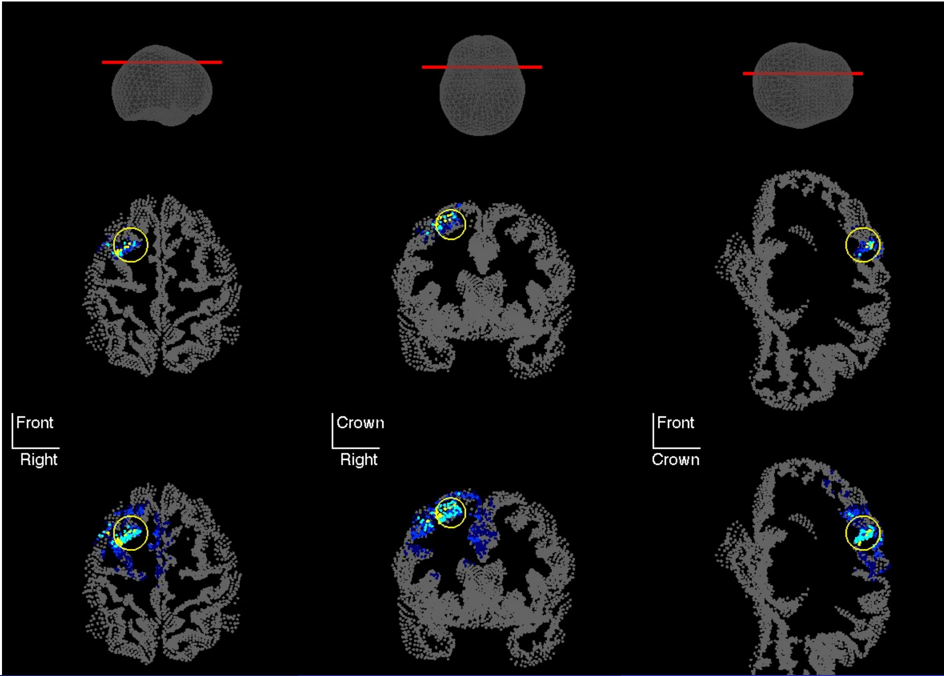

Rel change in x
Iterations
Rel change in θ
t
Singular Values
Singular Values
Outer Iterations
Inner Iterations per Outer Iteration
Inner Iterations
Outer Iterations


Basis vector 1
×10^-3
3
2
1
0
0
0.5
1
t
0
5
10
Outer Iterations
Basis vector 2
×10^-3
2
1
0
0
0.5
1
t
0
5
10
Outer Iterations
Basis vector 3
×10^-3
2
1
0
0
0.5
1
t
0
5
10
Outer Iterations
Basis vector 4
×10^-3
1.5
1
0.5
0
0
0.5
1
t
0
5
10
Outer Iterations

An application to MEG

In this example we see the effect of the focality parameter

- Data: 153 measurements at magnetometers
- A is the leadfield matrix 153×75000
- Sparsity prior


Front
Right
Crown
Right
Front
Crown