Sparse Bayesian Inference with Regularized Gaussian Distributions

@ AIP 2023, MS12 1: Fast optimization-based methods for inverse problems

Jasper M. Everink, Technical University of Denmark, jmev@dtu.dk

Joint work with Martin S. Andersen and Yiqiu Dong

Computational **U**ncertainty **Q**uantification for **I**nverse problems

DTU Compute
Department of Applied Mathematics and Computer Science

Sparsity in linear least squares estimation

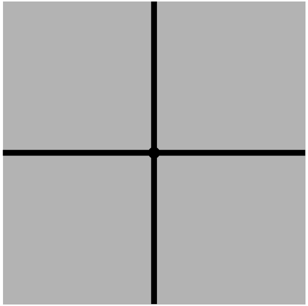
Parameters: $x \in \mathbb{R}^n$, linear forward operator: $A \in \mathbb{R}^{m\times n}$, and data: $b = Ax + e \in \mathbb{R}^m$.

Regularized linear least squares

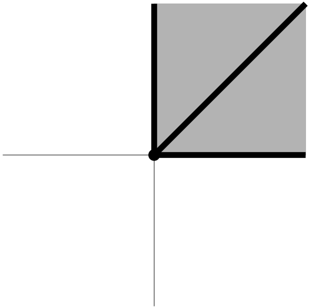
$$\min_{z\in\mathbb{R}^n} \left\{ \frac{\lambda}{2}\|Az - b\|_2^2 + f(z) \right\}$$

can give sparse solutions,
i.e., many zeros in x and/or Lx.

$f(x) = \|x\|_1$
LASSO



$f(x) = \|Lx\|_1 + \chi_{\mathbb{R}^n_{\geq 0}}(x)$
Constrained generalized LASSO



Sparsity with continuous distributions

Given a continuous posterior distribution, e.g.,

$$\pi(x \mid b) \propto \exp\left(-\frac{\lambda}{2}\|Ax-b\|_2^2 - f(x)\right),$$

then, for "approximate" sparsity,

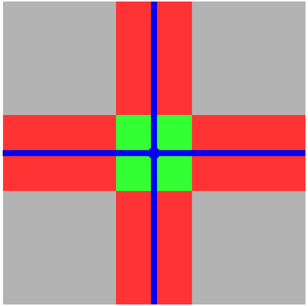
$$\mathbb{P}(\blacksquare) > 0, \quad \mathbb{P}(\blacksquare) > 0,$$

but for "true" sparsity,

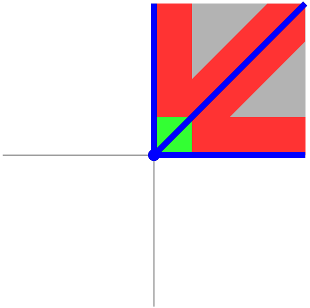
$$\mathbb{P}(\blacksquare) = 0.$$

Does the difference matter?

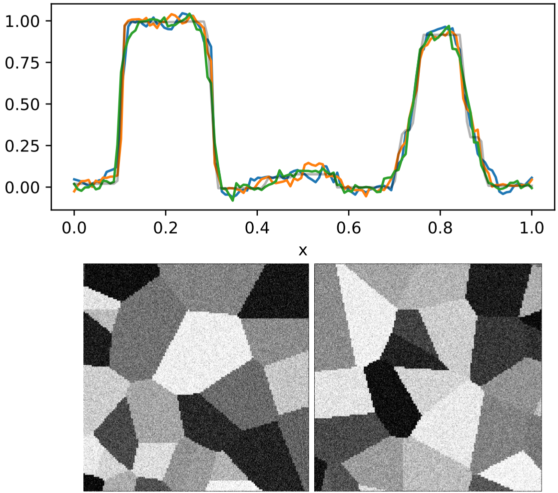
$f(x) = \|x\|_1$
Bayesian LASSO



$f(x) = \|Lx\|_1 + \chi_{\mathbb{R}^n_{\geq 0}}(x)$
Bayesian constrained generalized LASSO

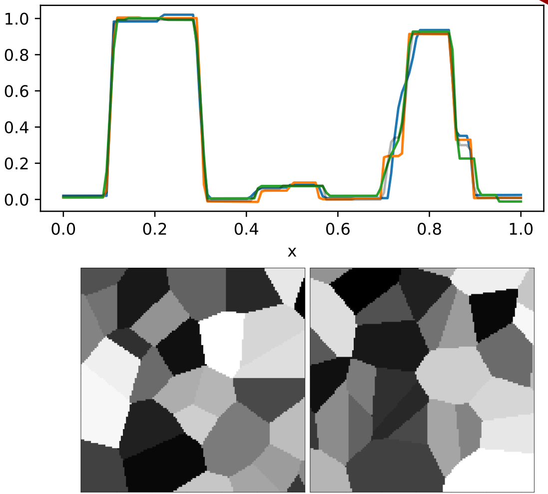


Samples we get from a continuous distribution



exaggerated

Sparse samples we (might) want



Sparsity with distributions of varying dimensions

Varying dimension model:
Partition the domain in different models

$$F_i,$$

define model priors

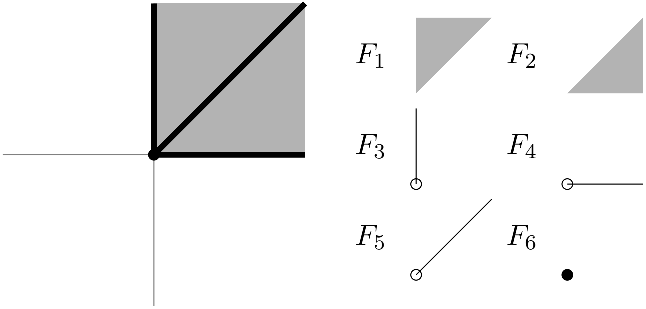
$$\mathbb{P}(F_i) \quad \text{and}$$

define conditional densities

$$\pi(x \mid F_i),$$

of dimension $\dim(F_i)$.

$$f(x) = \|Lx\|_1 + \chi_{\mathbb{R}^n_{\geq 0}}(x)$$



Sampling with sparsity

Sampling from varying dimension models

Reversible-Jump Markov Chain Monte Carlo (RJMCMC), e.g., STMALA, which can be difficult to work with in high-dimensional problems.

Our method: regularized Gaussian distribution

$$x \mid b := \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \| Az - \hat{b} \|^2_{\Sigma^{-1}} + f(z) \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \Sigma).$$

Trade-off:

Disadvantage: Implicit assumptions on the prior behind $x \mid b$.
Advantage: Use tools from optimization theory to analyze and sample.

Regularized Gaussian distribution

(Linear) Randomize-Then-Optimize (RTO)/Perturbation Optimization (PO):
If $\pi(x \mid b) \propto \exp\left(-\frac{1}{2}\|Ax - b\|^2_{\Sigma^{-1}}\right)$, i.e., Gaussian posterior, then

$$x \mid b = \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Az - \hat{b}\|^2_{\Sigma^{-1}} \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \Sigma).$$

Randomization/Perturbation: sample from $\hat{b}$,
Optimization: solve the optimization problem with the $\hat{b}$ sample.

Regularized Gaussian Distribution:

$$x \mid b := \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2}\|Az - \hat{b}\|^2_{\Sigma^{-1}} + f(z) \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \Sigma).$$

Is $x \mid b$ is a well-defined probability distribution? How does $x \mid b$ look like?

Rank issue

$$x \mid b := \min_{z\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Az - \hat{b}\|^2_{\Sigma^{-1}} + f(z) \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \Sigma).$$

If $\mathrm{rank}(A) < n$, i.e., there is not enough (artificial) data,
then **the probability distribution is not always well-defined***.

For the remainder of this talk, assume $\mathrm{rank}(A) = n$.

Regularization is post-processing

Formulation 1: Regularized Gaussian distribution

$$x \mid b := \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Az - \hat{b}\|^2_{\Sigma^{-1}} + f(z) \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \Sigma).$$

Formulation 2: Proximal post-processed posterior

$$x^\star := \min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|Az - \hat{b}\|^2_{\Sigma^{-1}} \right\}, \quad \text{i.e} , \quad \pi(x^\star) \propto \exp\left(-\frac{1}{2} \|Ax^\star - b\|^2_{\Sigma^{-1}} \right),$$

then

$$x \mid b = \mathrm{prox}_f^{\Lambda^{-1}}(x^\star) := \operatorname*{argmin}_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \|z - x^\star\|^2_{\Lambda^{-1}} + f(z) \right\}, \quad \text{with} \quad \Lambda^{-1} = \mathrm{Cov}(x^\star)^{-1}.$$

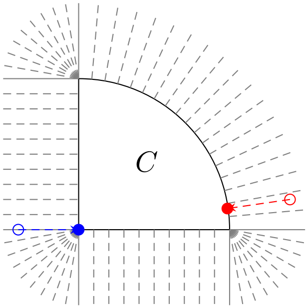
Why not use a cheaper post-processor?

Example - Constrained

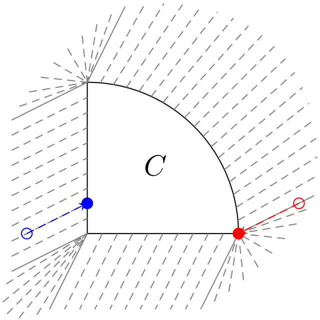
If $f(x) = \chi_C(x)$, then $\text{prox}_f^{\Lambda^{-1}}(x^\star) = \Pi_C^{\Lambda^{-1}}(x^\star)$ is an oblique projection.

$$\Pi_C^I(x^\star) := \underset{x \in C}{\operatorname{argmin}} \|x - x^\star\|_2^2 \quad \text{or} \quad \Pi_C^{\Lambda^{-1}}(x^\star) := \underset{x \in C}{\operatorname{argmin}} \|x - x^\star\|_{\Lambda^{-1}}^2$$

Π_C^I



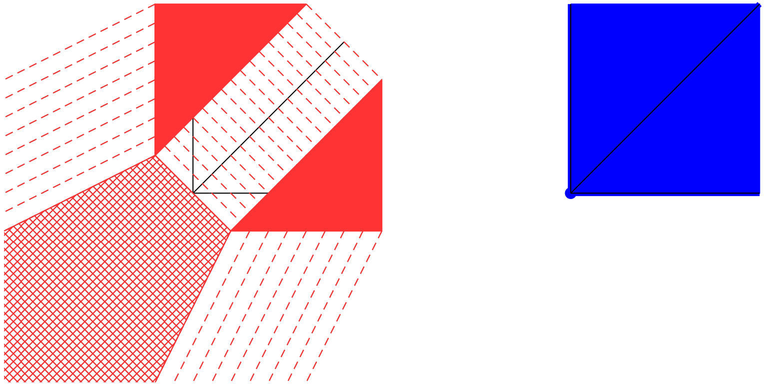
$\Pi_C^{\Lambda^{-1}}$



Example - Nonnegative Total Variation

$$f(x) = |x_2 - x_1| + \chi_{\mathbb{R}^2_{\geq 0}}(x)$$

$x^\star$ $\qquad\qquad$ $\mathrm{prox}_f^{\Lambda^{-1}}(x^\star)$



$$\mathbb{P}(x^\star \in \text{[red region]}) > 0 \quad \Longrightarrow \quad \mathbb{P}(\mathrm{prox}_f^{\Lambda^{-1}}(x^\star) \in \text{[blue region]}) > 0$$

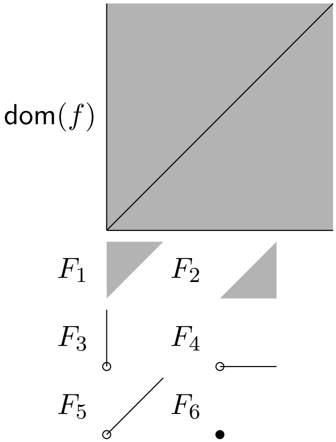
Low-dimensional subspaces

$\text{prox}_f^{\Lambda^{-1}}(x^\star)$ has positive probability on low-dimension subspaces representing sparsity.

Similar results hold if f is:

- convex piecewise linear/polyhedral
- curved constraints, e.g., ball
- group lasso, e.g., $\sum_{i=1}^{k} \|D_i x\|_2$

$$f(x) = |x_2 - x_1| + \chi_{\mathbb{R}^2_{\geq 0}}(x)$$



A Bayesian look

$$f(x) = |x_2 - x_1| + \chi_{\mathbb{R}^2_{\geq 0}}(x) = \|Lx\|_1 + \chi_{\mathbb{R}^n_{\geq 0}}(x)$$

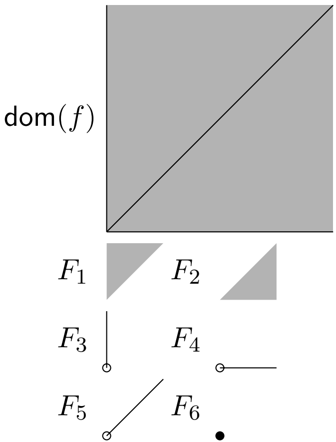
For f convex piecewise linear:

Probability distribution satisfies

$$\pi(x \mid b, F_j) \propto \exp\left(-\frac{1}{2}\|Ax - b\|^2_{\Sigma^{-1}} - f(x)\right).$$

Conditional prior:

$$\pi(x \mid F_i) \propto \frac{\pi(x \mid b, F_i)}{\pi(b \mid x, F_i)} \propto \exp\left(-f(x)\right).$$



Hierarchical model

Assume f is positive homogeneous and convex piecewise linear, e.g, $f(x) = \|Lx\|_1 + \chi_{\mathbb{R}^n_{\geq 0}}$.

$$x \mid b, \lambda, \gamma := \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{\lambda}{2} \|Ax - \hat{b}\|_2^2 + \gamma f(x) \right\}, \quad \text{with} \quad \hat{b} \sim \mathcal{N}(b, \lambda^{-1} I) \tag{1}$$

Add hyperpriors: $\lambda \sim \Gamma(\alpha_\lambda, \beta_\lambda)$ and $\gamma \sim \Gamma(\alpha_\gamma, \beta_\gamma)$.

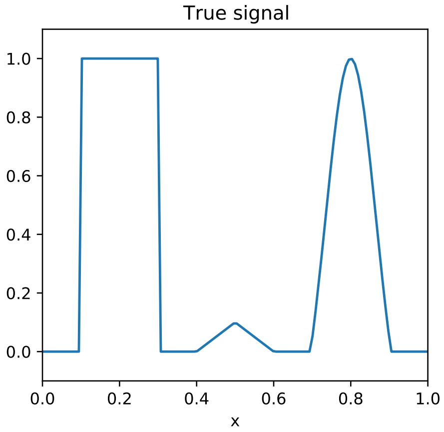
Hierarchical Gibbs Sampler

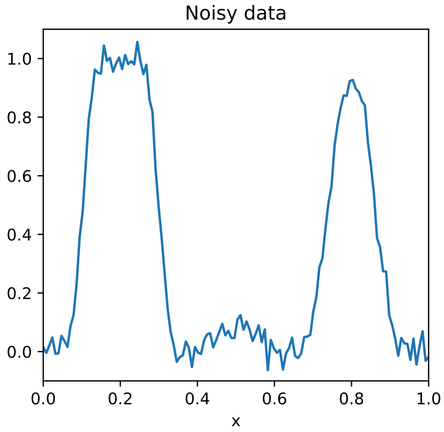
Repeat:

1. λ: $\lambda_k \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|Ax^{k-1} - b\|_2^2 + \beta_\lambda\right)$,
2. γ: $\gamma_k \sim \Gamma\left(\dim(F(x^{k-1})) + \alpha_\gamma, f(x^{k-1}) + \beta_\gamma\right)$, $\dim(F(x^{k-1}))$ is the level of sparsity
3. x: Sample $x|b, \lambda^k, \gamma^k$ through (1).

Sampling from (1) can only be done efficiently **approximately**, e.g., using ADMM.

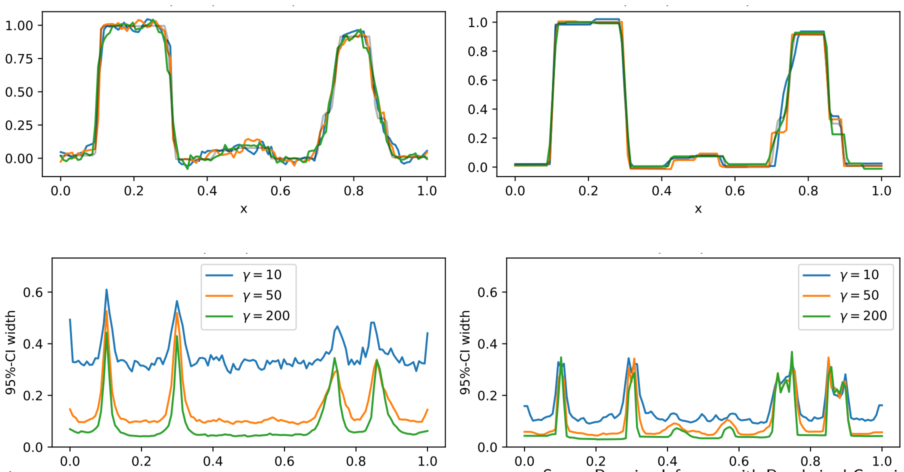
Numerical example: 1D deblurring with TV





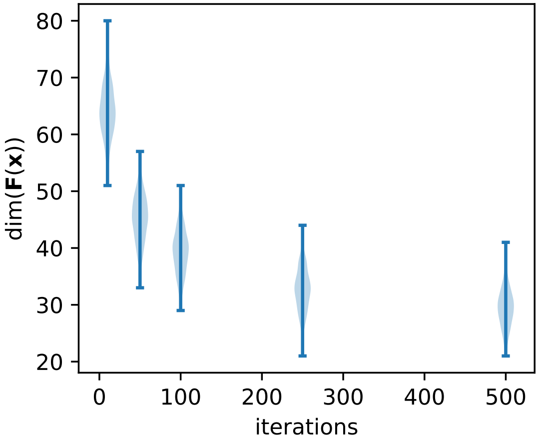
Numerical example: deblurring with TV (no Gibbs)

$$\pi(x \mid b) \propto \exp\left(-\frac{\lambda}{2}\|Ax - b\|_2^2 - \gamma\|Lx\|_1\right) \quad \textit{versus} \quad x \mid b = \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \left\{\frac{\lambda}{2}\|Ax - \hat{b}\|_2^2 + \gamma f(x)\right\}$$



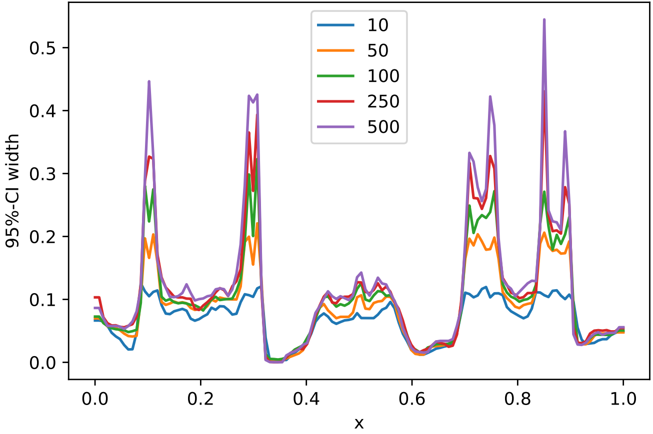
Numerical example: Gibbs denoising with TV

Computational cost (iterations) *versus* accuracy (sparsity)



Numerical example: Gibbs denoising

Computational cost (iterations) *versus* accuracy (componentwise credibility interval width)



To be continued

Overview

Sample efficiently from an implicit varying dimension model:

> Regularized Gaussian distribution:
>
> $$x \mid b := \min_{z\in\mathbb{R}^n} \left\{ \frac{1}{2}\|Az - \hat{b}\|^2_{\Sigma^{-1}} + f(z) \right\},$$
>
> with $\hat{b} \sim \mathcal{N}(b, \Sigma)$.

Assuming $\operatorname{rank}(A) = n$, then

- Positive probability on subspaces
- Conditional prior/conditionally Bayesian
- Gibbs sampler for a hierarchical model

Well-behaved subdifferential

$\gamma \sum_i \|D_i x - d_i\|_p$

$\chi_{D^n \cap \mathbb{R}^n_{\geq 0}}(x)$

Piecewise linear

$\chi_{[0,1]^n}(x)$ $\gamma\|Lx - c\|_1$

$\max_i\{a_i^T x + b_i\}$

Positive homogeneous

$\chi_{\mathbb{R}^n_{\geq 0}}(x)$ $\gamma\|Lx\|_1$

$\max_i\{a_i^T x\}$

Preprint on ArXiv: *Sparse Bayesian inference with regularized Gaussian distributions*

Sparse Bayesian Inference with Regularized Gaussian Distributions

@ AIP 2023, MS12 1: Fast optimization-based methods for inverse problems

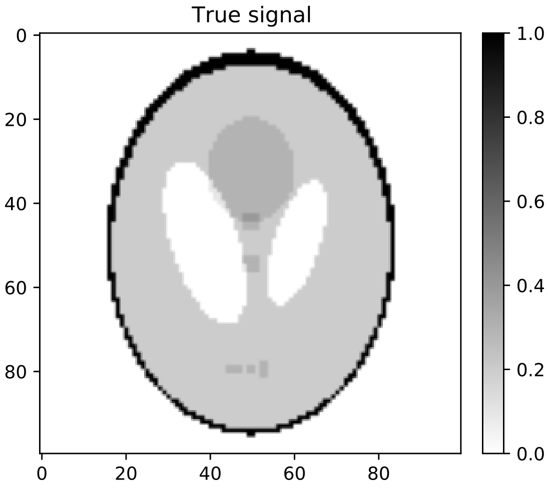
Jasper M. Everink, Technical University of Denmark, jmev@dtu.dk

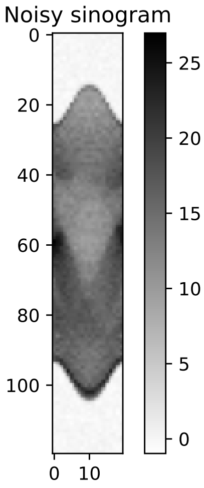
Joint work with Martin S. Andersen and Yiqiu Dong

Computational **U**ncertainty **Q**uantification for **I**nverse problems

DTU Compute
Department of Applied Mathematics and Computer Science

Numerical example: NNTV regularized CT (rank$(A) < n$)





Numerical example: NNTV regularized CT ($\mathrm{rank}(A) < n$)