



GENERALIZED HYBRID SOLVERS

for sparsity-promoting Bayesian inverse problems

Advancement exam talk

17th November 2022



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

OUTLINE

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

BAYESIAN INVERSE PROBLEMS

Goal: we would like to reconstruct an unknown signal $\mathbf{x} \in \mathbb{R}^n$ given an indirect noisy observation

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

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 $\mathbf{Q}_{\text{noise}} \in \mathbb{R}^{m \times m}$.

Goal: we would like to reconstruct an unknown signal $\mathbf{x} \in \mathbb{R}^n$ given an indirect noisy observation

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

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

LINEAR FORWARD OPERATORS

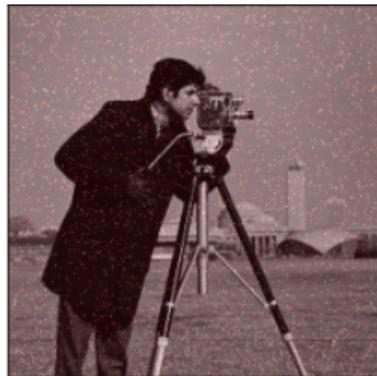
Ground truth



Blur



Under-sampling



Up-sampling



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

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

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

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

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

LEAST-SQUARES

Observation (no noise)



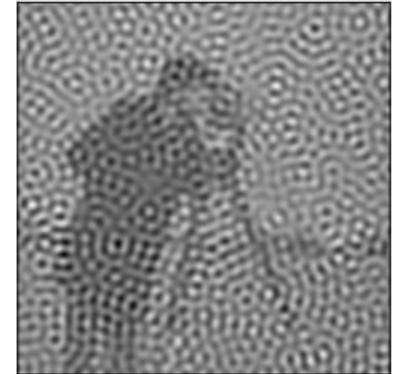
Least-squares reconstruction



Observation (with noise)



Least-squares reconstruction



Fundamental idea: introduce a prior distribution with density $\pi(\mathbf{x})$ to supplement the data, and apply Bayes' theorem.

Bayes' theorem

$$\overbrace{\pi(\mathbf{x} | \mathbf{y})}^{\text{posterior}} = \frac{\pi(\mathbf{y} | \mathbf{x})\pi(\mathbf{x})}{\int \pi(\mathbf{y} | \mathbf{x})\pi(\mathbf{x}) d\mathbf{x}}$$
$$\propto \underbrace{\pi(\mathbf{y} | \mathbf{x})}_{\text{likelihood}} \underbrace{\pi(\mathbf{x})}_{\text{prior}}$$

The “solution” is the posterior measure with density π_{post} , and the ability to make queries w.r.t. the posterior, e.g., estimate expectations $\mathbb{E}_{\pi_{\text{post}}} [f(\mathbf{x})]$ or seek point estimates such as the mode.

Maximum a posteriori (MAP) estimate

$$\begin{aligned}\mathbf{x}_{\text{MAP}} &= \arg \max_{\mathbf{x}} \{\pi(\mathbf{y} | \mathbf{x})\pi(\mathbf{x})\} \\ &= \arg \min_{\mathbf{x}} \{-\log \pi(\mathbf{y} | \mathbf{x}) - \log \pi(\mathbf{x})\}\end{aligned}$$

HOW TO PICK A PRIOR?

A common choice is to pick

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

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

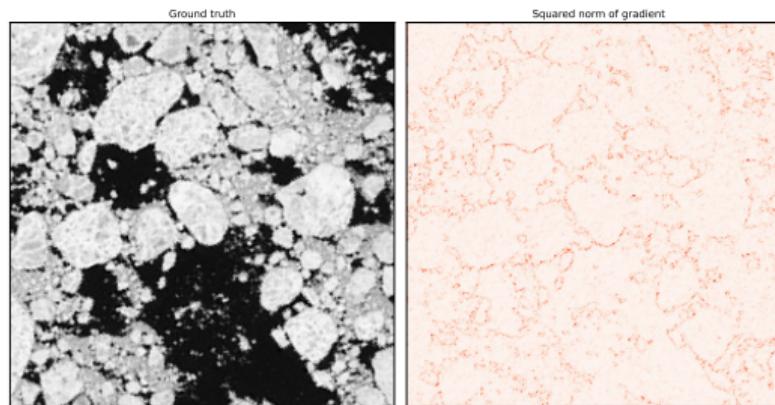


Figure: Discrete gradient operator.

GAUSSIAN PRIOR + GAUSSIAN LIKELIHOOD

Suppose that

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

GAUSSIAN PRIOR + GAUSSIAN LIKELIHOOD

Suppose that

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

Then

$$\mathbf{x} | \mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\text{post}}, \mathbf{Q}_{\text{post}}^{-1}\right)$$

with

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

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

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

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

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

Gaussian posterior

The posterior is the Gaussian $\mathcal{N}(\boldsymbol{\mu}_{\text{post}}, \boldsymbol{Q}_{\text{post}}^{-1})$ with

$$\boldsymbol{Q}_{\text{post}} = \sum_{j=1}^K \boldsymbol{F}_j^T \boldsymbol{Q}_j \boldsymbol{F}_j,$$
$$\boldsymbol{\mu}_{\text{post}} = \boldsymbol{Q}_{\text{post}}^{-1} \left(\sum_{j=1}^K \boldsymbol{F}_j^T \boldsymbol{Q}_j \boldsymbol{y}_j \right).$$

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

TIKHONOV DE-BLURRING

Ground truth



Observation

 $\lambda = 1$  $\lambda = 100$  $\lambda = 1000$ 

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

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 PRIOR SAMPLES

GAUSSIAN POSTERIOR SAMPLES

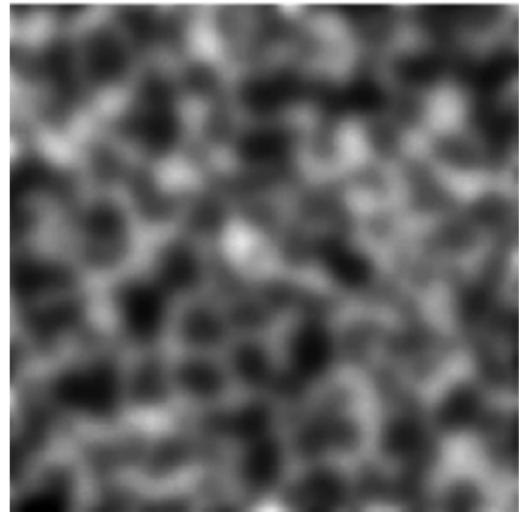


Figure: Gaussian prior sample.

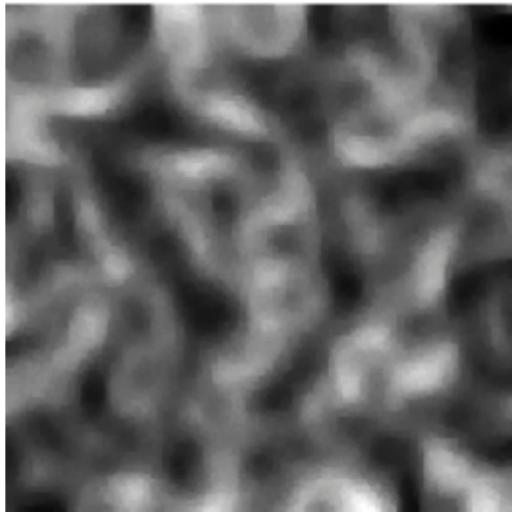


Figure: Total variation (TV) prior sample.

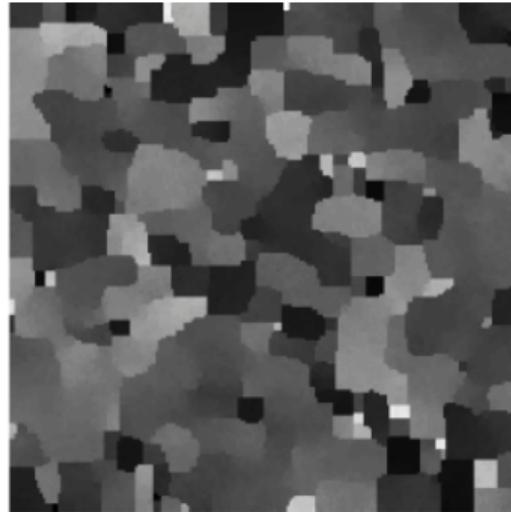


Figure: Cauchy difference prior sample.

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. 

NON-GAUSSIAN PRIORS

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

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

NON-GAUSSIAN PRIORS

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

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

- 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 VS. TV (MAP)

Gaussian prior



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

SCALE MIXTURES

An interesting fact:

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

SCALE MIXTURES

An interesting fact:

$$w \sim \text{Exponential} \left(2^{-1} b^{-2} \right),$$
$$x | w \sim \mathcal{N}(0, w),$$

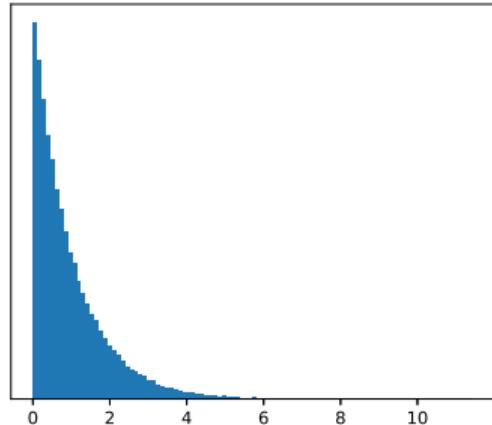
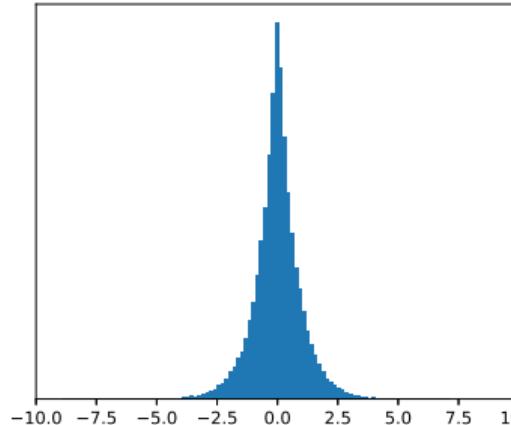
SCALE MIXTURES

An interesting fact:

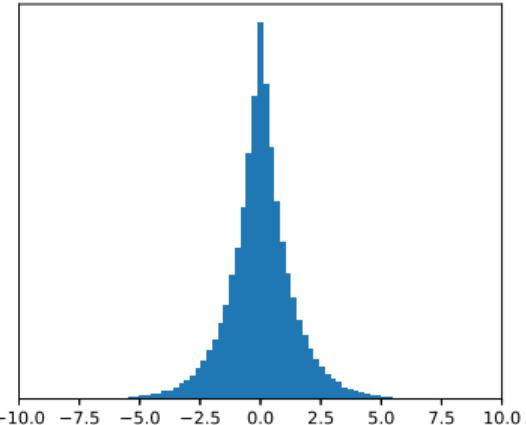
$$\begin{aligned} w &\sim \text{Exponential}\left(2^{-1}b^{-2}\right), \\ x \mid w &\sim \mathcal{N}(0, w), \\ \Rightarrow x &\sim \text{Laplace}(0, b). \end{aligned}$$

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

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

w samples $x | w$ samples

Laplace samples



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

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

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

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

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

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

- Encompasses exponential, Gamma, inverse Gamma distributions

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

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

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

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

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

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

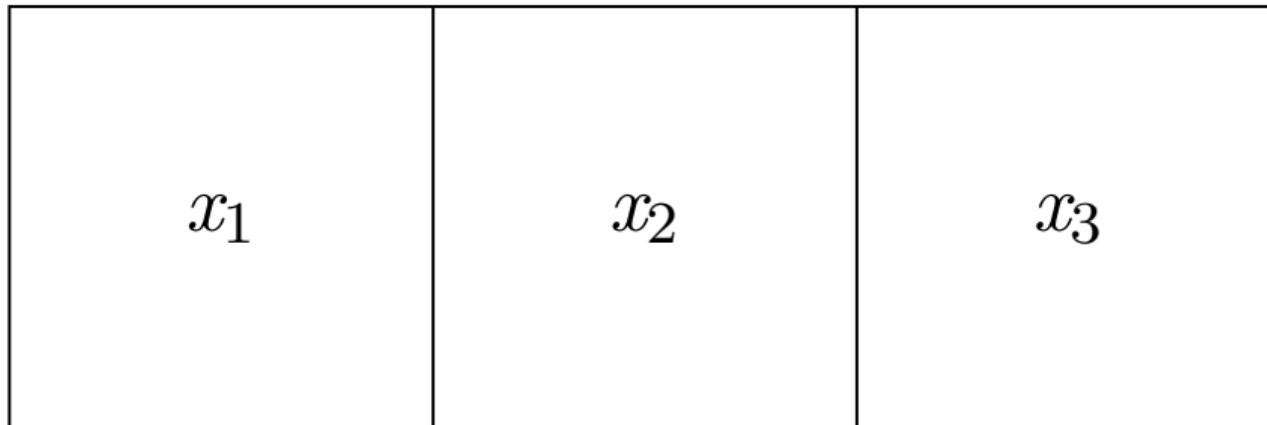
$$\pi_{\text{GG}}(\theta | r, s, \vartheta) = \frac{|r|}{\vartheta \Gamma(s)} \left(\frac{\theta}{\vartheta} \right)^{rs-1} \exp \left\{ - \left(\frac{\theta}{\vartheta} \right)^r \right\} \mathbb{1}_{>0}(\theta)$$

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

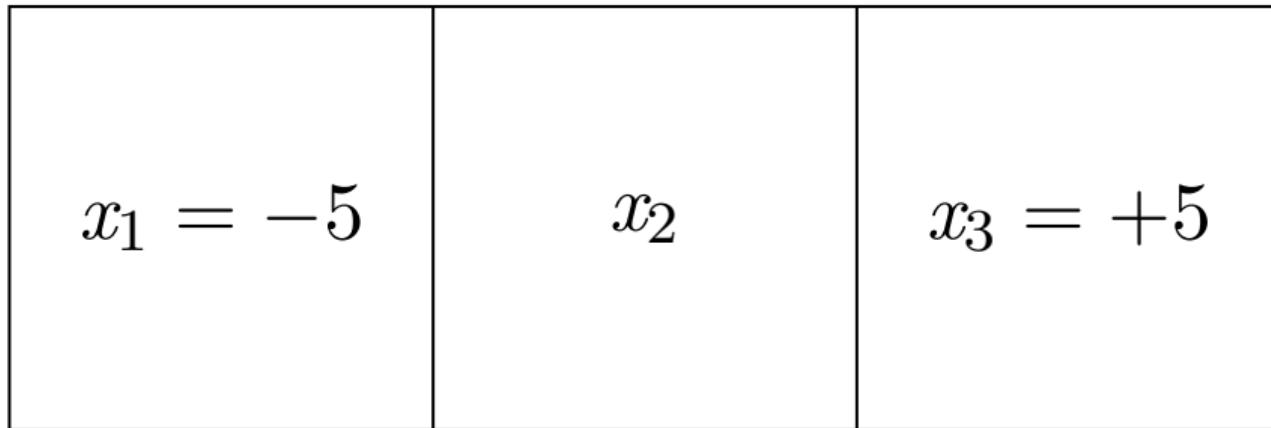
- Encompasses exponential, Gamma, inverse Gamma distributions
- Parameters (r, s) set the parameteric family of marginal prior, ϑ controls scale
- In certain limits, captures ℓ_p -norm priors for $0 < p < 2$, Student- t , Cauchy
- Detailed convexity analysis provided by [5]

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



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

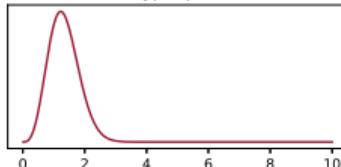


What does the (univariate) conditional

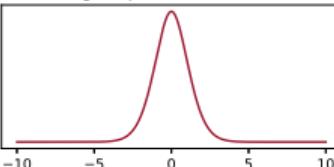
$$\pi(x_2 \mid x_1 = -5, x_3 = 5)$$

look like?

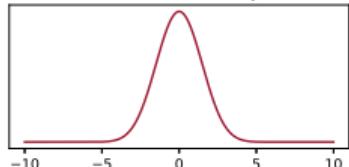
Hyper-prior



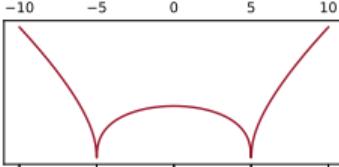
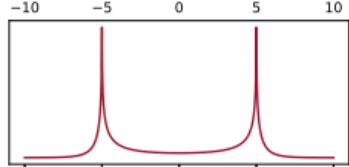
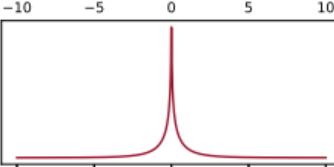
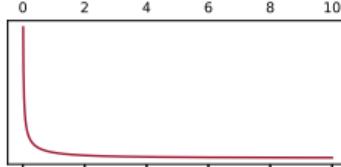
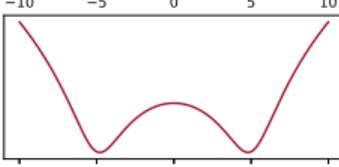
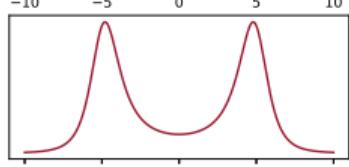
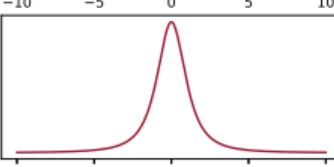
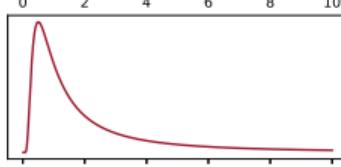
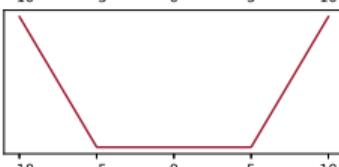
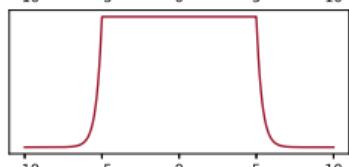
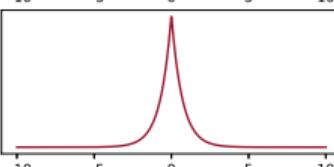
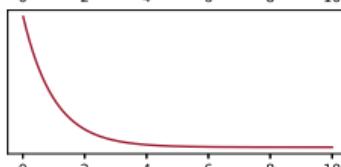
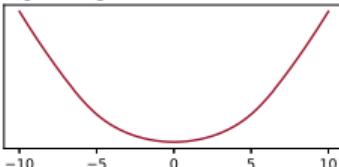
Marginal prior on differences



Conditional density



Negative logarithm of conditional density



THE HIERARCHICAL MODEL

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\beta^{-1} \boldsymbol{R}^T \boldsymbol{R}\right)^{-1}\right) \\ \boldsymbol{y} | \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{F}\boldsymbol{x}, \alpha \boldsymbol{I}) \end{aligned} \quad \text{vs.} \quad \begin{aligned} \alpha &\sim \text{GG}(r_1, s_1, \vartheta_1) \\ \beta_j &\sim \text{GG}(r_2, s_2, \vartheta_2) \quad j = 1, \dots, k \\ \boldsymbol{x} | \boldsymbol{\beta} &\sim \mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{R}^T \boldsymbol{D}_{\boldsymbol{\beta}}^{-1} \boldsymbol{R}\right)^{-1}\right) \\ \boldsymbol{y} | \boldsymbol{x}, \alpha &\sim \mathcal{N}(\boldsymbol{F}\boldsymbol{x}, \alpha \boldsymbol{I}) \end{aligned}$$

THE HIERARCHICAL MODEL

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\beta^{-1} \boldsymbol{R}^T \boldsymbol{R}\right)^{-1}\right) & \text{vs.} & \quad \alpha \sim \text{GG}(r_1, s_1, \vartheta_1) \\ \boldsymbol{y} | \boldsymbol{x} &\sim \mathcal{N}(\boldsymbol{F}\boldsymbol{x}, \alpha \boldsymbol{I}) & & \beta_j \sim \text{GG}(r_2, s_2, \vartheta_2) \quad j = 1, \dots, k \\ & & & \boldsymbol{x} | \boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \left(\boldsymbol{R}^T \boldsymbol{D}_{\boldsymbol{\beta}}^{-1} \boldsymbol{R}\right)^{-1}\right) \\ & & & \boldsymbol{y} | \boldsymbol{x}, \alpha \sim \mathcal{N}(\boldsymbol{F}\boldsymbol{x}, \alpha \boldsymbol{I}) \end{aligned}$$

$$\pi(\boldsymbol{x} | \boldsymbol{y}) \propto \pi(\boldsymbol{y} | \boldsymbol{x})\pi(\boldsymbol{x}) \quad \text{vs.} \quad \pi(\boldsymbol{x}, \alpha, \boldsymbol{\beta} | \boldsymbol{y}) \propto \pi(\boldsymbol{y} | \boldsymbol{x}, \alpha)\pi(\boldsymbol{x} | \boldsymbol{\beta})\pi(\alpha)\pi(\boldsymbol{\beta})$$

NEGATIVE LOG POSTERIOR

In the simpler Gaussian prior/likelihood case:

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

In our hierarchical model:

$$E(\boldsymbol{x}, \alpha, \boldsymbol{\beta}) = E_1(\boldsymbol{x}, \alpha) + E_2(\boldsymbol{x}, \boldsymbol{\beta}),$$

$$E_1(\boldsymbol{x}, \alpha) = \frac{1}{2\alpha} \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_2^2 - \left(r_1 s_1 - \frac{m+2}{2} \right) \log \alpha + \left(\frac{\alpha}{\vartheta_1} \right)^{r_1},$$

$$E_2(\boldsymbol{x}, \boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{R}\boldsymbol{x}\|_{\boldsymbol{D}_{\boldsymbol{\beta}}^{-1}}^2 - \left(r_2 s_2 - \frac{3}{2} \right) \sum_{j=1}^k \log \beta_j + \sum_{j=1}^k \left(\frac{\beta_j}{\vartheta_2} \right)^{r_2}$$

SOLVERS

Hierarchical Solvers for Inverse Problems

User Guide

[Overview](#)

[Installation](#)

[Mathematical Background](#)

[Multivariate Gaussians](#)

[Bayesian inference](#)

[Gaussian Factors and Posteriors](#)

[Examples](#)

 1. [Working with `hsip.linops`](#)

 2. [Solving simple inverse problems](#)

 3. [Basic hierarchical examples](#)

 4. [Using CuPy](#)

 5. [Global hybrid IAS](#)

 6. [MURI Annual Review 2022 - Software Demo](#)

[Projects](#)

 1. [Generalized hybrid solvers](#)

≡ On this page

[User Guide](#)

[Developer Guide](#)

>Show Source

Developer Guide

[API Reference](#)

[hsip.data](#)

[hsip.linops](#)

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, "Hierachical 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 $\boldsymbol{x}^0, \alpha^0, \beta^0$.

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

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

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

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

end for

BAYESIAN COORDINATE DESCENT (BCD) ALGORITHM

BCD Algorithm

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

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

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

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

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

end for

UPDATES

Each \boldsymbol{x} -update is the solution to

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

UPDATES

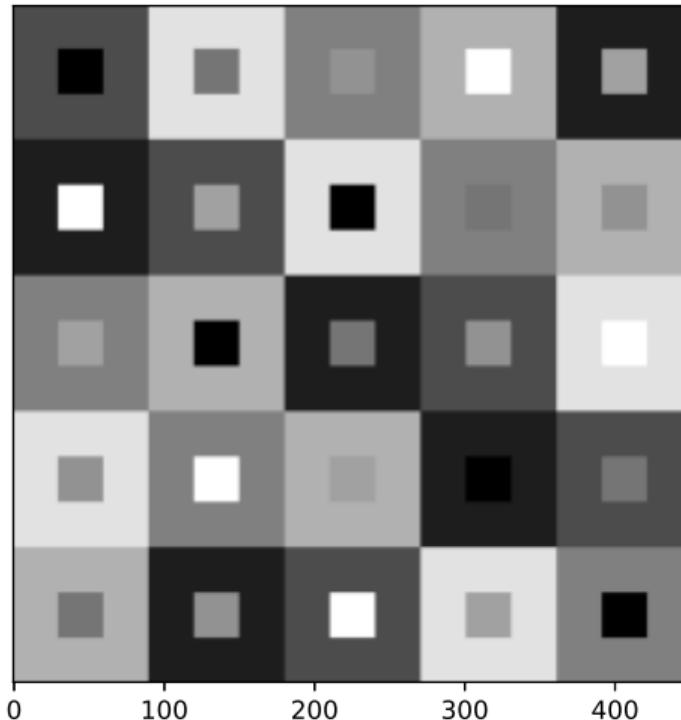
Each \boldsymbol{x} -update is the solution to

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

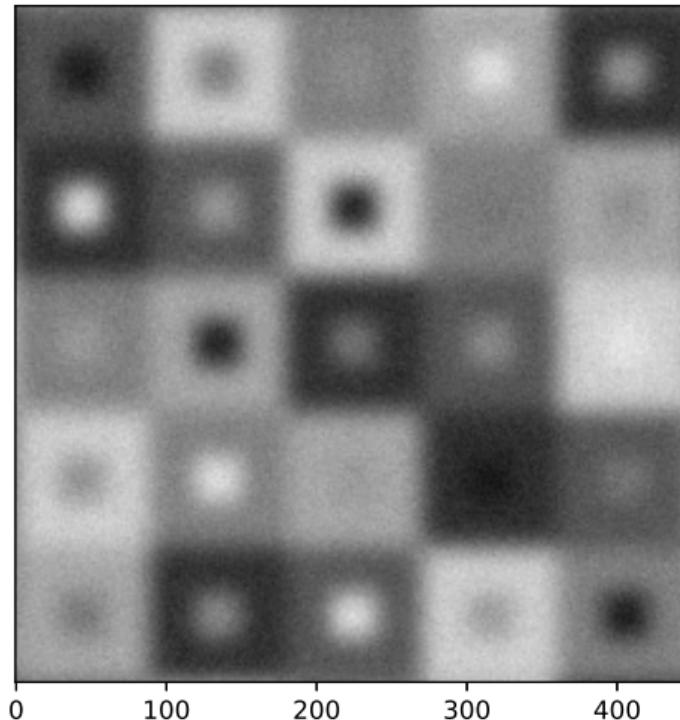
Updating α, β ?

- 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

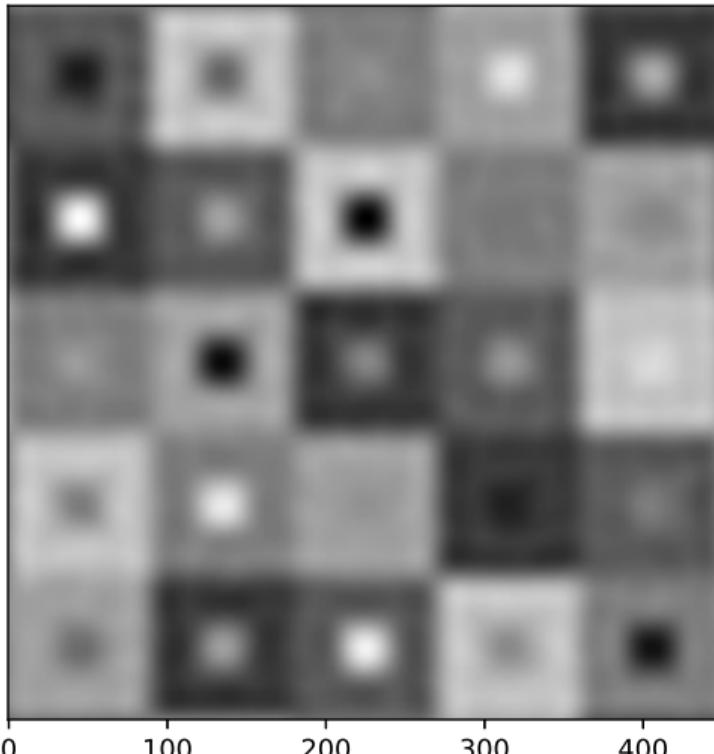
Ground truth



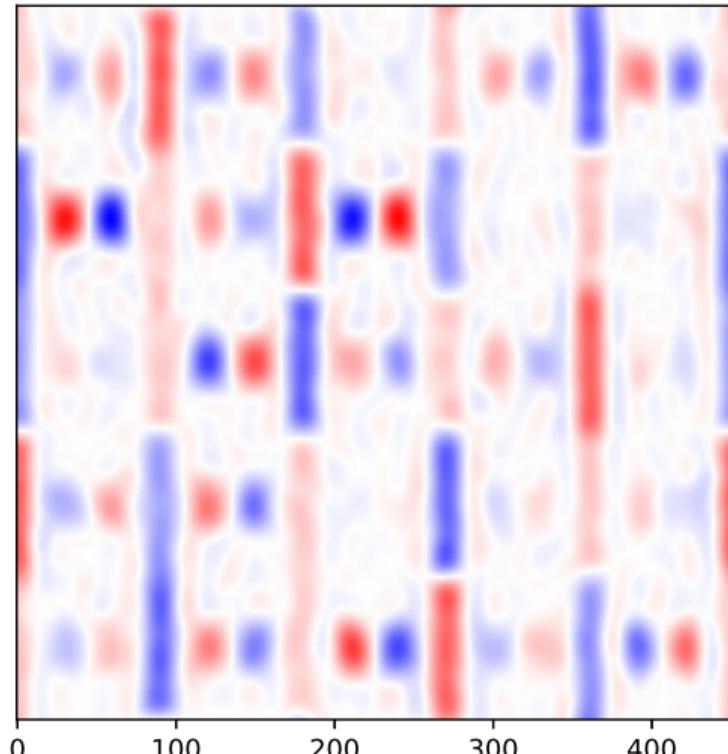
Observation

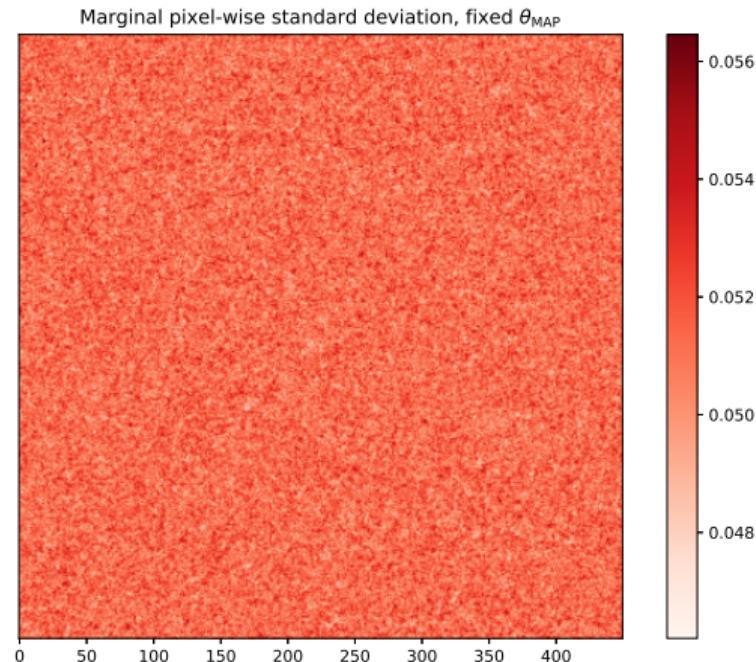


Reconstruction, single-parameter for prior

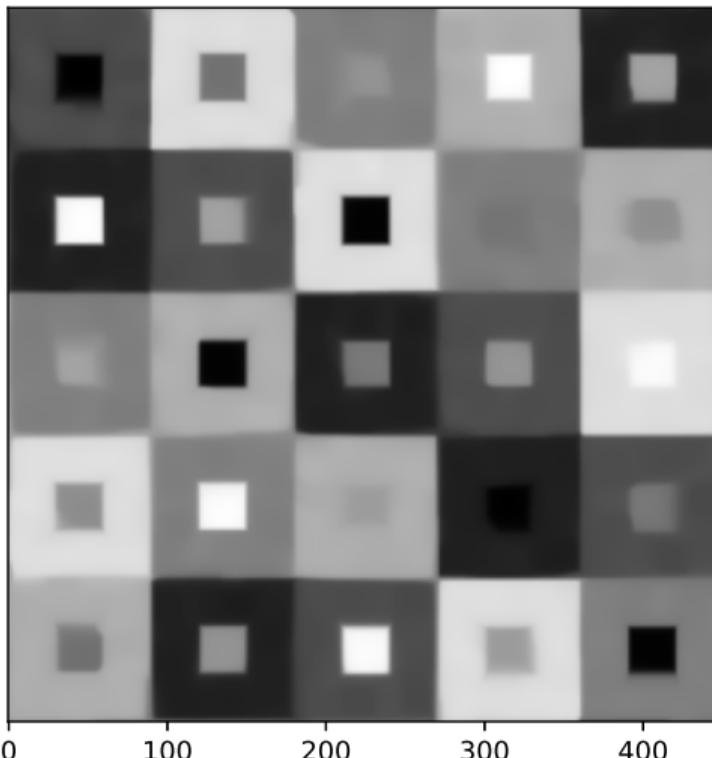


Horizontal gradient

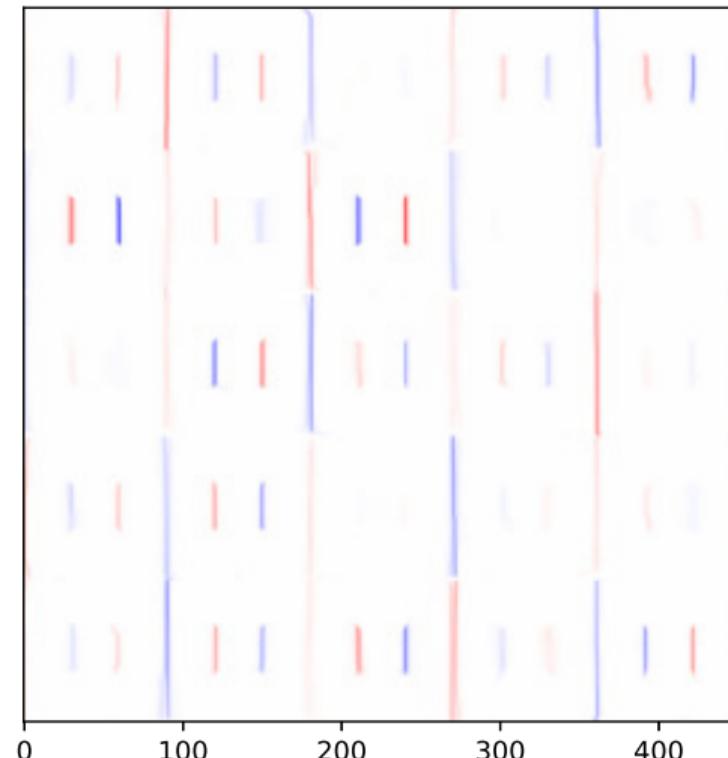


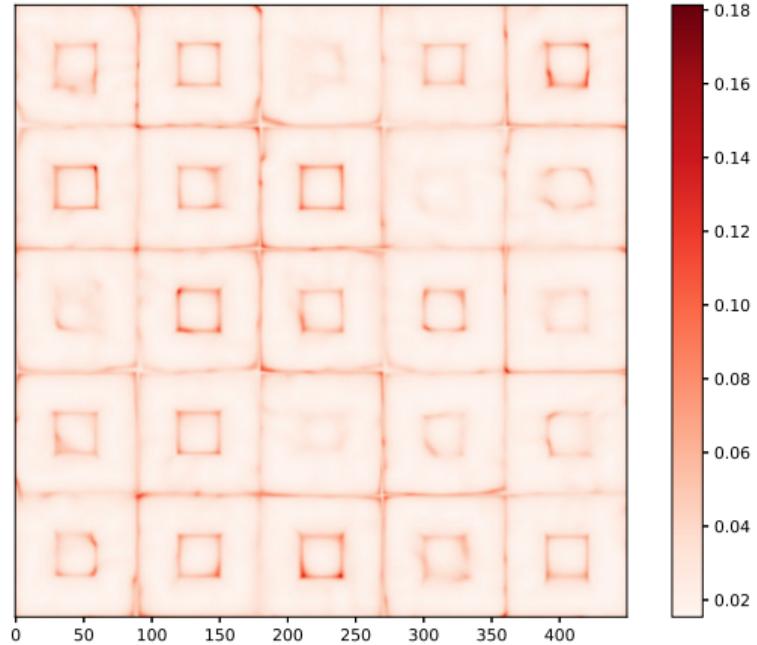


Reconstruction, many-parameter for prior



Horizontal gradient



Marginal pixel-wise standard deviation, fixed θ_{MAP} 

Observation



Observation



IAS Reconstruction



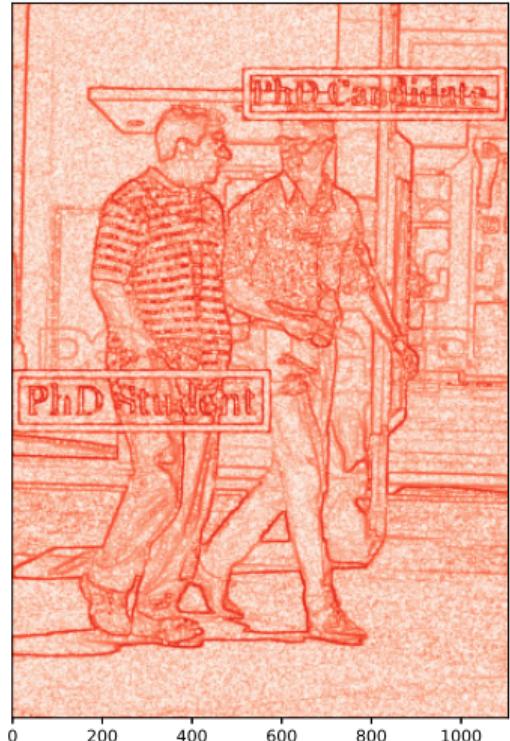
Observation



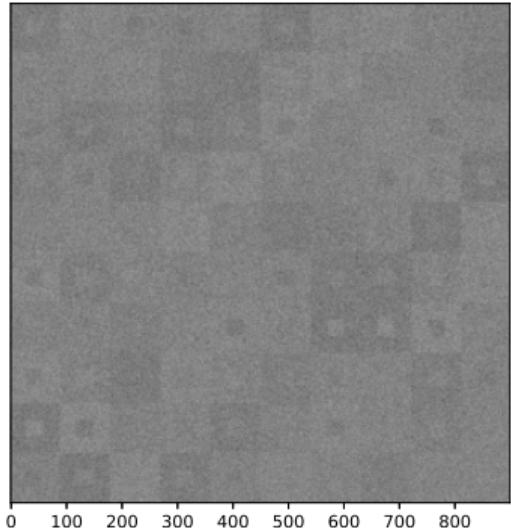
IAS Reconstruction



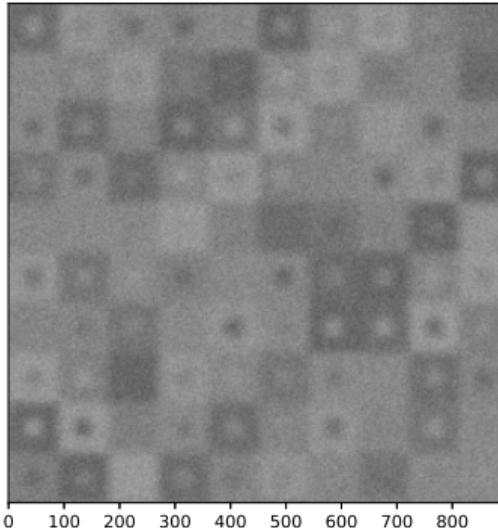
Norm of variance weights



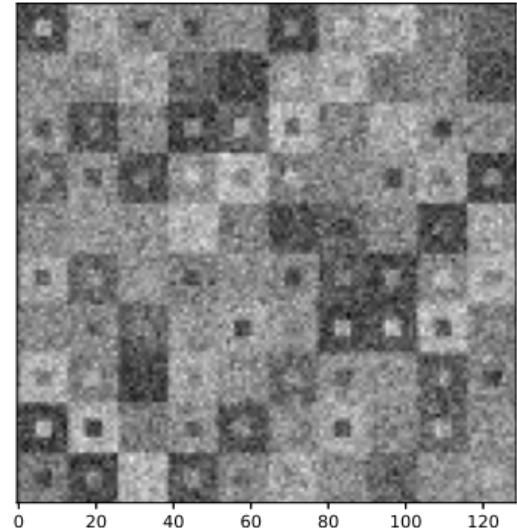
Point-wise observation (high noise)



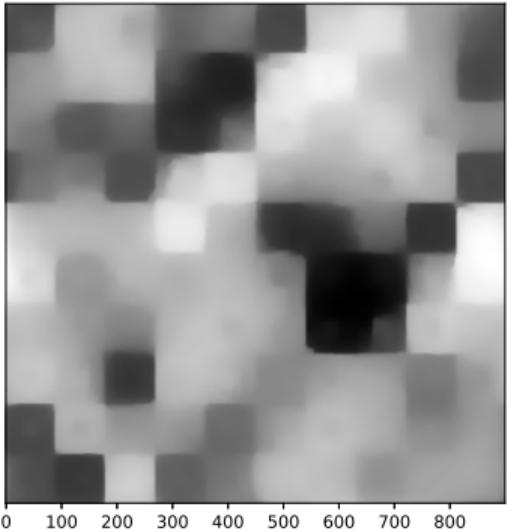
Blurred observation (medium noise)



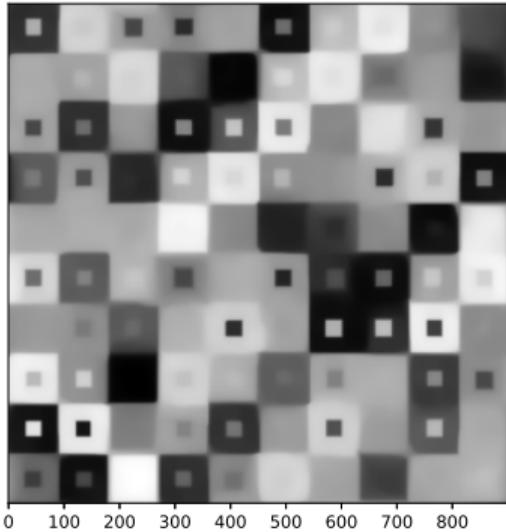
Up-sampled observation (low noise)



From point-wise observation



From blurred observation



From up-sampled observation

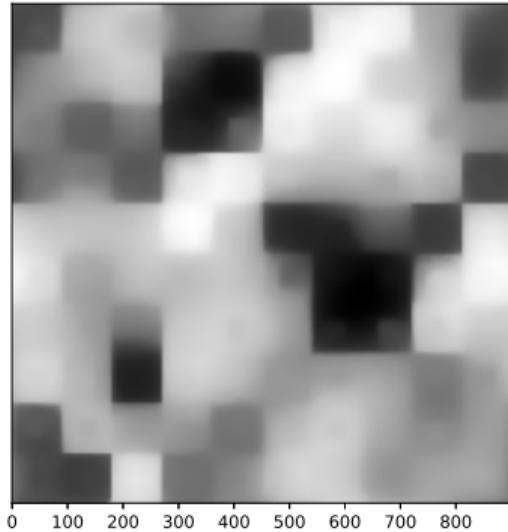


Figure: Separate reconstructions by each observation. % error in learned noise standard deviations (left to right): 0.56%, 0.25%, 48.98%.

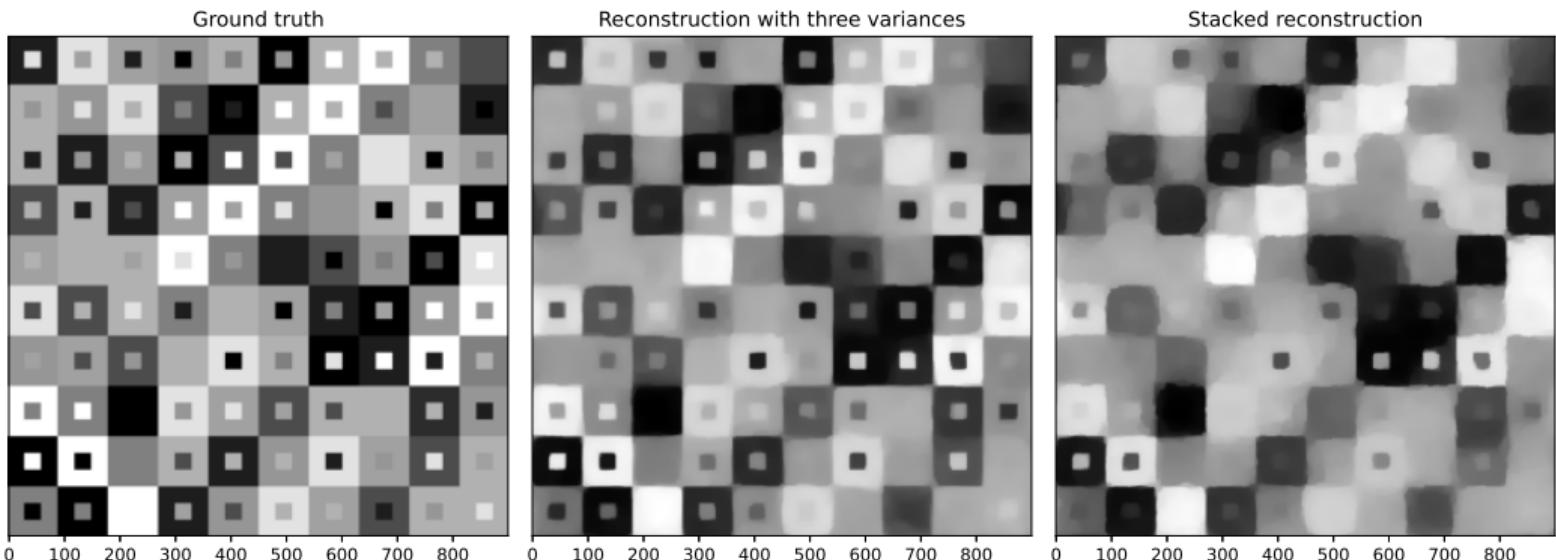
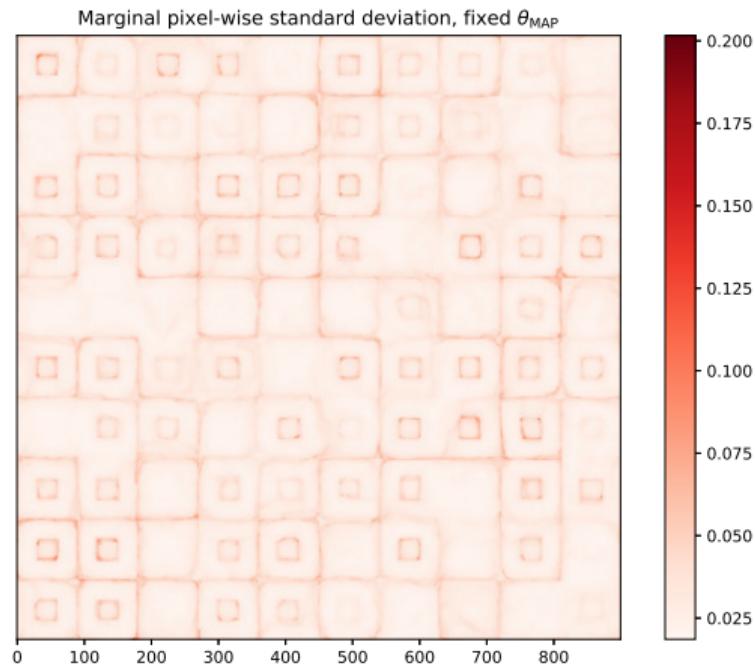


Figure: Reconstructions with three learned variance parameters vs. “stacked” reconstruction with a single learned variance parameter. % error in three learned noise standard deviations: 0.12%, 0.03%, 0.76%.

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

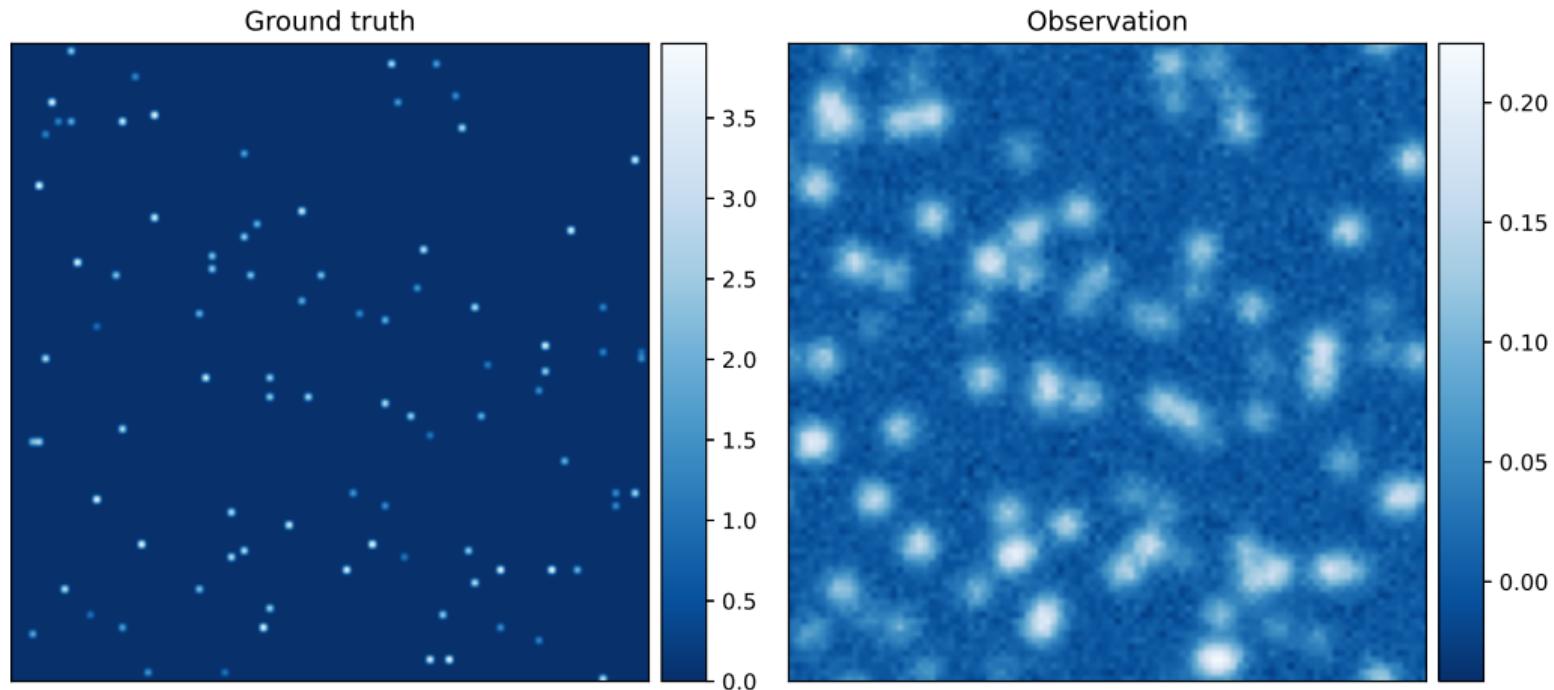
However, the same issue arises in joint MLE for mean and variance of Gaussian data.

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

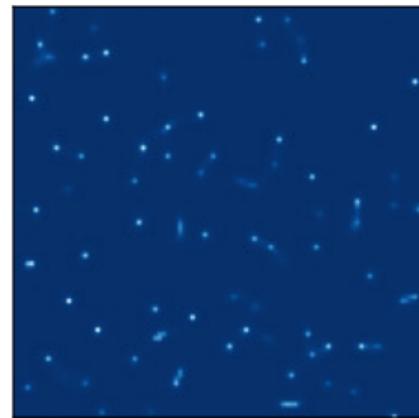
⁸Calvetti, Pragliola and Somersalo, “Sparsity Promoting Hybrid Solvers for Hierarchical Bayesian Inverse Problems”, 2020.



Vanilla IAS



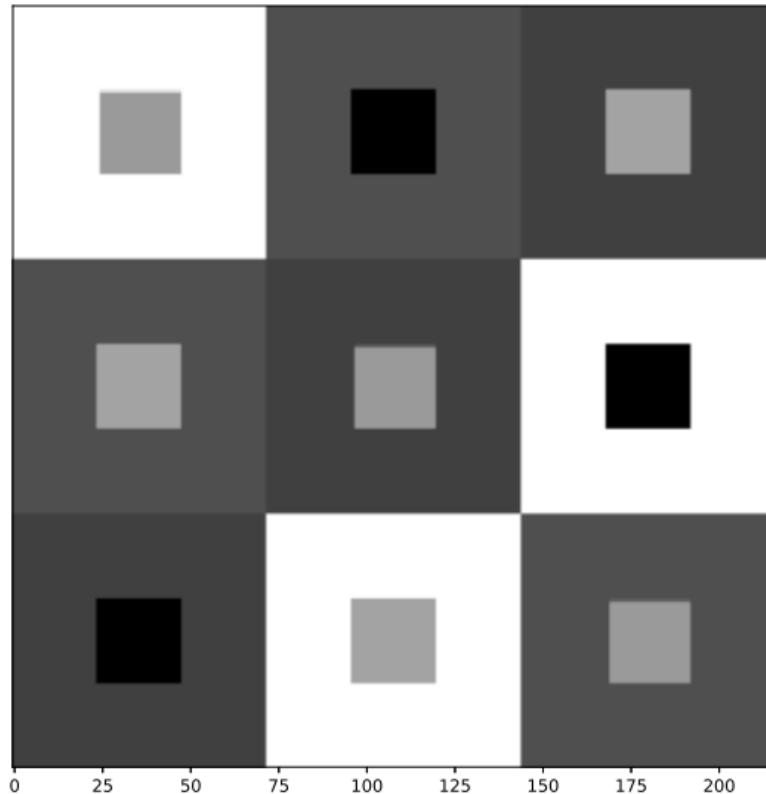
Globally convex IAS solution



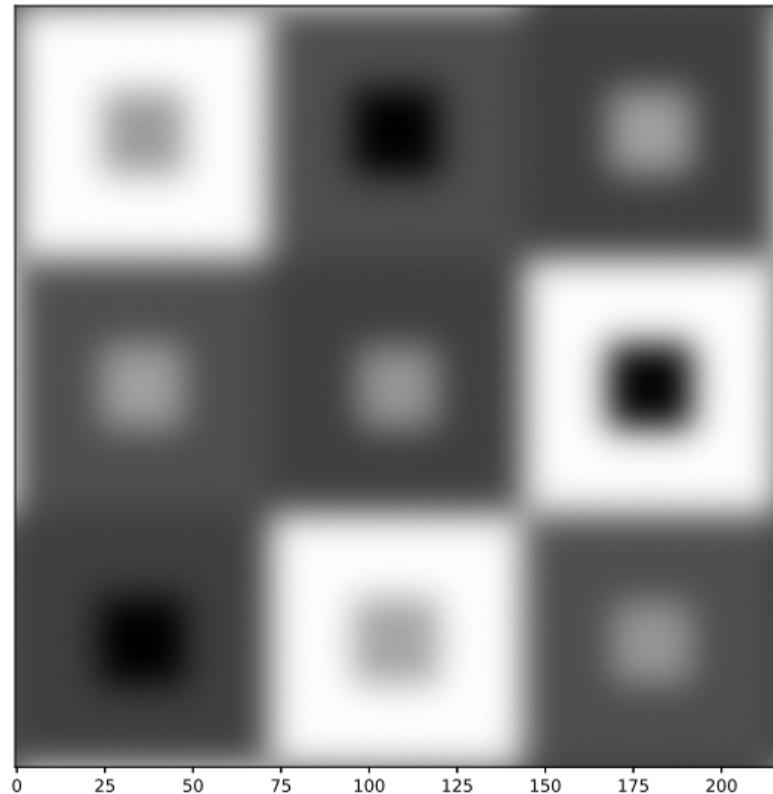
Global hybrid IAS



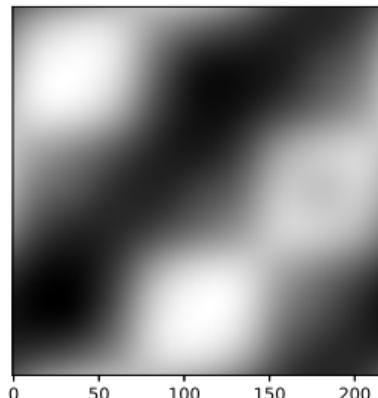
Ground truth



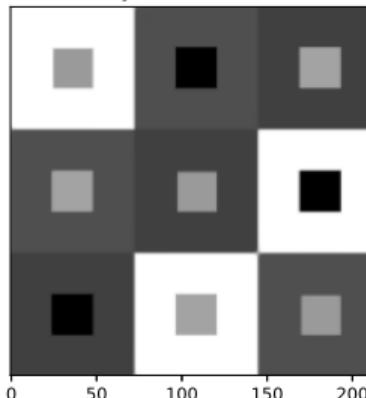
Observation



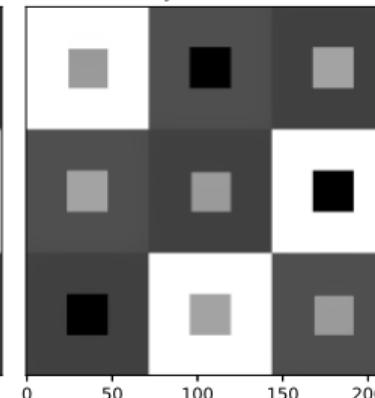
Vanilla IAS



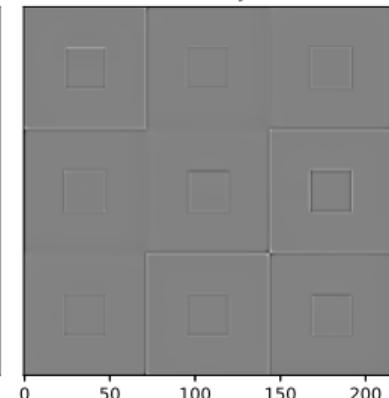
Globally convex IAS solution



Global hybrid IAS solution



Difference in 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

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

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

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

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

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

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

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

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

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

- 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.

$$\boldsymbol{v}_2 | \boldsymbol{x}, \boldsymbol{\beta} \sim \mathcal{N} \left(\boldsymbol{\mu}_2, \boldsymbol{Q}_2^{-1} \right), \quad \boldsymbol{Q}_2^{-1} = \frac{1}{\lambda_2(\boldsymbol{\beta})} \boldsymbol{I} - \boldsymbol{D}_{\boldsymbol{\beta}}^{-1}, \quad \boldsymbol{\mu}_2 = \boldsymbol{Q}^{-1} \boldsymbol{R} \boldsymbol{x},$$

$$\Rightarrow \boldsymbol{x} | \boldsymbol{v}, \alpha, \boldsymbol{\beta}, \boldsymbol{y} \sim \mathcal{N} \left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1} \right), \quad \boldsymbol{Q} = \frac{1}{\alpha} \boldsymbol{F}^T \boldsymbol{F} + \frac{1}{\lambda_2} \boldsymbol{R}^T \boldsymbol{R}, \quad \boldsymbol{\mu} = \boldsymbol{Q}^{-1} \left(\frac{1}{\alpha} \boldsymbol{F}^T \boldsymbol{y} + \boldsymbol{R}^T \boldsymbol{v}_2 \right)$$

PCG-IAS Algorithm

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

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

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

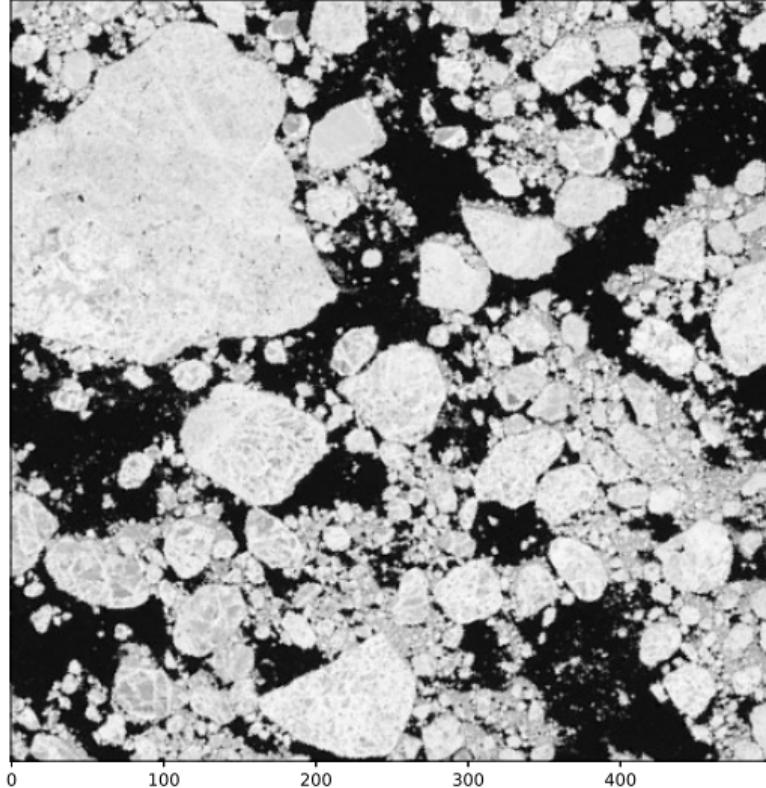
$$\boldsymbol{\beta}^k \leftarrow \arg \max_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta} | \alpha^k, \boldsymbol{x}^{k-1}, \boldsymbol{y}).$$

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

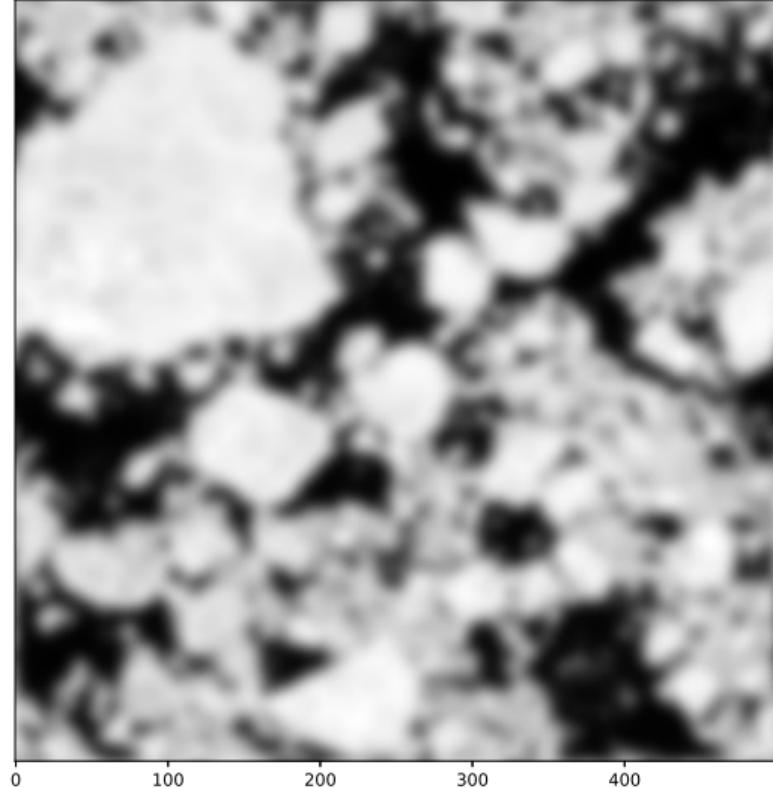
$$\boldsymbol{x}^k \leftarrow \arg \max_{\boldsymbol{x}} \pi(\boldsymbol{x} | \alpha^k, \boldsymbol{\beta}^k, \boldsymbol{v}^k, \boldsymbol{y}).$$

end for

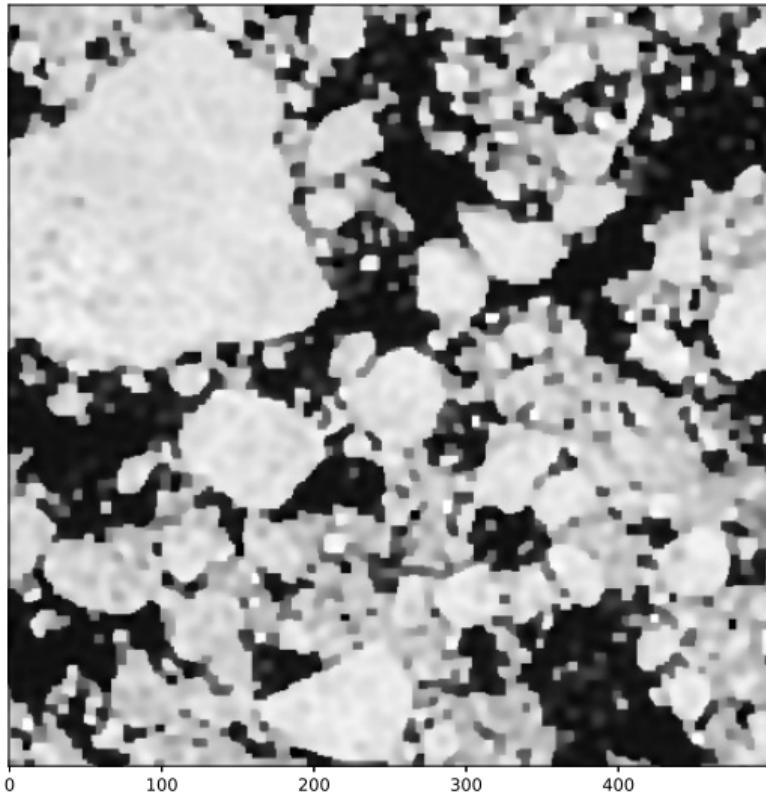
Ground truth



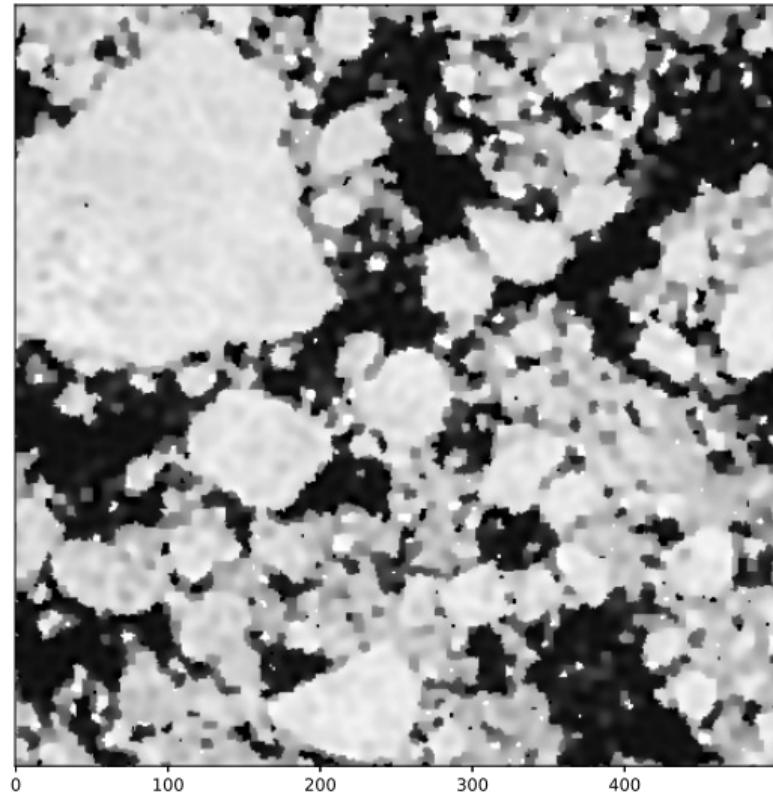
Observation



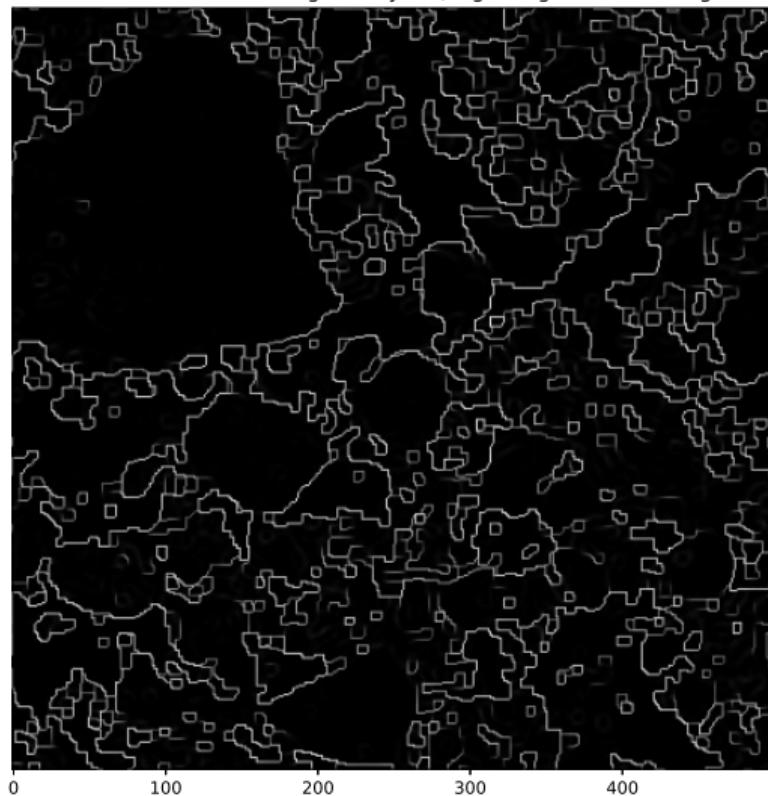
Non-convex solution without global hybrid, log of variance weights



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

