A Bayesian model for Structured Sparse Sequences

Ingmar Schuster (FU Berlin)

(with Jean-Bernard Salomond, Université Paris Est - Créteil)

March 7, 2017

OUTLINE

Introduction

BAYESIAN STATISTICS

- Want to explain observed data X from parameters μ
- given likelihood model $p(X|\mu)$ and prior distribution $p(\mu)$ use Bayes theorem to get posterior distribution

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)}$$

- *Bayes estimators* are integrals wrt $p(\mu|X)$

Bayesian Statistics

- Want to explain observed data X from parameters μ
- given likelihood model $p(X|\mu)$ and prior distribution $p(\mu)$ use Bayes theorem to get posterior distribution

$$p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)}$$

- *Bayes estimators* are integrals wrt $p(\mu|X)$

MD	**Bayes**	
Boltzmann dens.	$p(\mu|X)$	
unnorm. Boltzmann dens.	$p(X|\mu)p(\mu)$	prop. to $p(\mu|X)$
free energy	$p(X)$	model selection

PROBLEM SETUP

Given $X \in \mathbb{R}^n$ we want

- denoised version of X
- test: which dimensions contain signal, which noise?
- contribution
  - use structure information about X
  - proof good theoretical sparsity properties

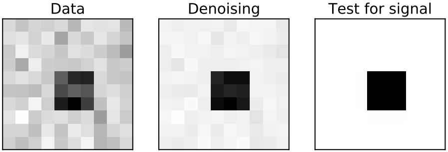


Figure: Images inherently structured by pixel adjacency

Problem setup

Given $X \in \mathbb{R}^n$ we want

- denoised version of X
- test: which dimensions contain signal, which noise?
- contribution
  - use structure information about X
  - proof good theoretical sparsity properties

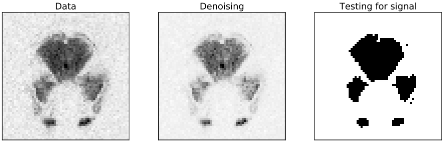


Figure: Images inherently structured by pixel adjacency

Model and statistical test

Normal-Gamma Shrinkage Model

$$\sigma_i^2 \sim \mathrm{Gamma}(\alpha, \beta) \qquad \text{for } i \in \{1, \dots, n\}$$
$$\mu \sim \mathcal{N}(0, \mathrm{Diag}[\sigma^2])$$
$$X \sim \mathcal{N}(\mu, \delta I_n)$$

- σ_i^2 is scale of signal at component i
- μ denoised version of data X
- assume Gaussian noise with variance δ
- example of a shrinkage model (like LASSO)

Dependent Scale Mixture Model (DSM)

$$\sigma_i^2 \sim \text{Gamma}(\alpha, \beta) \qquad \text{for } i \in \{1, \ldots, n\}$$

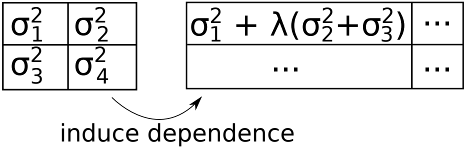
$$\mu \sim \mathcal{N}(0, \text{Diag}[A\sigma^2])$$

$$X \sim \mathcal{N}(\mu, \delta I_n)$$

Dependent Scale Mixture Model (DSM)

$$\sigma_i^2 \sim \mathrm{Gamma}(\alpha, \beta) \qquad \text{for } i \in \{1, \dots, n\}$$
$$\mu \sim \mathcal{N}(0, \mathrm{Diag}[A\sigma^2])$$
$$X \sim \mathcal{N}(\mu, \delta I_n)$$

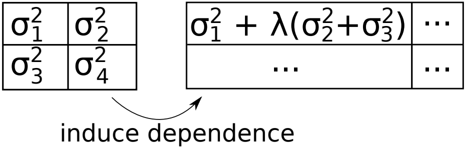
Grid structure: Prior on μ_1 is Gaussian with variance



Dependent Scale Mixture Model (DSM)

$$\sigma_i^2 \sim \text{Gamma}(\alpha, \beta) \qquad \text{for } i \in \{1, \dots, n\}$$
$$\mu \sim \mathcal{N}(0, \text{Diag}[A\sigma^2])$$
$$X \sim \mathcal{N}(\mu, \delta I_n)$$

Grid structure: Prior on μ_1 is Gaussian with variance

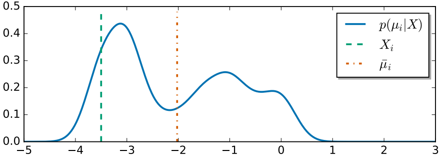


- dependence between adjacent signals induced using A
- still obtain good theoretical properties for sparsity test

TEST (1)

$$\sigma_i^2 \sim \text{Gamma}(\alpha, \beta) \quad \mu \sim \mathcal{N}(0, \text{Diag}[A\sigma^2]) \quad X \sim \mathcal{N}(\mu, \delta I_n)$$

- let $\bar{\mu} = \mathbb{E}_{p(\mu|X)}[\mu]$



- average μ will be closer to 0 than X (shrinkage effect)
- test: what's the best explanation for X_i, signal or noise?
  - after average denoising closer to data than to 0?
  - there is signal if

$$\bar{\mu}_i / X_i > 0.5$$

TEST (2)

$$\sigma_i^2 \sim \text{Gamma}(\alpha, \beta) \quad \mu \sim \mathcal{N}(0, \text{Diag}[A\sigma^2]) \quad X \sim \mathcal{N}(\mu, \delta I_n)$$

- can integrate over μ in closed form given σ^2

$$\frac{\bar{\mu}_i}{X_i} = \cdots = \int \frac{(A\sigma^2)_i}{(A\sigma^2)_i + \delta} p(\sigma^2|X) \mathrm{d}\sigma^2$$

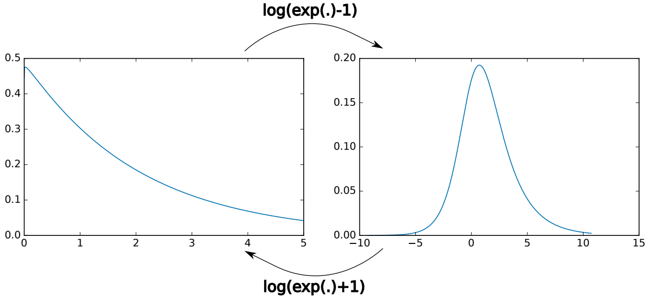
- Test for signal: $\int \frac{(A\sigma^2)_i}{(A\sigma^2)_i+\delta} p(\sigma^2|X)\mathrm{d}\sigma^2 > 0.5$

Inference

Change of variables

- Want to compute integral wrt $p(\sigma^2|X)$
- supported on $\mathbb{R}^n_+$
- no signal means maximum probability close to 0
- for easier numeric integration transform σ^2 to

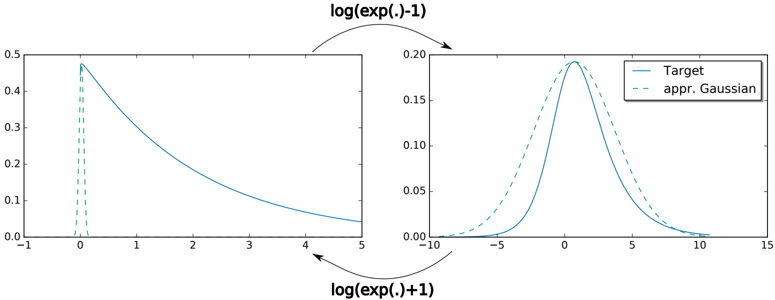
$$\tau = \log(\exp(\sigma^2) - 1) \in \mathbb{R}^n$$



Gaussian Approximation

- idea: 2nd order Taylor approxmation
- find $\widehat{\tau} = \arg\max_\tau \log p(\tau|X)$
- compute curvature $\widehat{C} = \nabla^2 \log p(\widehat{\tau}|X)$
- Monte Carlo with Gaussian approximation to $p(\tau|X)$:

$$\mathcal{N}(\widehat{\tau}, (-\widehat{C})^{-1})$$



ADAPTIVE METROPOLIS

Adaptive Metropolis (AM)

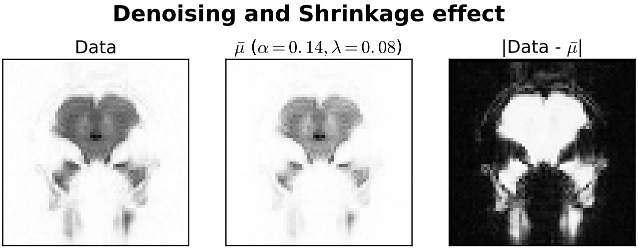
- uses Metropolis-Hastings with proposal $\mathcal{N}(\boldsymbol{\tau}_t, K_t)$
- K_t is scaled down empirical covariance of $\{\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{t-1}\}$

Comparison

- AM takes long to mix on high-dimensional posterior
- AM gives theoretical convergence guarantees
- Gaussian approximation very fast for inference, but no theoretical guarantees

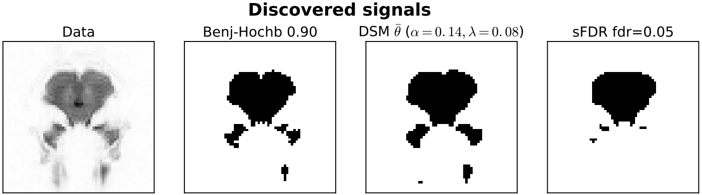
Results

DENOISING



- Denoising removes artifacts from the scull in fMRI
- Shrinking effect visible - $\bar{\mu}$ closer to 0
- Integration in 4096 dimensions

TESTING



- Benjamini-Hochberg doesn't use adjacency (no smoothing)
- sFDR method by Tansey et al. (2014) oversmoothes
- DSM gives good results with theoretical guarantees in increasing dimensionality regime

Conclusion

Conclusion and outlook

- Bayesian method for denoising and testing
  - using structure information
  - fully Bayesian (using integral over posterior)
  - guarantees in increasing dimensionality regime
  - applicable to any data with graph structure
- Inference
  - transform space to make distribution more gaussian
  - adaptive numeric integration schemes
- typical MD estimators reusable for Bayesian inference

Thank you