



DARTMOUTH
Guarini School of Graduate
and Advanced Studies

GENERALIZED HYBRID SOLVERS

for sparsity-promoting Bayesian inverse problems

Advancement exam talk

17th November 2022

Jonathan Lindbloom[†], Jan Glaubitz, and Anne Gelb



OUTLINE

1. Bayesian inverse problems
2. Hierarchical models
3. Solvers

BAYESIAN INVERSE PROBLEMS

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\text{noise}}^{-1}).$$

$$\mathbf{y} = \mathbf{F}\mathbf{x} + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{\text{noise}}^{-1}).$$

Here $\mathbf{y}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ with forward/measurement operator $\mathbf{F} \in \mathbb{R}^{m \times n}$ and SPD noise precision $\mathbf{Q}_{\text{noise}} \in \mathbb{R}^{m \times m}$.

Obstacles:

1. Presence of noise ϵ
2. $\mathbf{F}^{-1}$ may not exist
3. We may be given $\mathbf{y}$ but not $\mathbf{Q}_{\text{noise}}$

LINEAR FORWARD OPERATORS

Ground truth



Blur



Under-sampling



Up-sampling



Many more, such as Fourier, linear ODE/PDE, etc.

$$\pi(\mathbf{y} \mid \mathbf{x}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 \right\}.$$

Limiting to constant-times-identity precisions of the form $\mathbf{Q}_{\text{noise}} = \sigma^{-2} \mathbf{I}$, the likelihood for the observed data $\mathbf{y}$ is Gaussian with

$$\pi(\mathbf{y} \mid \mathbf{x}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 \right\}.$$

Why not just maximize w.r.t $\mathbf{x}$?

Maximum likelihood estimate

$$\begin{aligned}\mathbf{x}_{\text{MLE}} &= \arg \max_x \left\{ \pi(\mathbf{y} \mid \mathbf{x}) \right\} \\ &= \left(\mathbf{F}^T \mathbf{F} \right)^{-1} \mathbf{F}^T \mathbf{y}\end{aligned}$$

LEAST-SQUARES

Observation (no noise)



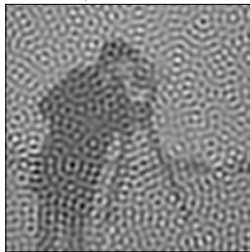
Least-squares reconstruction



Observation (with noise)



Least-squares reconstruction



Bayes' theorem

$$\overbrace{\pi(\mathbf{x} \mid \mathbf{y})}^{\text{posterior}} = \frac{\pi(\mathbf{y} \mid \mathbf{x}) \pi(\mathbf{x})}{\int \pi(\mathbf{y} \mid \mathbf{x}) \pi(\mathbf{x}) d\mathbf{x}}$$

$$\propto \underbrace{\pi(\mathbf{y} \mid \mathbf{x})}_{\text{likelihood}} \underbrace{\pi(\mathbf{x})}_{\text{prior}}$$

Maximum a posteriori (MAP) estimate

C

A common choice is to pick

$$\pi_{\text{prior}}(\mathbf{x}) \propto \exp \left\{ -\frac{\lambda}{2} \|\mathbf{R}\mathbf{x}\|_2^2 \right\}$$

with and $\mathbf{R} \in \mathbb{R}^{k \times n}$ a transformation we believe should sparsify the latent signal $\mathbf{x}$.

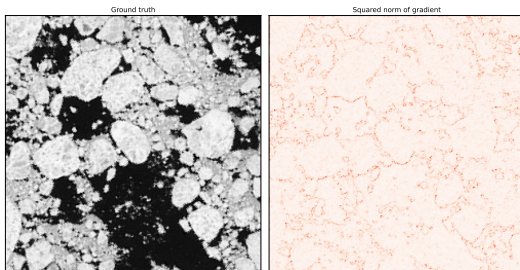


Figure: Discrete gradient operator.

Suppose that

$$\begin{aligned} \mathbf{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\lambda \mathbf{R}^T \mathbf{R}\right)^{-1}\right), \\ \mathbf{y} \mid \mathbf{x} &\sim \mathcal{N}\left(\mathbf{F} \mathbf{x}, \mathbf{Q}_{\text{noise}}^{-1}\right). \end{aligned}$$

Suppose that

$$\begin{aligned} \mathbf{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\lambda \mathbf{R}^T \mathbf{R}\right)^{-1}\right), \\ \mathbf{y} \mid \mathbf{x} &\sim \mathcal{N}\left(\mathbf{F} \mathbf{x}, \mathbf{Q}_{\text{noise}}^{-1}\right). \end{aligned}$$

$$\boldsymbol{x} \mid \boldsymbol{y} \sim \mathcal{N} \left(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{Q}_{\text{post}}^{-1} \right)$$
$$\begin{aligned} \mathbf{Q}_{\text{post}} &= \mathbf{F}^T \mathbf{Q}_{\text{noise}} \mathbf{F} + \lambda \mathbf{R}^T \mathbf{R}, \\ \boldsymbol{\mu}_{\text{post}} &= \mathbf{Q}_{\text{post}}^{-1} \left(\mathbf{F}^T \mathbf{Q}_{\text{noise}} \mathbf{y} \right). \end{aligned}$$

$$\pi_{\text{post}}(\mathbf{x}) \propto \pi_1(\mathbf{x}) \times \cdots \times \pi_K(\mathbf{x}), \quad \pi_j(\mathbf{x}) \propto \exp \left\{ -\frac{1}{2} \|\mathbf{F}_j \mathbf{x} - \mathbf{y}_j\|_{\mathbf{Q}_j}^2 \right\}.$$

More generally, suppose we are interested in a posterior of the form

$$\pi_{\text{post}}(\mathbf{x}) \propto \pi_1(\mathbf{x}) \times \cdots \times \pi_K(\mathbf{x}), \quad \pi_j(\mathbf{x}) \propto \exp \left\{ -\frac{1}{2} \|\mathbf{F}_j \mathbf{x} - \mathbf{y}_j\|_{\mathbf{Q}_j}^2 \right\}.$$

Gaussian posterior

The posterior is the Gaussian $\mathcal{N}(\mu_{\text{post}}, Q_{\text{post}}^{-1})$ with

$$\begin{aligned} Q_{\text{post}} &= \sum_{j=1}^K \mathbf{F}_j^T \mathbf{Q}_j \mathbf{F}_j, \\ \mu_{\text{post}} &= Q_{\text{post}}^{-1} \left(\sum_{j=1}^K \mathbf{F}_j^T \mathbf{Q}_j \mathbf{y}_j \right). \end{aligned}$$

Additionally, $\mathbf{x}_{\text{MAP}} = \boldsymbol{\mu}_{\text{post}}$.

$\lambda = 1000$ 

$$\arg \min_x \left\{ \frac{1}{2\sigma^2} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{\lambda}{2} \|\mathbf{R}\mathbf{x}\|_2^2 \right\}$$

WHY WE LIKE GAUSSIANS

- Explicit expression for normalized posterior density and moments

¹Vono, Dobigeon and Chainais, "High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm", 2022.

²Rue and Held, Gaussian Markov Random Fields: Theory and Applications, 2005.

³Parker and Fox, "Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients", 2012.

WHY WE LIKE GAUSSIANS

- Explicit expression for normalized posterior density and moments
- For MAP estimate, just need to solve system $\mathbf{Q}_{\text{post}} \boldsymbol{\mu} = \text{"rhs"}$

¹Vono, Dobigeon and Chainais, “High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm”, 2022.

²Rue and Held, Gaussian Markov Random Fields: Theory and Applications, 2005.

³Parker and Fox, "Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients", 2012.

WHY WE LIKE GAUSSIANS

- Explicit expression for normalized posterior density and moments
- For MAP estimate, just need to solve system $\mathbf{Q}_{\text{post}}\boldsymbol{\mu} = \text{"rhs"}$
- Sampling high-dimensional Gaussians has been studied extensively [1]
 - Factorization approaches, e.g., the Cholesky approach costing $\mathcal{O}(pn^2)$ [2]
 - Inverse square root approximations
 - Conjugate gradient approaches, e.g., [3]
 - Markov chain Monte Carlo (MCMC)

¹Vono, Dobigeon and Chainais, "High-Dimensional Gaussian Sampling: A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm", 2022.

²Rue and Held, Gaussian Markov Random Fields: Theory and Applications, 2005.

³Parker and Fox, "Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients", 2012.

GAUSSIAN POSTERIOR SAMPLES



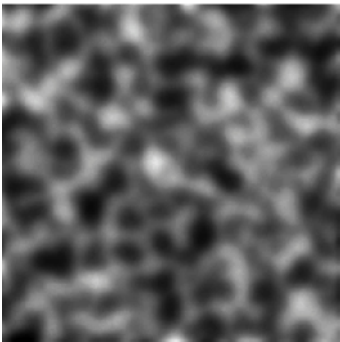


Figure: Gaussian prior sample.

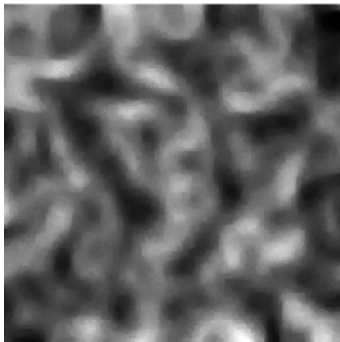


Figure: Total variation (TV)
prior sample.



Figure: Cauchy difference prior sample.

⁴Markkanen et al., "Cauchy difference priors for edge-preserving Bayesian inversion", 2019.

NON-GAUSSIAN PRIORS

- Can more strongly promote sparsity under R 👍

NON-GAUSSIAN PRIORS

- Can more strongly promote sparsity under R 👍
- No longer have explicit expression for the posterior, MAP, etc. 👎
- Posterior may no longer be log-concave 👎

$$E(\mathbf{x}) := -\log \pi(\mathbf{x} \mid \mathbf{y}) = \underbrace{\frac{1}{2} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_Q^2}_{\text{strongly convex}} + \underbrace{-\log \pi(\mathbf{x})}_{\text{convex?}}$$

NON-GAUSSIAN PRIORS

- Can more strongly promote sparsity under R 👍
- No longer have explicit expression for the posterior, MAP, etc. 👎
- Posterior may no longer be log-concave 👎

$$E(\mathbf{x}) := -\log \pi(\mathbf{x} \mid \mathbf{y}) = \underbrace{\frac{1}{2} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_Q^2}_{\text{strongly convex}} \underbrace{- \log \pi(\mathbf{x})}_{\text{convex?}}$$

- A common example: ℓ_p -norm priors for $0 < p < 2$,

$$\pi(\mathbf{x}) \propto \exp \left\{ -\lambda \|\mathbf{R}\mathbf{x}\|_p^p \right\}$$

Gaussian prior

A grayscale image showing a person in a dark coat standing next to a camera on a tripod. The image is very blurry, representing a Gaussian prior. The person is holding the camera, and the background is indistinct.

Laplace prior (TV)



Laplace prior (TV)

HIERARCHICAL MODELS

MOTIVATION

1. Provides a framework for working with non-Gaussian, sparsity-promoting priors
2. Can also allow us to learn noise covariance parameters with uninformative priors

An interesting fact:

$$w \sim \text{Exponential} \left(2^{-1} b^{-2} \right),$$

An interesting fact:

$$w \sim \text{Exponential} \left(2^{-1} b^{-2} \right),$$

$$x \mid w \sim \mathcal{N}(0, w),$$

$$w \sim \text{Exponential} \left(2^{-1} b^{-2} \right),$$

$$x \mid w \sim \mathcal{N}(0, w),$$

$$\Rightarrow x \sim \text{Laplace}(0, b).$$

x is marginally Laplace, but conditionally-Gaussian. In terms of densities,

$$\pi(x) = \int \pi(x, w) dw = \int \underbrace{\pi(x | w)}_{\text{Gaussian}} \times \underbrace{\pi(w)}_{\text{Exponential}} dw.$$

THE GENERALIZED GAMMA DISTRIBUTION

We can represent a wide family of sparsity-promoting densities as a scale mixture of a Gaussian and a Generalized Gamma random variable with density function

$$\pi_{\text{GG}}(\theta \mid r, s, \vartheta) = \frac{|r|}{\vartheta \Gamma(s)} \left(\frac{\theta}{\vartheta}\right)^{rs-1} \exp\left\{-\left(\frac{\theta}{\vartheta}\right)^r\right\} \mathbb{1}_{>0}(\theta)$$

for parameters $r \in \mathbb{R} \setminus \{0\}$, $s > 0$, $\vartheta > 0$.

- Encompasses exponential, Gamma, inverse Gamma distributions
- Parameters (r, s) set the parameteric family of marginal prior, ϑ controls scale

⁵Calvetti et al., "Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors", 2020.

THE GENERALIZED GAMMA DISTRIBUTION

We can represent a wide family of sparsity-promoting densities as a scale mixture of a Gaussian and a Generalized Gamma random variable with density function

$$\pi_{\text{GG}}(\theta \mid r, s, \vartheta) = \frac{|r|}{\vartheta \Gamma(s)} \left(\frac{\theta}{\vartheta}\right)^{rs-1} \exp\left\{-\left(\frac{\theta}{\vartheta}\right)^r\right\} \mathbb{1}_{>0}(\theta)$$

for parameters $r \in \mathbb{R} \setminus \{0\}$, $s > 0$, $\vartheta > 0$.

- Encompasses exponential, Gamma, inverse Gamma distributions
- Parameters (r, s) set the parameteric family of marginal prior, ϑ controls scale
- In certain limits, captures ℓ_p -norm priors for $0 < p < 2$, Student- t , Cauchy
- Detailed convexity analysis provided by [5]

⁵Calvetti et al., "Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors", 2020.

An unknown $\mathbf{x} = (x_1, x_2, x_3)$, with a prior on pixel differences

$x_1 = -5$	x_2	$x_3 = +5$
------------	-------	------------

What does the (univariate) conditional

$$\pi(x_2 \mid x_1 = -5, x_3 = 5)$$

look like?



VS.

$$\mathbf{y} \mid \mathbf{x}, \alpha \sim \mathcal{N}(\mathbf{F}\mathbf{x}, \alpha \mathbf{I})$$

VS.

$$x|\beta \sim \mathcal{N}\left(\mathbf{0}, \left(R^T D_{\beta}^{-1} R\right)^{-1}\right)$$

VS.

NEGATIVE LOG POSTERIOR

In the simpler Gaussian prior/likelihood case:

$$E(\mathbf{x}) = \frac{1}{2\alpha} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{2\beta} \|\mathbf{R}\mathbf{x}\|_2^2$$

In our hierarchical model:

$$\begin{aligned} E(\mathbf{x}, \alpha, \beta) &= E_1(\mathbf{x}, \alpha) + E_2(\mathbf{x}, \beta), \\ E_1(\mathbf{x}, \alpha) &= \frac{1}{2\alpha} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 - \left(r_1 s_1 - \frac{m+2}{2}\right) \log \alpha + \left(\frac{\alpha}{\vartheta_1}\right)^{r_1}, \\ E_2(\mathbf{x}, \beta) &= \frac{1}{2} \|\mathbf{R}\mathbf{x}\|_{\mathbf{D}_\beta^{-1}}^2 - \left(r_2 s_2 - \frac{3}{2}\right) \sum_{j=1}^k \log \beta_j + \sum_{j=1}^k \left(\frac{\beta_j}{\vartheta_2}\right)^{r_2} \end{aligned}$$

SOLVERS

☰ On this page

User Guide

Developer Guide

[Show Source](#)

Overview

Installation

Mathematical Background

Multivariate Gaussians

Bayesian inference

Gaussian Factors and Posteriors

Examples

Projects

- ## 1. Generalized hybrid solvers

API Reference

ALGORITHMS FOR POINT ESTIMATION

- Seeking posterior mean:
 - For inverse Gamma variance hyper-prior, conjugacy relations exist that motivate a Bayesian coordinate descent (BCD) algorithm⁶ (block updates to conditional means)
 - Approximation via MCMC samples, design of transition kernel critical for high dimensions
- Seeking MAP point:
 - Iterating alternating sequential (IAS) algorithm, many relevant works (e.g., [7])
 - Simulated annealing, gradient descent, proximal methods, etc.

⁶Glaubitz, Gelb and Song, Generalized sparse Bayesian learning and application to image reconstruction, 2022.

⁷Calvetti, Somersalo and Strang, "Hierarchical Bayesian models and sparsity: ℓ_2 -magic", 2019; Calvetti et al., "Sparse reconstructions from few noisy data: analysis of hierarchical Bayesian models with generalized gamma hyperpriors", 2020; Calvetti, Pragliola and Somersalo, "Sparsity Promoting Hybrid Solvers for Hierarchical Bayesian Inverse Problems", 2020.

ITERATING ALTERNATING SEQUENTIAL (IAS) ALGORITHM

IAS Algorithm

Choose initial $\mathbf{x}^0, \alpha^0, \beta^0$.

for $k = 1, \dots, n_{\text{maxits}}$ do

$\alpha^k \leftarrow \arg \max_{\alpha} \pi(\alpha \mid \mathbf{x}^{k-1}, \beta^{k-1}, \mathbf{y})$.

$\beta^k \leftarrow \arg \max_{\beta} \pi(\beta \mid \alpha^k, \mathbf{x}^{k-1}, \mathbf{y})$.

$\mathbf{x}^k \leftarrow \arg \max_{\mathbf{x}} \pi(\mathbf{x} \mid \alpha^k, \beta^k, \mathbf{y})$.

end for

BAYESIAN COORDINATE DESCENT (BCD) ALGORITHM

BCD Algorithm

Choose initial $\mathbf{x}^0, \alpha^0, \beta^0$.

for $k = 1, \dots, n_{\text{maxits}}$ do

$$\alpha^k \leftarrow \mathbb{E} \left[\alpha \mid \mathbf{x}^{k-1}, \boldsymbol{\beta}^{k-1}, \mathbf{y} \right].$$

$$\boldsymbol{\beta}^k \leftarrow \mathbb{E} \left[\boldsymbol{\beta} \mid \mathbf{x}^{k-1}, \alpha^k, \mathbf{y} \right].$$

$$\mathbf{x}^k \leftarrow \mathbb{E} \left[\mathbf{x} \mid \alpha^k, \beta^k, \mathbf{y} \right].$$

end for

UPDATES

Each $\mathbf{x}$ -update is the solution to

$$\arg \min_{\mathbf{x}} \left\{ \frac{1}{2\alpha} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{2} \|\mathbf{R}\mathbf{x}\|_{\mathbf{D}_{\beta}^{-1}}^2 \right\} = \left(\frac{1}{\alpha} \mathbf{F}^T \mathbf{F} + \mathbf{R}^T \mathbf{D}_{\beta}^{-1} \mathbf{R} \right)^{-1} \left(\frac{1}{\alpha} \mathbf{F}^T \mathbf{y} \right)$$

UPDATES

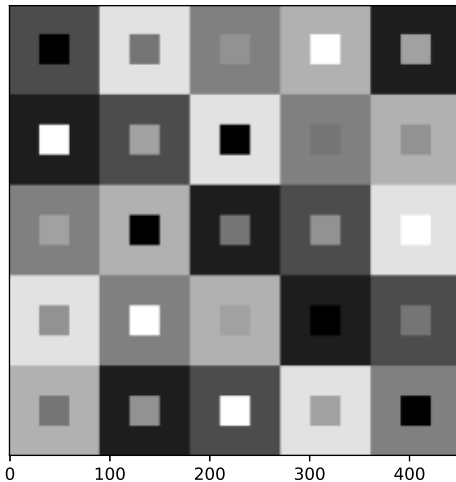
Each $\mathbf{x}$ -update is the solution to

$$\arg \min_{\mathbf{x}} \left\{ \frac{1}{2\alpha} \|\mathbf{F}\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{2} \|\mathbf{R}\mathbf{x}\|_{\mathbf{D}_{\beta}^{-1}}^2 \right\} = \left(\frac{1}{\alpha} \mathbf{F}^T \mathbf{F} + \mathbf{R}^T \mathbf{D}_{\beta}^{-1} \mathbf{R} \right)^{-1} \left(\frac{1}{\alpha} \mathbf{F}^T \mathbf{y} \right)$$

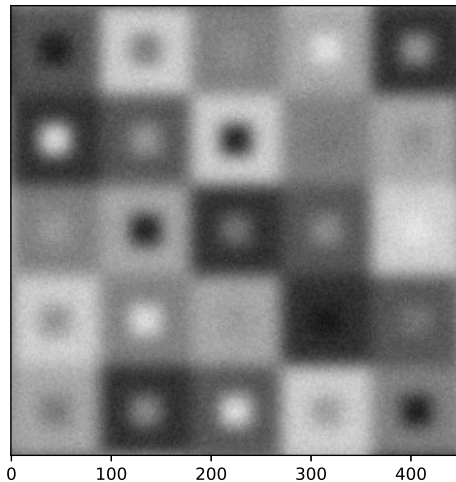
Updating α, β ?

- Conjugacy relations can give analytic expressions for conditional mode/mean
- Without conjugacy relations, conditional modes are generally solutions to an associated IVP
- Without conjugacy relations, conditional means may still be known analytically

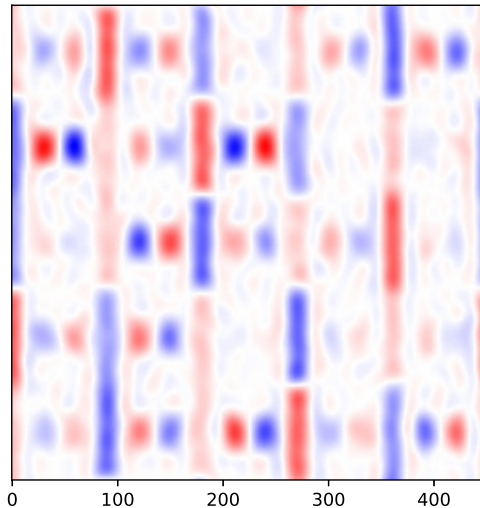
Ground truth

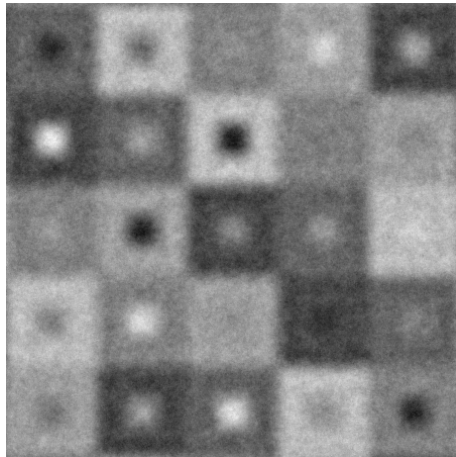
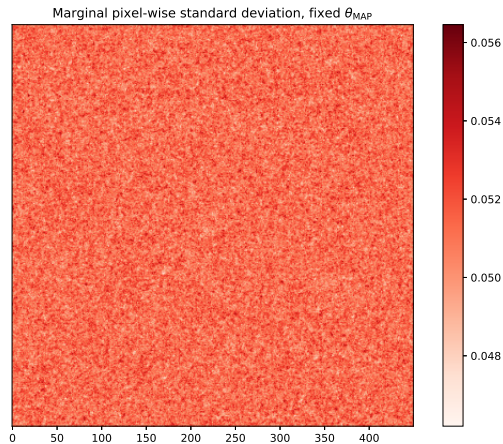


Observation

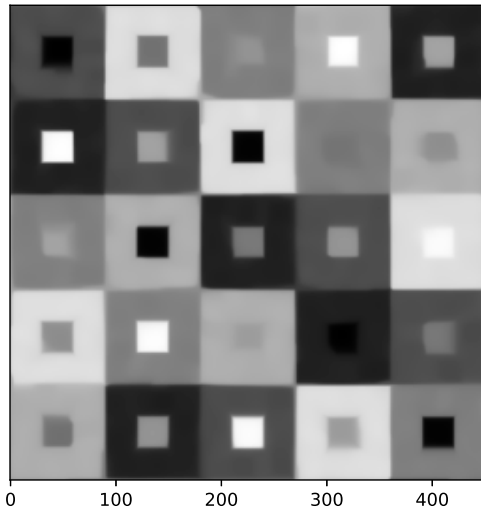


Horizontal gradient

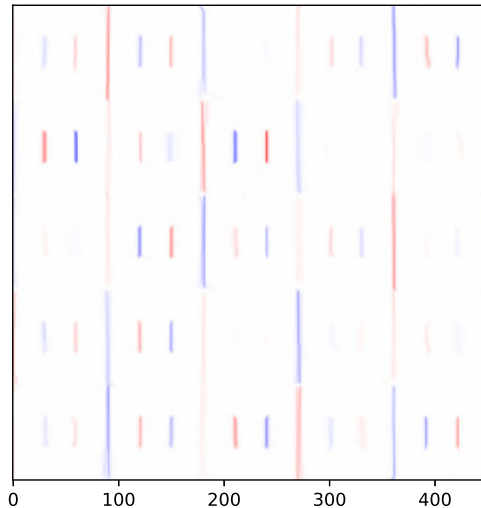


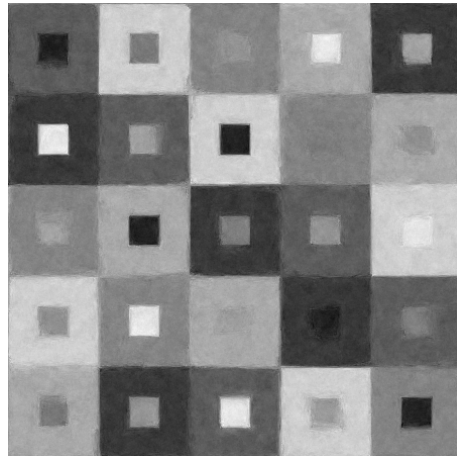
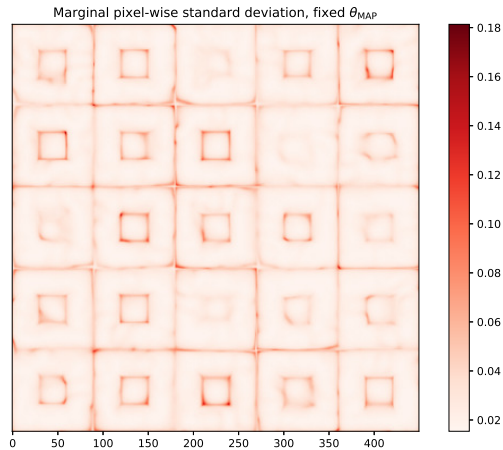


Reconstruction, many-parameter for prior



Horizontal gradient





Observation



0 200 400 600 800 1000

IAS Reconstruction



0 200 400 600 800 1000

Observation



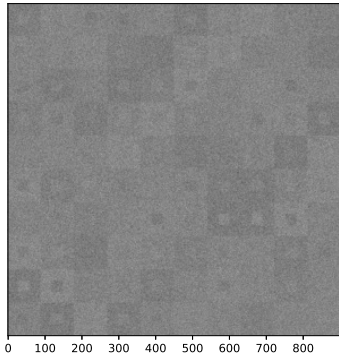
IAS Reconstruction



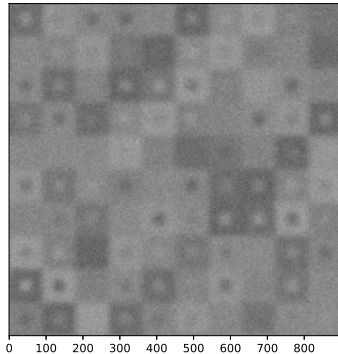
Norm of variance weights



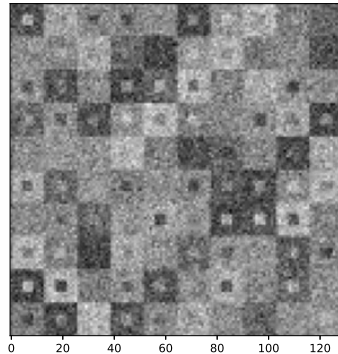
Point-wise observation (high noise)



Blurred observation (medium noise)



Up-sampled observation (low noise)



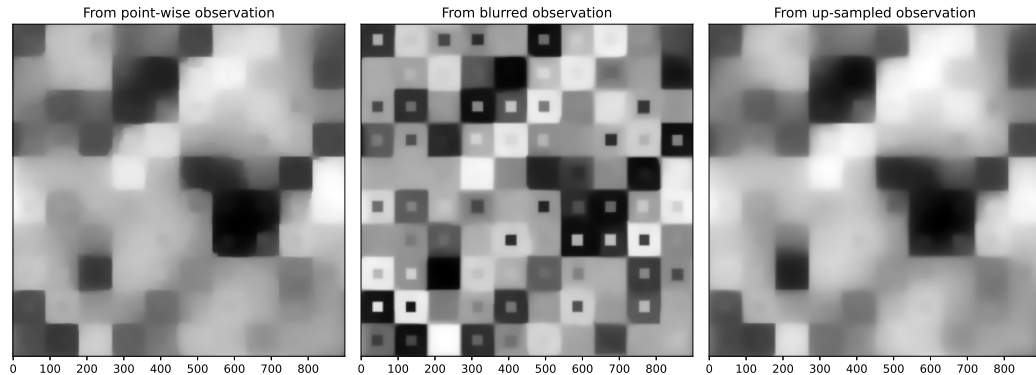


Figure: Separate reconstructions by each observation. % error in learned noise standard deviations (left to right): 0.56%, 0.25%, 48.98%.

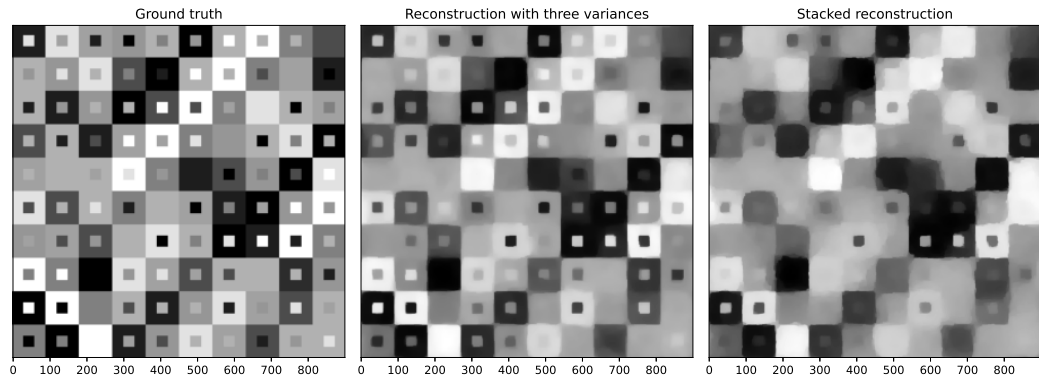
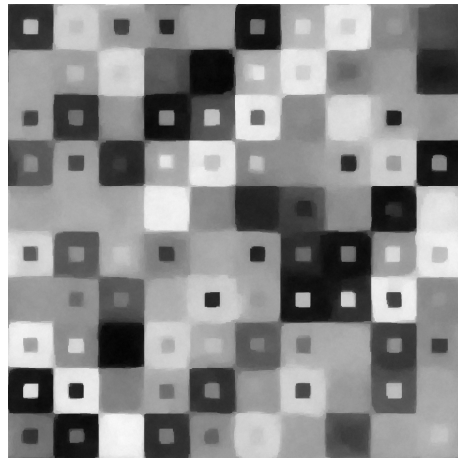
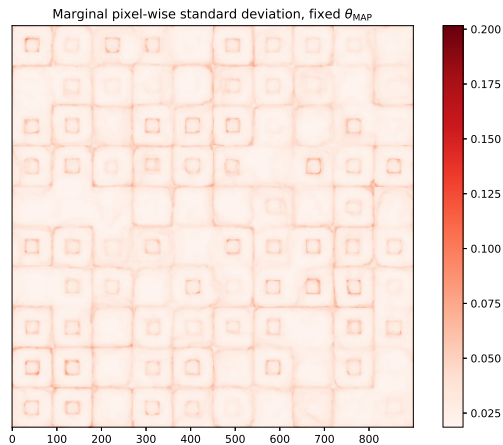


Figure: Reconstructions with three learned variance parameters vs. “stacked” reconstruction with a single learned variance parameter. % error in three learned noise standard deviations: 0.12%, 0.03%, 0.76%.

RECONSTRUCTION ACCURACY

Reconstruction	MSE	SSIM
From point-wise	0.0531	0.7111
From blurred	0.0114	0.7833
From up-sampled	0.0591	0.7039
With three learned variances	0.0063*	0.8302*
Stacked	0.0199	0.7595



CONVERGENCE

- Log-concavity of posterior will depend on the hyper-hyper parameters of the hyper-priors $(r_{1/2}, s_{1/2}, \vartheta_{1/2})$
- Previous work on IAS derived convexity conditions for fixed noise covariance and assuming $\mathbf{R}^T \mathbf{R}$ invertible
- If posterior strongly log-concave, IAS converges to unique global minimizer (block coordinate descent, convex + differentiable)
- Convergence of BCD unknown
- First contribution: extended convexity analysis to unknown noise covariance and $\mathbf{R}^T \mathbf{R}$ singular
- Second contribution: extension to the data fusion setting.

INTERPRETING CONVEXITY CONDITIONS

For the prior, sparsity is controlled by (r_2, s_2, ϑ_2) . Stronger sparsity requires parameters that yield a non-log concave prior (e.g., Student- t , Cauchy, etc.)

Unexpectedly, interpreting conditions for noise covariance much more difficult.

- “Uninformative” priors always break convexity
- For increasing m , prior must approach a Dirac delta to maintain global convexity

However, the same issue arises in joint MLE for mean and variance of Gaussian data.

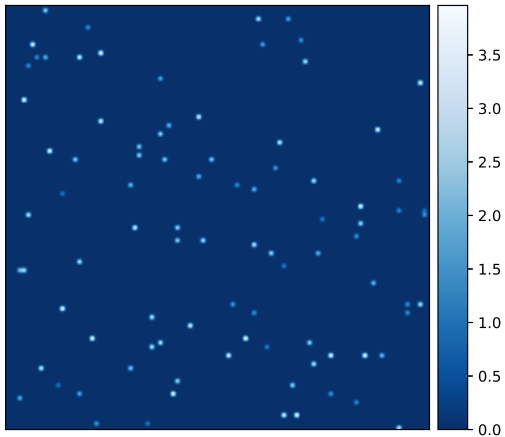
Next steps: analyze in terms of fixed point iteration, rather than convexity.

GLOBAL HYBRID IAS/BCD

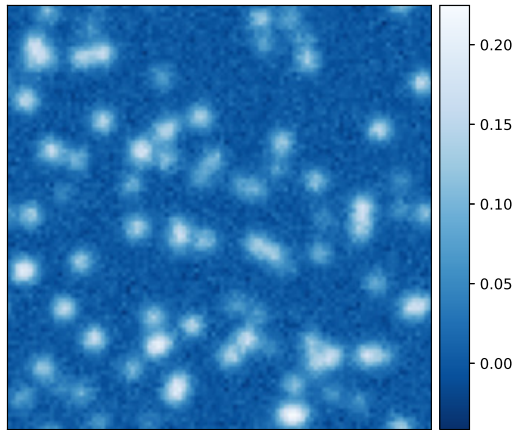
- We may like to impose a non-convex prior, but global convergence no longer guaranteed
- Idea: first solve a “close” convex problem, then use as initialization of a second solve using true prior of interest⁸
- No global convergence guarantees, but empirically a good strategy for obtaining better local minima
- Third contribution: global hybrid BCD; no conjugacy, yet conditional mean still known analytically

⁸Calvetti, Pragliola and Somersalo, “Sparsity Promoting Hybrid Solvers for Hierarchical Bayesian Inverse Problems”, 2020.

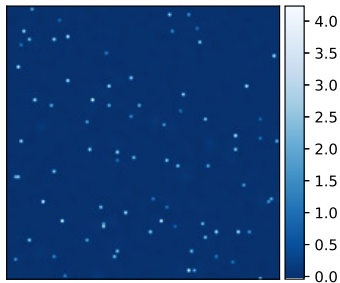
Ground truth



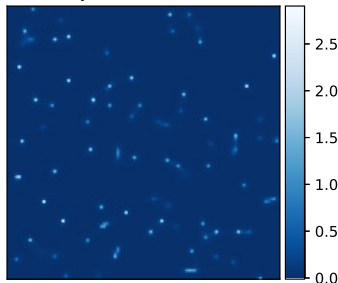
Observation



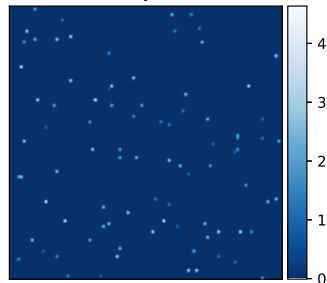
Vanilla IAS

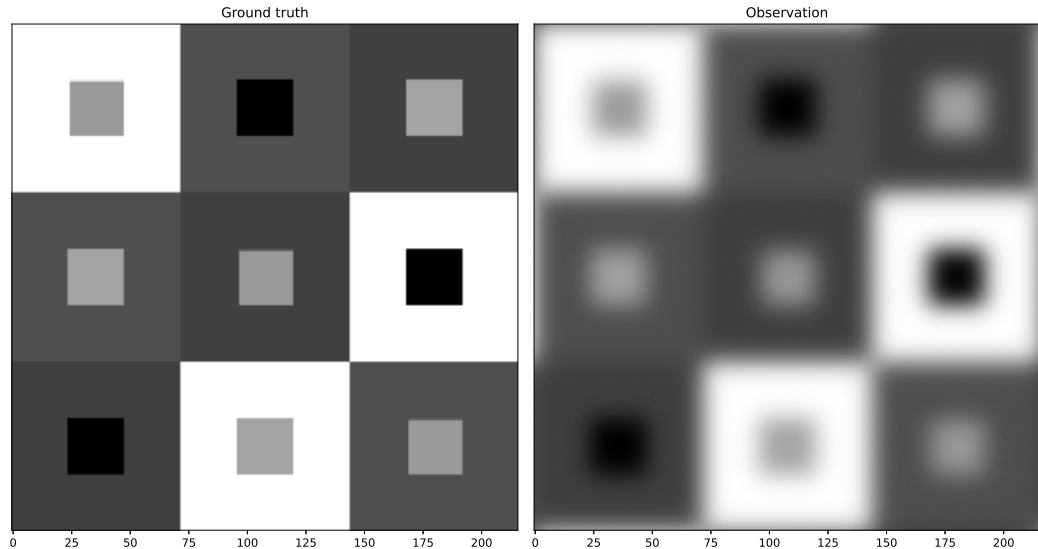


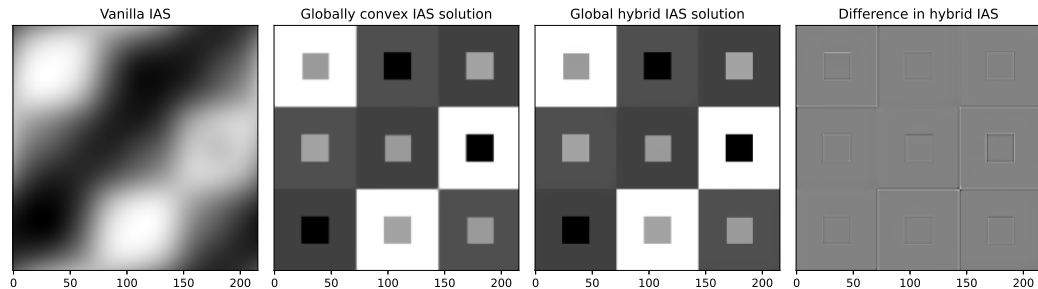
Globally convex IAS solution



Global hybrid IAS







PCG-IAS/BCD

Fourth contribution:

- When using a Gaussian prior, we are able to make use of a common diagonalization of $\mathbf{F}^T \mathbf{F}$ and $\mathbf{R}^T \mathbf{R}$ to directly compute an $\mathbf{x}$ -update, e.g., if both block-circulant with circulant blocks (BCCB) then

$$\mathbf{F}^T \mathbf{F} = \mathcal{F}^H \mathbf{\Lambda}_F \mathcal{F}, \quad \mathbf{R}^T \mathbf{R} = \mathcal{F}^H \mathbf{\Lambda}_R \mathcal{F}$$

⁹Marnissi et al., "A Data Augmentation Approach for Sampling Gaussian Models in High Dimension", 2019.

PCG-IAS/BCD

Fourth contribution:

- When using a Gaussian prior, we are able to make use of a common diagonalization of $\mathbf{F}^T \mathbf{F}$ and $\mathbf{R}^T \mathbf{R}$ to directly compute an $\mathbf{x}$ -update, e.g., if both block-circulant with circulant blocks (BCCB) then

$$\mathbf{F}^T \mathbf{F} = \mathcal{F}^H \mathbf{\Lambda}_F \mathcal{F}, \quad \mathbf{R}^T \mathbf{R} = \mathcal{F}^H \mathbf{\Lambda}_R \mathcal{F}$$

- When doing IAS/BCD with a sparse prior, $\mathbf{D}_\beta^{-1}$ “gets-in-the-way”, since

$$\mathbf{R}^T \mathbf{D}_\beta^{-1} \mathbf{R} = \textcolor{red}{?}$$

⁹Marnissi et al., “A Data Augmentation Approach for Sampling Gaussian Models in High Dimension”, 2019.

PCG-IAS/BCD

Fourth contribution:

- When using a Gaussian prior, we are able to make use of a common diagonalization of $\mathbf{F}^T \mathbf{F}$ and $\mathbf{R}^T \mathbf{R}$ to directly compute an $\mathbf{x}$ -update, e.g., if both block-circulant with circulant blocks (BCCB) then

$$\mathbf{F}^T \mathbf{F} = \mathcal{F}^H \mathbf{\Lambda}_F \mathcal{F}, \quad \mathbf{R}^T \mathbf{R} = \mathcal{F}^H \mathbf{\Lambda}_R \mathcal{F}$$

- When doing IAS/BCD with a sparse prior, $\mathbf{D}_\beta^{-1}$ “gets-in-the-way”, since

$$\mathbf{R}^T \mathbf{D}_\beta^{-1} \mathbf{R} = \textcolor{red}{?}$$

- We have made use of recent data augmentation techniques⁹ to get around this

⁹Marnissi et al., “A Data Augmentation Approach for Sampling Gaussian Models in High Dimension”, 2019.

$$v_2 | x, \beta \sim \mathcal{N}(\mu_2, Q_2^{-1}), \quad Q_2^{-1} = \frac{1}{\lambda_2(\beta)} I - D_\beta^{-1}, \quad \mu_2 = Q^{-1} R x,$$

$$\Rightarrow x | v, \alpha, \beta, y \sim \mathcal{N}(\mu, Q^{-1}), \quad Q = \frac{1}{\alpha} F^T F + \frac{1}{\lambda_2} R^T R, \quad \mu = Q^{-1} \left(\frac{1}{\alpha} F^T y + R^T v_2 \right)$$

PCG-IAS Algorithm

Choose initial $x^0, \alpha^0, \beta^0, v^0$.

for $k = 1, \dots, n_{\text{maxits}}$ do

$$\alpha^k \leftarrow \arg \max_{\alpha} \pi(\alpha | x^{k-1}, \beta^{k-1}, y).$$

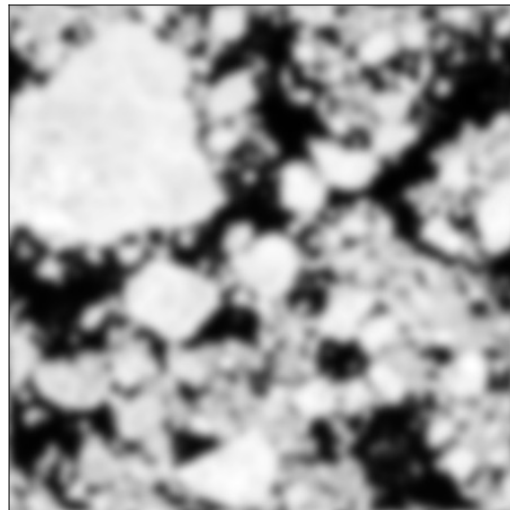
$$\beta^k \leftarrow \arg \max_{\beta} \pi(\beta | \alpha^k, x^{k-1}, y).$$

$$v^k \leftarrow \arg \max_v \pi(v | \alpha^k, \beta^k, x^{k-1}, y)$$

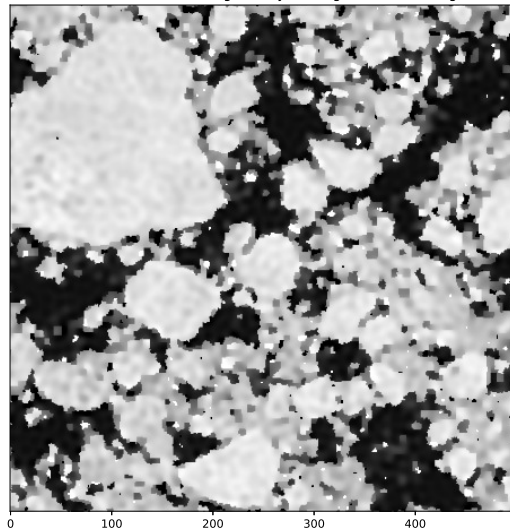
$$x^k \leftarrow \arg \max_x \pi(x | \alpha^k, \beta^k, v^k, y).$$

end for

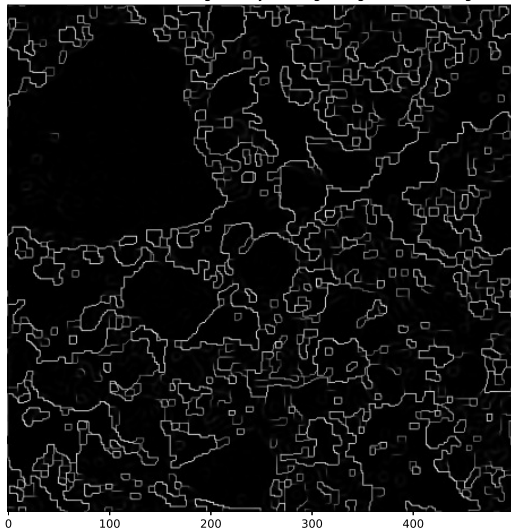
Observation



Non-convex solution with global hybrid, log of variance weights



Non-convex solution with global hybrid, log of regularization weights



Non-convex solution without global hybrid, log of regularization weights

