DARTMOUTH
Department of
Mathematics

HIERARCHICAL BAYESIAN INVERSE PROBLEMS

Graduate Student Seminar

17th January 2023

Jonathan Lindbloom

OUTLINE

1. Bayesian inverse problems

2. Hierarchical models

3. Solvers

BAYESIAN INVERSE PROBLEMS

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

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

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

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

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

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

Obstacles:

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

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 $\boldsymbol{Q}_{\text{noise}} = \sigma^{-2}\boldsymbol{I}$, the likelihood for the observed data $\boldsymbol{y}$ is Gaussian with

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

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

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

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

Maximum likelihood estimate

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

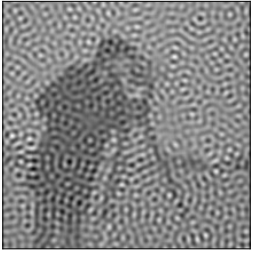
LEAST-SQUARES

Observation (no noise)

Least-squares reconstruction

Observation (with noise)

Least-squares reconstruction



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

Bayes' theorem

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

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 $\mathbb{E}_{\pi_{\text{post}}}[f(\boldsymbol{x})]$ or seek point estimates such as the mode.

Maximum a posteriori (MAP) estimate

$$\begin{aligned}\boldsymbol{x}_{\text{MAP}} &= \arg\max_{\boldsymbol{x}} \left\{\pi(\boldsymbol{y} \mid \boldsymbol{x})\pi(\boldsymbol{x})\right\} \\ &= \arg\min_{\boldsymbol{x}} \left\{-\log \pi(\boldsymbol{y} \mid \boldsymbol{x}) - \log \pi(\boldsymbol{x})\right\}\end{aligned}$$

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 $\boldsymbol{x}$.

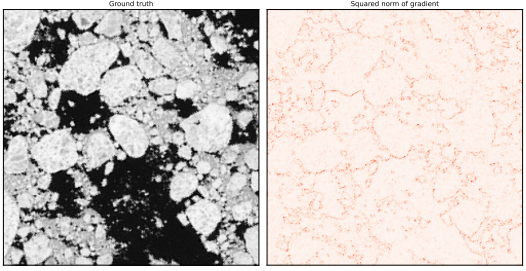


Figure: Discrete gradient operator.

GAUSSIAN PRIOR + GAUSSIAN LIKELIHOOD

Suppose that

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

GAUSSIAN PRIOR + GAUSSIAN LIKELIHOOD

Suppose that

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

Then

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

with

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

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

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

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

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

Gaussian posterior

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

$$\boldsymbol{Q}_{\text{post}} = \sum_{j=1}^{K} \boldsymbol{F}_j^T \boldsymbol{Q}_j \boldsymbol{F}_j,$$

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

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

TIKHONOV DE-BLURRING

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

WHY WE LIKE GAUSSIANS

- Explicit expression for normalized posterior density and moments

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

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

[3]Parker 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 $\boldsymbol{Q}_{\text{post}}\boldsymbol{\mu} =$ “rhs”

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

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

[3]Parker and Fox, “Sampling Gaussian Distributions in Krylov Spaces with Conjugate Gradients”, 2012.

WHY WE LIKE GAUSSIANS

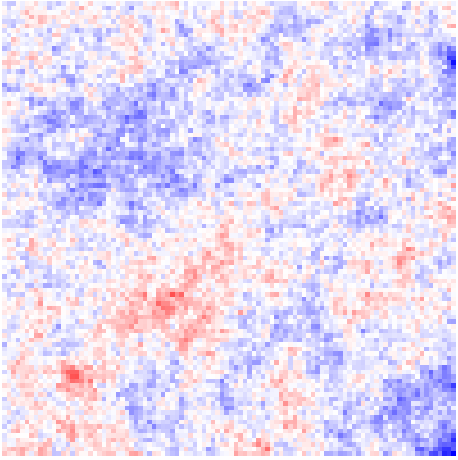
- Explicit expression for normalized posterior density and moments
- For MAP estimate, just need to solve system $\boldsymbol{Q}_{\text{post}}\boldsymbol{\mu} =$ "rhs"
- Sampling high-dimensional Gaussians has been studied extensively [1]
  - Factorization approaches, e.g., the Cholesky approach costing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

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

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

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

GAUSSIAN PRIOR SAMPLES



GAUSSIAN POSTERIOR SAMPLES

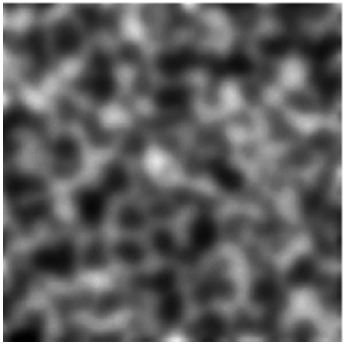


Figure: Gaussian prior sample.

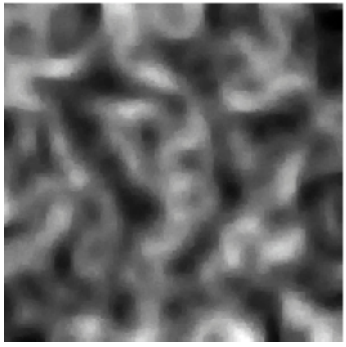


Figure: Total variation (TV) prior sample.

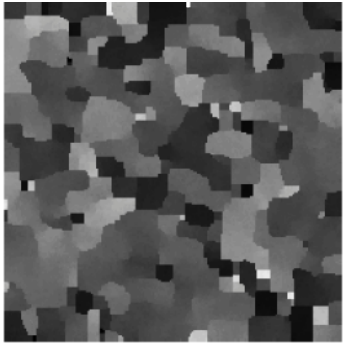


Figure: Cauchy difference prior sample.

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

NON-GAUSSIAN PRIORS

- Can more strongly promote sparsity under $\boldsymbol{R}$ 👍

NON-GAUSSIAN PRIORS

- Can more strongly promote sparsity under $\boldsymbol{R}$ 👍
- No longer have explicit expression for the posterior, MAP, etc. 👎

NON-GAUSSIAN PRIORS

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

$$E(\boldsymbol{x}) \coloneqq -\log \pi(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{1}{2} \underbrace{\|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{Q}}^2}_{\text{strongly convex}} \quad \underbrace{-\log \pi(\boldsymbol{x})}_{\text{convex?}}$$

NON-GAUSSIAN PRIORS

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

$$E(\boldsymbol{x}) \coloneqq -\log \pi(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{1}{2} \underbrace{\|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{Q}}^2}_{\text{strongly convex}} \quad \underbrace{-\log \pi(\boldsymbol{x})}_{\text{convex?}}$$

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

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

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:

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

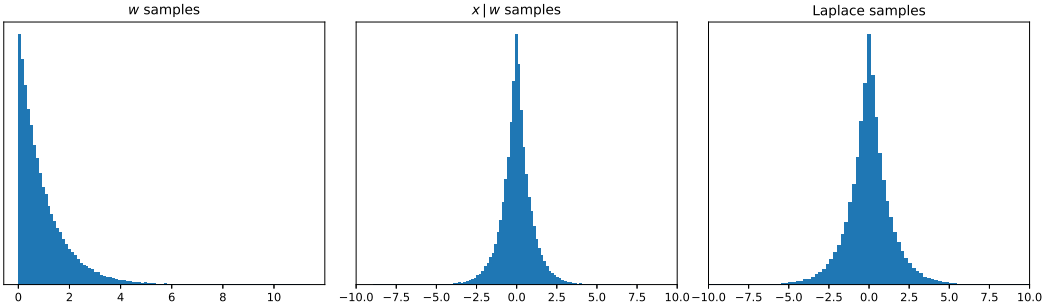
SCALE MIXTURES

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
$x \mid w$ samples
Laplace samples
0
2
4
6
8
10
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

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

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

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

[5] Calvetti 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

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

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

- Encompasses exponential, Gamma, inverse Gamma distributions

[5] Calvetti 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

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

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

[5] Calvetti 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

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

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

[5] Calvetti 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

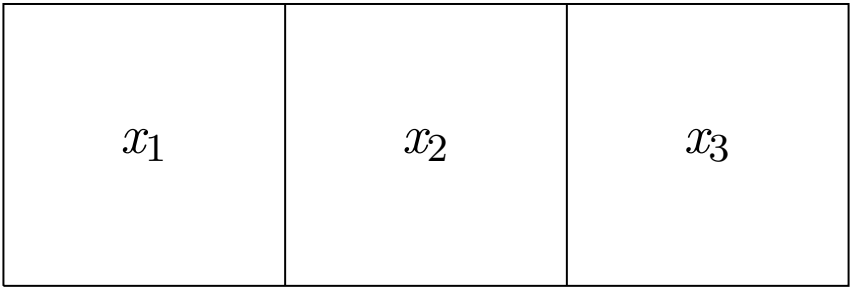
$$\pi_{\mathrm{GG}}(\theta \mid r, s, \vartheta) = \frac{|r|}{\vartheta \Gamma(s)} \left(\frac{\theta}{\vartheta}\right)^{rs-1} \exp\left\{-\left(\frac{\theta}{\vartheta}\right)^{r}\right\} \mathbb{1}_{>0}(\theta)$$

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

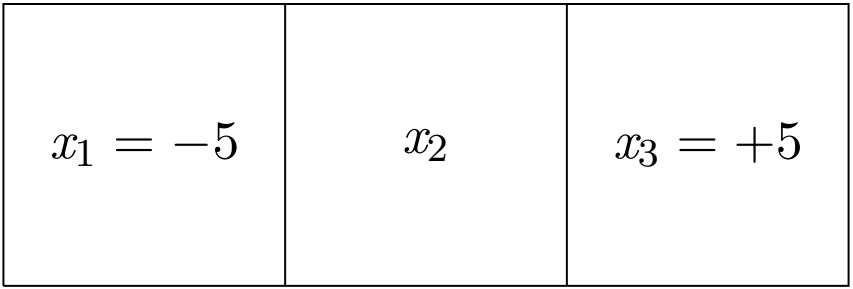
- Encompasses exponential, Gamma, inverse Gamma distributions
- Parameters (r, s) set the parameteric family of marginal prior, ϑ controls scale
- In certain limits, captures ℓ_p-norm priors for $0 < p < 2$, Student-t, Cauchy
- Detailed convexity analysis provided by [5]

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

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



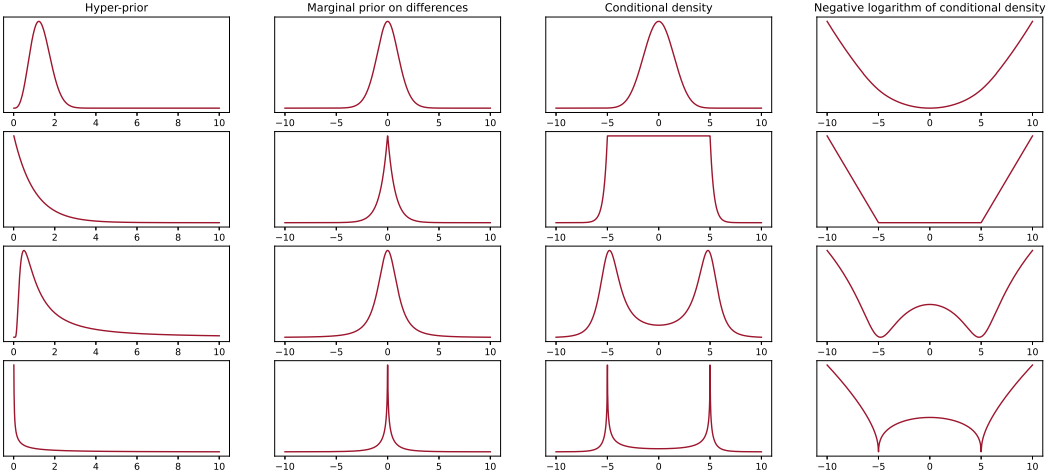
An unknown $\boldsymbol{x} = (x_1, x_2, x_3)$, with a prior on pixel differences



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

$$\begin{aligned}\boldsymbol{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\beta^{-1}\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\right)\\ \boldsymbol{y}\,|\,\boldsymbol{x} &\sim \mathcal{N}\left(\boldsymbol{F}\boldsymbol{x}, \alpha\boldsymbol{I}\right)\end{aligned}$$

vs.

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

THE HIERARCHICAL MODEL

$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}\left(\mathbf{0}, \left(\beta^{-1}\boldsymbol{R}^T\boldsymbol{R}\right)^{-1}\right) \\ \boldsymbol{y} \,|\, \boldsymbol{x} &\sim \mathcal{N}\left(\boldsymbol{F}\boldsymbol{x}, \alpha\boldsymbol{I}\right) \end{aligned}$$

vs.

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

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

NEGATIVE LOG POSTERIOR

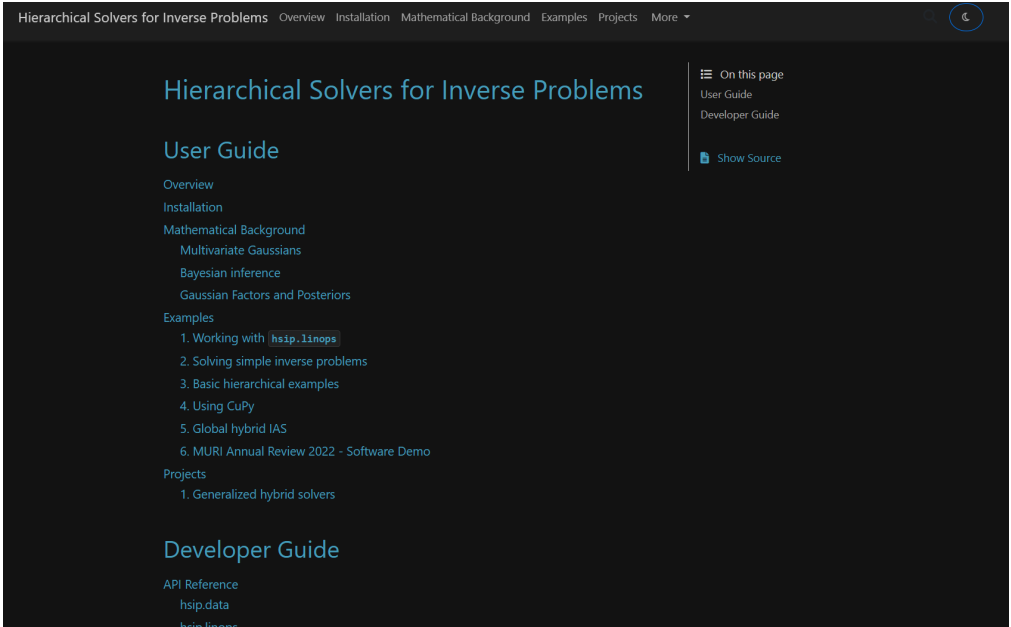
In the simpler Gaussian prior/likelihood case:

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

In our hierarchical model:

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

SOLVERS


Hierarchical Solvers for Inverse Problems
Overview
Installation
Mathematical Background
Examples
Projects
More
Hierarchical Solvers for Inverse Problems
On this page
User Guide
Developer Guide
Show Source
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
Developer Guide
API Reference
hsip.data

MAP ESTIMATION

- For both sampling and MAP estimation, we care about the convexity of of $E(\boldsymbol{x}, \alpha, \boldsymbol{\beta})$.
- Pre-existing work considered known noise covariance and showed convexity with the assumption $\ker(\boldsymbol{R}) = \{\mathbf{0}\}$, which is more restrictive than the common kernel condition

$$\ker(\boldsymbol{F}) \cap \ker(\boldsymbol{R}) = \{\mathbf{0}\}$$

- We have shown that the same convexity conditions hold under this less restrictive assumption, and extended to unknown noise covariance

ALGORITHMS FOR POINT ESTIMATION

- Seeking posterior mean:
  - For inverse Gamma variance hyper-prior, conjugacy relations exist that motivate a Bayesian coordinate descent (BCD) algorithm[6] (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.

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

[7] Calvetti, 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 $\boldsymbol{x}^0, \alpha^0, \boldsymbol{\beta}^0$.
for $k = 1, \ldots, n_{\text{maxits}}$ do
$\alpha^k \leftarrow \arg\max_\alpha \pi(\alpha \mid \boldsymbol{x}^{k-1}, \boldsymbol{\beta}^{k-1}, \boldsymbol{y})$.
$\boldsymbol{\beta}^k \leftarrow \arg\max_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta} \mid \alpha^k, \boldsymbol{x}^{k-1}, \boldsymbol{y})$.
$\boldsymbol{x}^k \leftarrow \arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x} \mid \alpha^k, \boldsymbol{\beta}^k, \boldsymbol{y})$.
end for

BAYESIAN COORDINATE DESCENT (BCD) ALGORITHM

BCD Algorithm

Choose initial $\boldsymbol{x}^0, \alpha^0, \boldsymbol{\beta}^0$.
for $k = 1, \ldots, n_{\text{maxits}}$ do

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

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

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

end for

UPDATES

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

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

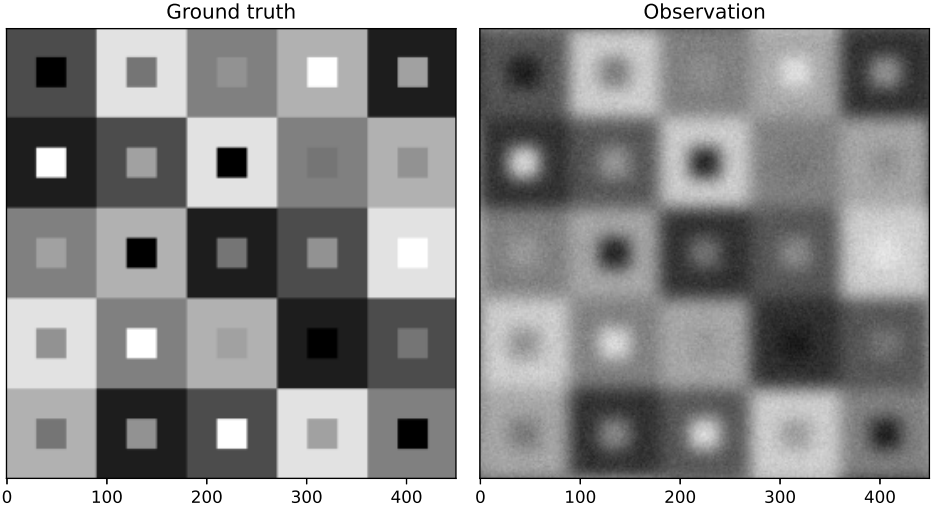
UPDATES

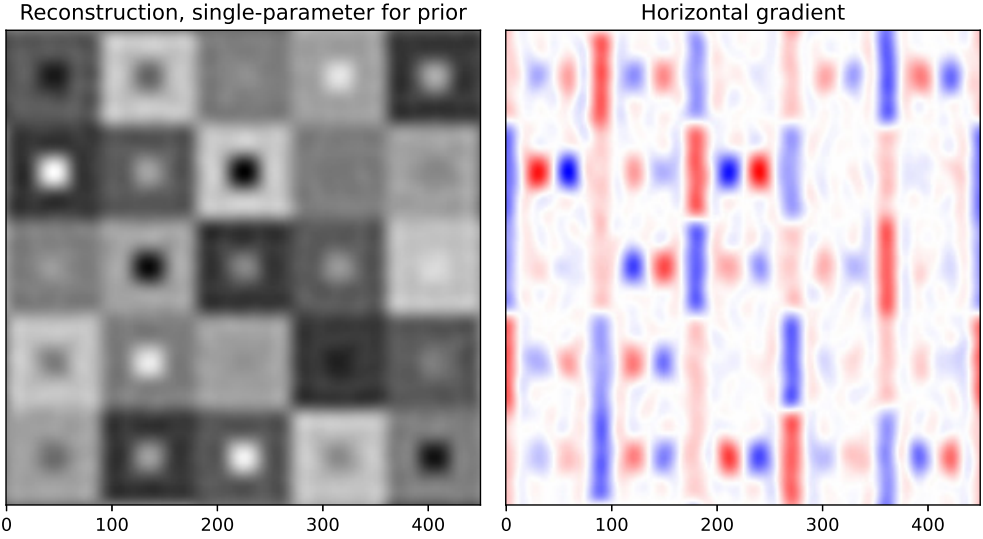
Each $\boldsymbol{x}$-update is the solution to

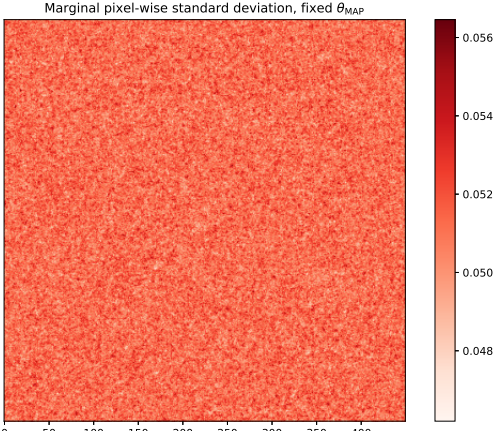
$$\arg\min_{\boldsymbol{x}} \left\{ \frac{1}{2\alpha} \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{2} \|\boldsymbol{R}\boldsymbol{x}\|_{\boldsymbol{D}_{\boldsymbol{\beta}}^{-1}}^2 \right\} = \left(\frac{1}{\alpha} \boldsymbol{F}^T \boldsymbol{F} + \boldsymbol{R}^T \boldsymbol{D}_{\boldsymbol{\beta}}^{-1}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

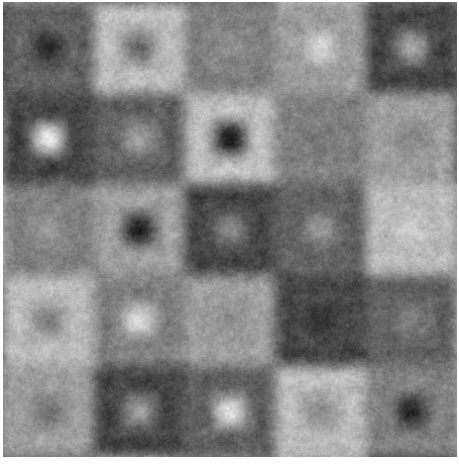
Updating $\alpha, \boldsymbol{\beta}$?

- 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

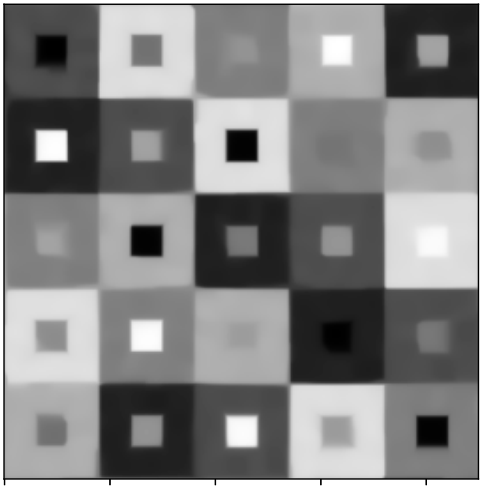

Ground truth
Observation
0
100
200
300
400
0
100
200
300
400


Reconstruction, single-parameter for prior
Horizontal gradient
0
100
200
300
400
0
100
200
300
400

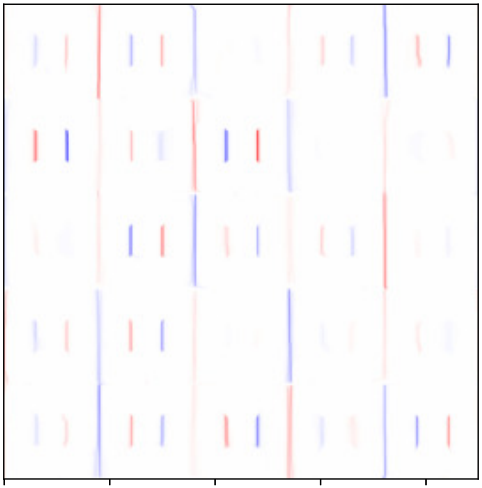

Marginal pixel-wise standard deviation, fixed θ_{MAP}
0.056
0.054
0.052
0.050
0.048
0
50
100
150
200
250
300
350
400

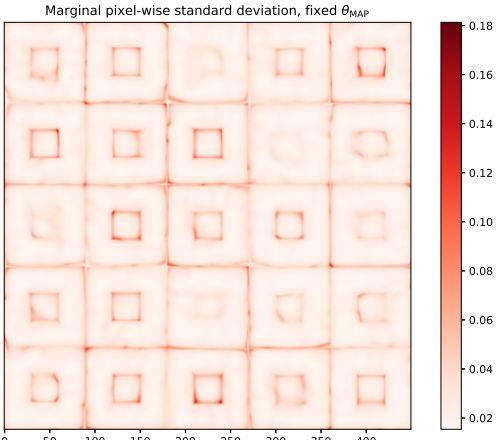


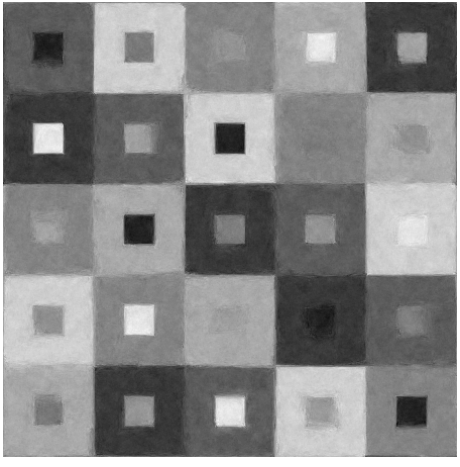
Reconstruction, many-parameter for prior

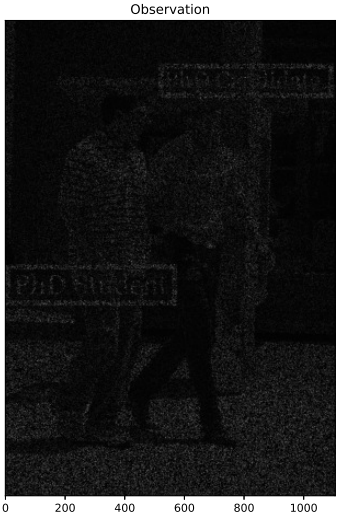


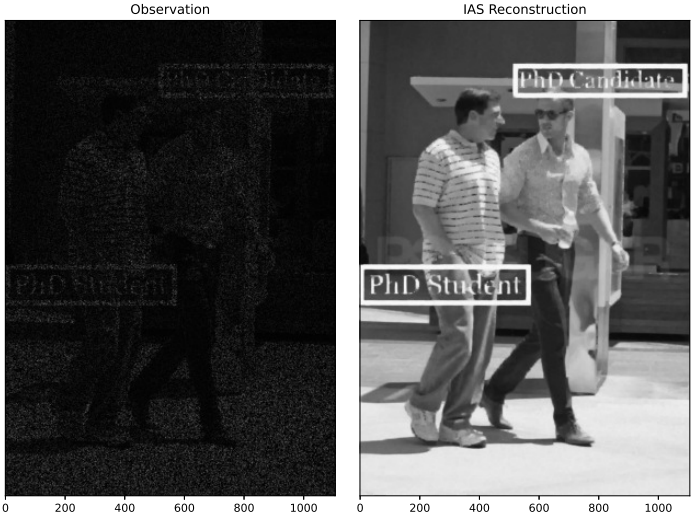
Horizontal gradient

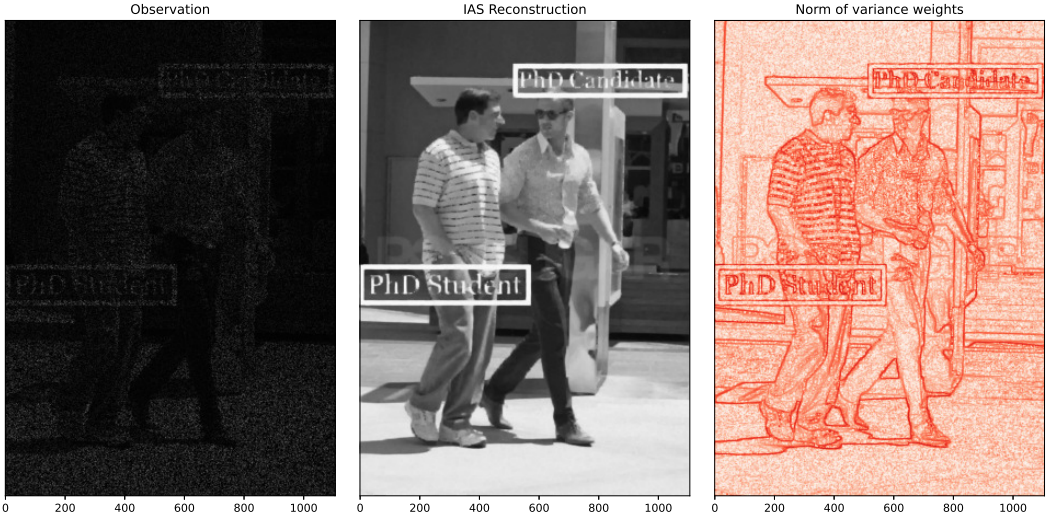


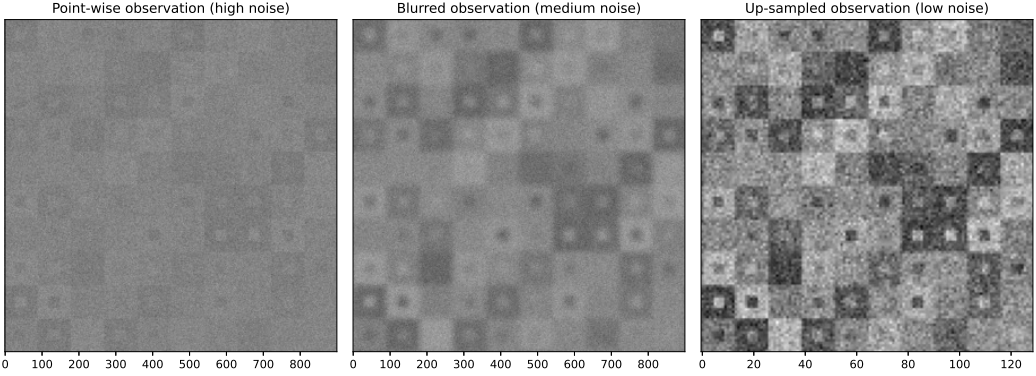

Marginal pixel-wise standard deviation, fixed θ_{MAP}
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
50
100
150
200
250
300
350
400




Observation
0
200
400
600
800
1000


Observation
IAS Reconstruction
PhD Candidate
PhD Student
0
200
400
600
800
1000
0
200
400
600
800
1000


Observation
IAS Reconstruction
Norm of variance weights
PhD Candidate
PhD Student
0
200
400
600
800
1000


Point-wise observation (high noise)
Blurred observation (medium noise)
Up-sampled observation (low noise)
0 100 200 300 400 500 600 700 800
0 100 200 300 400 500 600 700 800
0 20 40 60 80 100 120

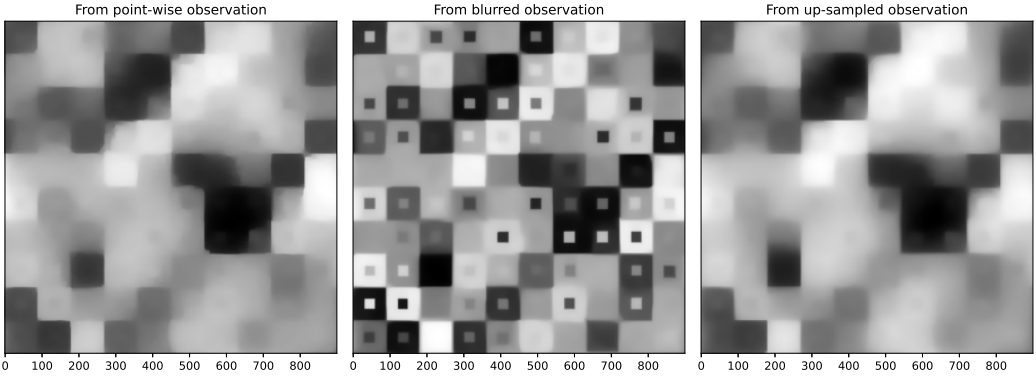


Figure: Separate reconstructions by each observation. % error in learned noise standard deviations (left to right): 0.56%, 0.25%, 48.98%.

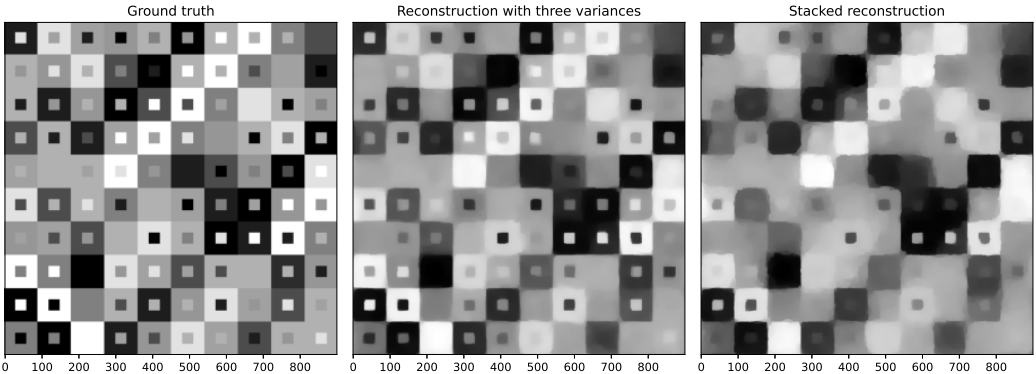
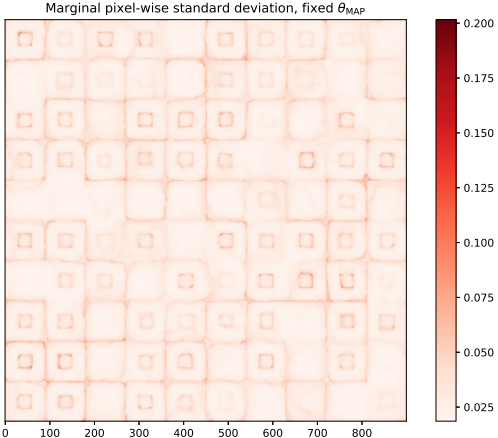
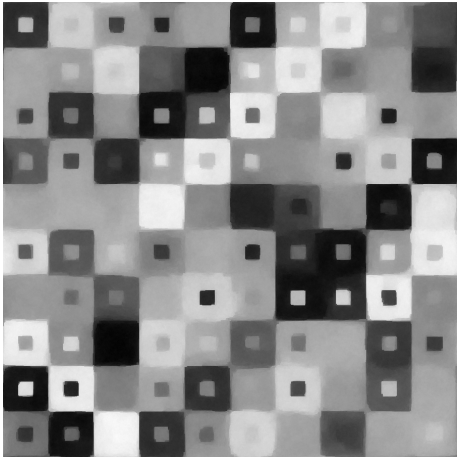


Figure: Reconstructions with three learned variance parameters vs. “stacked” reconstruction with a single learned variance parameter. % error in three learned noise standard deviations: 0.12%, 0.03%, 0.76%.

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


Marginal pixel-wise standard deviation, fixed θ_{MAP}
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0
100
200
300
400
500
600
700
800



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 $\boldsymbol{R}^T\boldsymbol{R}$ 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 $\boldsymbol{R}^T\boldsymbol{R}$ 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

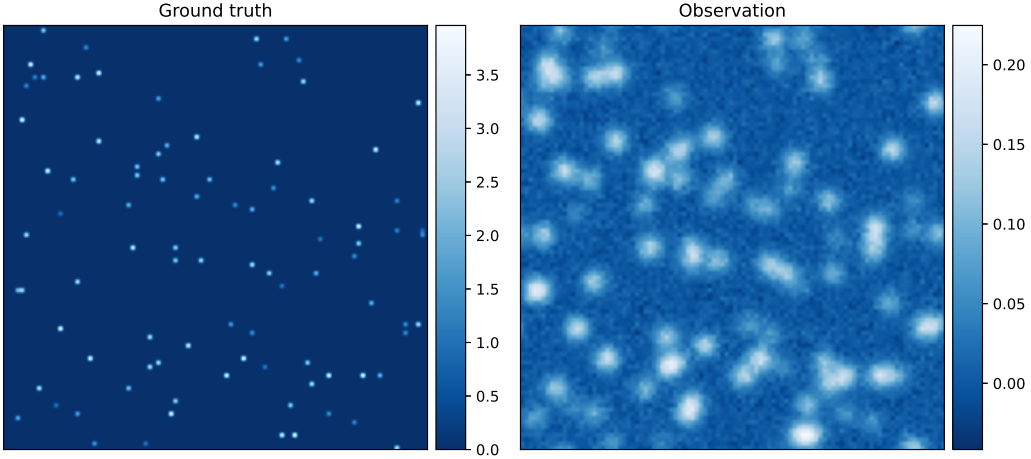
However, the same issue arises in joint MLE for mean and variance of Gaussian data.

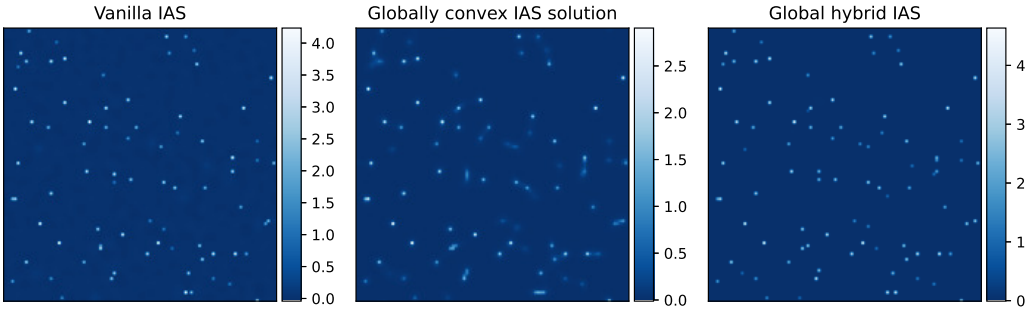
Next steps: analyze in terms of fixed point iteration, rather than convexity.

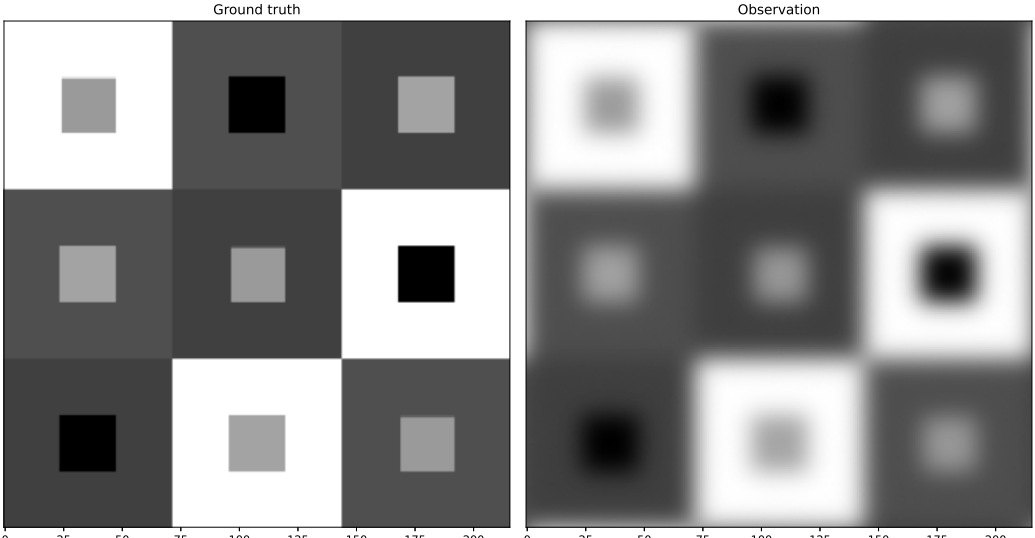
GLOBAL HYBRID IAS/BCD

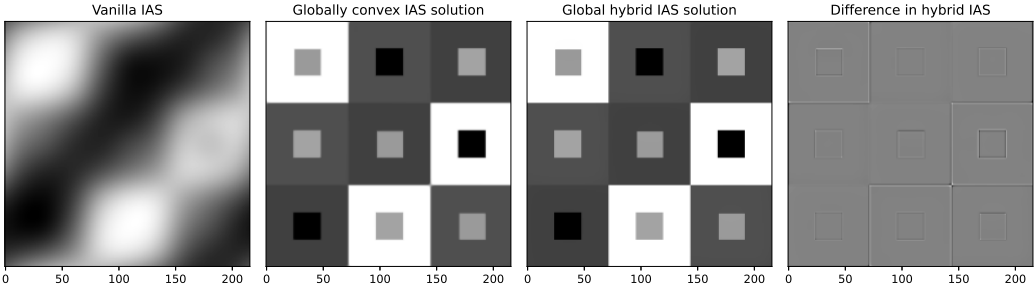
- We may like to impose a non-convex prior, but global convergence no longer guaranteed
- Idea: first solve a "close" convex problem, then use as initialization of a second solve using true prior of interest[8]
- No global convergence guarantees, but empirically a good strategy for obtaining better local minima
- Third contribution: global hybrid BCD; no conjugacy, yet conditional mean still known analytically

[8] Calvetti, Pragliola and Somersalo, "Sparsity Promoting Hybrid Solvers for Hierarchical Bayesian Inverse Problems", 2020.


Ground truth
Observation
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.20
0.15
0.10
0.05
0.00


Vanilla IAS
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Globally convex IAS solution
2.5
2.0
1.5
1.0
0.5
0.0
Global hybrid IAS
4
3
2
1
0


Ground truth
Observation
0
25
50
75
100
125
150
175
200


Vanilla IAS
Globally convex IAS solution
Global hybrid IAS solution
Difference in hybrid IAS
0
50
100
150
200

PCG-IAS/BCD

Fourth contribution:

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

$$\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{F}}\boldsymbol{\mathcal{F}}, \quad \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{R}}\boldsymbol{\mathcal{F}}$$

[9] Marnissi 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 $\boldsymbol{F}^T\boldsymbol{F}$ and $\boldsymbol{R}^T\boldsymbol{R}$ to directly compute an $\boldsymbol{x}$-update, e.g., if both block-circulant with circulant blocks (BCCB) then

$$\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{F}}\boldsymbol{\mathcal{F}}, \quad \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{R}}\boldsymbol{\mathcal{F}}$$

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

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

[9] Marnissi 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 $\boldsymbol{F}^T\boldsymbol{F}$ and $\boldsymbol{R}^T\boldsymbol{R}$ to directly compute an $\boldsymbol{x}$-update, e.g., if both block-circulant with circulant blocks (BCCB) then

$$\boldsymbol{F}^T\boldsymbol{F} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{F}}\boldsymbol{\mathcal{F}}, \quad \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{\mathcal{F}}^H\boldsymbol{\Lambda}_{\boldsymbol{R}}\boldsymbol{\mathcal{F}}$$

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

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

- We have made use of recent data augmentation techniques[9] to get around this

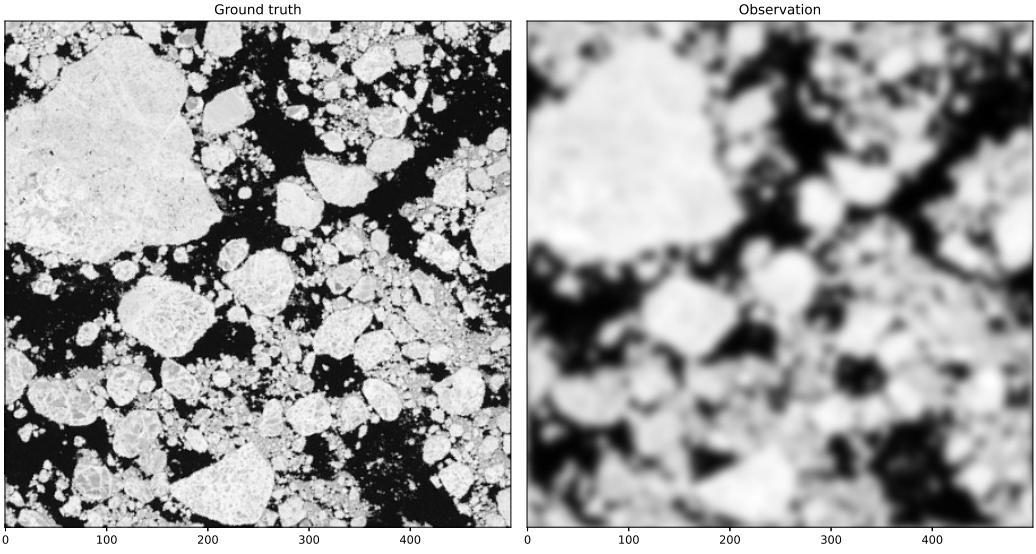
[9] Marnissi et al., "A Data Augmentation Approach for Sampling Gaussian Models in High Dimension", 2019.

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

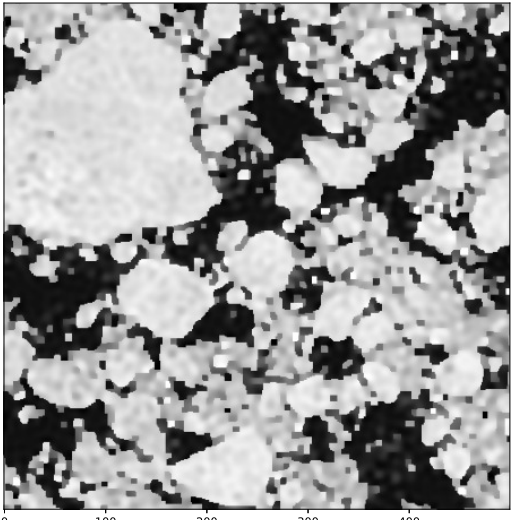
$$\Rightarrow \boldsymbol{x} \,|\, \boldsymbol{v}, \alpha, \boldsymbol{\beta}, \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right), \quad \boldsymbol{Q} = \frac{1}{\alpha}\boldsymbol{F}^T\boldsymbol{F} + \frac{1}{\lambda_2}\boldsymbol{R}^T\boldsymbol{R}, \quad \boldsymbol{\mu}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

PCG-IAS Algorithm

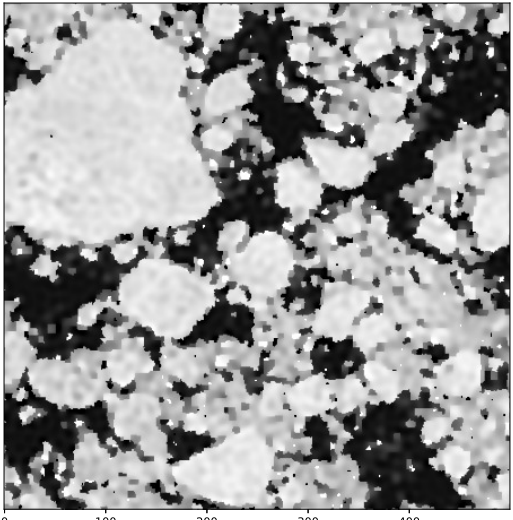
Choose initial $\boldsymbol{x}^0, \alpha^0, \boldsymbol{\beta}^0, \boldsymbol{v}^0$.
for $k = 1, \ldots, n_{\text{maxits}}$ do
  $\alpha^k \leftarrow \arg\max_\alpha \pi(\alpha \mid \boldsymbol{x}^{k-1}, \boldsymbol{\beta}^{k-1}, \boldsymbol{y})$.
  $\boldsymbol{\beta}^k \leftarrow \arg\max_{\boldsymbol{\beta}} \pi(\boldsymbol{\beta} \mid \alpha^k, \boldsymbol{x}^{k-1}, \boldsymbol{y})$.
  $\boldsymbol{v}^k \leftarrow \arg\max_{\boldsymbol{v}} \pi(\boldsymbol{v} \mid \alpha^k, \boldsymbol{\beta}^k, \boldsymbol{x}^{k-1}, \boldsymbol{y})$
  $\boldsymbol{x}^k \leftarrow \arg\max_{\boldsymbol{x}} \pi(\boldsymbol{x} \mid \alpha^k, \boldsymbol{\beta}^k, \boldsymbol{v}^k, \boldsymbol{y})$.
end for


Ground truth
Observation
0
100
200
300
400
0
100
200
300
400

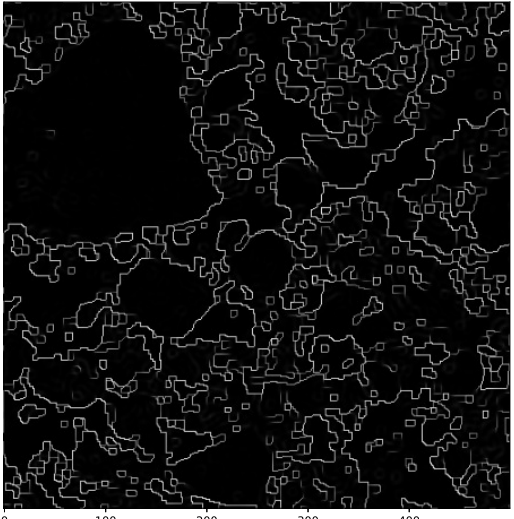
Non-convex solution without global hybrid, log of variance weights



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