Large Scale Sparse Inference: Bayesian Sampling Optimization for Magnetic Resonance Imaging

Matthias Seeger

Saarland University and MPI for Informatics, Saarbrücken
Joint work with Hannes Nickisch, Rolf Pohmann, Bernhard Schölkopf
(MPI for Biological Cybernetics, Tübingen)

13 September 2009
Image Reconstruction Pipeline

Ideal Image $u$

Measurement

$y \approx X u$

Design

Data $P(y|u)$

Reconstruction
Whatever images are . . .

they are not Gaussian!
Whatever images are . . .

they are not Gaussian!

- Wavelet transform coefficients super-Gaussian, “sparse”
- Spatial smoothness: Image gradient super-Gaussian, “sparse”
Whatever images are . . .
they are not Gaussian!

- Wavelet transform coefficients super-Gaussian, “sparse”
- Spatial smoothness: Image gradient super-Gaussian, “sparse”
Compressive Sensing

Sparse Linear Model

\[ y = X u + \varepsilon \]

- **y**: Signal
- **X**: Design
- **u**: Noise

Seeger (MMCI)
Compressive Sensing

Sparse Estimation

Fixed $X$, $y$.

$$\hat{u} = \arg \max_u P(y|u) \times \mathcal{P}(u)$$
Compressive Sensing

Sparse Elimination
<table>
<thead>
<tr>
<th>Sparse Elimination</th>
<th>Improved Sensing</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Viking Warrior" /></td>
<td><img src="image" alt="Speed Limit Sign" /></td>
</tr>
</tbody>
</table>

SPEED LIMIT 55
Compressive Sensing

Improved Sensing

SPEED LIMIT 55
Compressive Sensing

Improved Sensing

Seeger (MMCI)
Compressive Sensing

Improved Sensing

YES WE CAN
Compressive Sensing

Improved Sensing

YES WE CAN
Design Score

How informative is $X$ for reconstruction of $u$?
### Sparse Estimation

- How informative is $X$ for reconstruction of $u$?

### Design Score

- 
  - Point estimate not enough
  - Reconstruction uncertainty
    - How good are you?
    - How could you improve?
Sparse Estimation

Design Score

How informative is $X$ for reconstruction of $u$?
- Point estimate not enough
- Reconstruction uncertainty
  - How good are you?
  - How could you improve?
Design Optimization Needs Sparse Inference

Insufficient for Uncertainty

Bayesian Posterior

Uncertainty representation (from same input)

\[ P(u|y) \propto P(y|u)P(u) \]

Sparse Inference
Bayesian Experimental Design

- **Posterior: Uncertainty in reconstruction**
- **Experimental design:** Find poorly determined directions
- **Sequential search with interjacent partial measurements**
Bayesian Experimental Design

- Posterior: **Uncertainty** in reconstruction
- Experimental design:
  - Find poorly determined directions
- Sequential search with interjacent partial measurements
Bayesian Experimental Design

- Posterior: **Uncertainty** in reconstruction
- Experimental design: Find poorly determined directions
- Sequential search with interjacent partial measurements
Sequential Optimization of MRI Trajectories

[Image of MRI trajectories and MRI scans]
Sequential Optimization of MRI Trajectories

(a) Image showing spiral trajectories.
(b) Graph plotting score value against angle.

12.79
Sequential Optimization of MRI Trajectories

Seeger (MMCI)
Sequential Optimization of MRI Trajectories

Seeger (MMCI)  
Bayesian Design Optimization  
13 September 2009
Sounds reasonable. What’s the deal?

1. At huge scale
   \[256^2\text{ complex pixels} \to \mathbb{R}^{131072}\text{ (just one slice).}\]
   Fourier measurements non-local: Part of image won’t do

2. Global covariances
   Marginals not enough: Scores depend on dominating covariances

3. Many times
   Posterior update after each of many sequential steps
Sounds reasonable. What's the deal?
Non-Gaussian Bayesian inference . . .

1. At huge scale
   $256^2$ complex pixels $\rightarrow \mathbb{R}^{131072}$ (just one slice).
   Fourier measurements non-local: Part of image won’t do

2. Global covariances
   Marginals not enough: Scores depend on dominating covariances

3. Many times
   Posterior update after each of many sequential steps
The Challenge

Sounds reasonable. What’s the deal?
Non-Gaussian Bayesian inference . . .

1. At huge scale
   \[256^2 \text{ complex pixels} \rightarrow \mathbb{R}^{131072}\] (just one slice).
   Fourier measurements non-local: Part of image won’t do

2. Global covariances
   Marginals not enough: Scores depend on dominating covariances

3. Many times
   Posterior update after each of many sequential steps
Sounds reasonable. What’s the deal?
Non-Gaussian Bayesian inference . . .

1. At huge scale
   256^2 complex pixels $\rightarrow \mathbb{R}^{131072}$ (just one slice).
   Fourier measurements non-local: Part of image won’t do

2. Global covariances
   Marginals not enough: Scores depend on dominating covariances

3. Many times
   Posterior update after each of many sequential steps
Variational Relaxation

\[ P(u) \propto \prod_{i=1}^{q} t_i(s_i) = e^{-\frac{\tau_w}{\sigma} \|B_w u\|_1} \times e^{-\frac{\tau_{tv}}{\sigma} \|B_{tv} u\|_1}, \quad s = Bu \]

\[ P(y|u) = \mathcal{N}(y|Xu, \sigma^2 I) \]

\[ P(u|y) \propto P(u)P(y|u) \]

wavelet

gradient
Variational Relaxation

\[ P(u) \propto \prod_{i=1}^{q} t_i(s_i) = e^{-\frac{\tau_u}{\sigma} \| B_w u \|_1} \times e^{-\frac{\tau_v}{\sigma} \| B_{tv} u \|_1}, \quad s = Bu \]

\[ P(y|u) = \mathcal{N}(y| Xu, \sigma^2 I) \]

\[ P(u|y) \propto P(u) P(y|u) \]

Log partition function

\[ \log P(y) = \log \int P(y|u) P(u) \, du, \quad u \in \mathbb{R}^{131072} \]

Legendre/Fenchel “Gaussianification”

\[ \log P(y) \geq \log \int P(y|u) Q(u; \gamma) e^{-h(\gamma)/2} \, du, \quad \gamma \in \mathbb{R}_+^{196096} \]

\[ \Rightarrow \text{Maximize lower bound w.r.t. } \gamma > 0 \]
Variational Relaxation

\[ P(u) \propto \prod_{i=1}^{q} t_i(s_i) = e^{-\frac{\tau w}{\sigma} \|B_w u\|_1} \times e^{-\frac{\tau v}{\sigma} \|B_v u\|_1}, \quad s = Bu \]

\[ P(y|u) = \mathcal{N}(y|Xu, \sigma^2 I) \]

\[ P(u|y) \propto P(u)P(y|u) \]

Log partition function

\[ \log P(y) = \log \int P(y|u)P(u) \, du, \quad u \in \mathbb{R}^{131072} \]

Legendre/Fenchel “Gaussianification”

\[ \log P(y) \geq \log \int \frac{P(y|u)Q(u; \gamma)}{Q(u; \gamma)} e^{-h(\gamma)/2} \, du, \quad \gamma \in \mathbb{R}_+^{196096} \]

⇒ Maximize lower bound w.r.t. \( \gamma \succ 0 \)
A Convex Problem

\[ Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag} \gamma, \]

\[
\min_{\gamma > 0} - \log \int P(y|u)Q(u; \gamma) \, du + h(\gamma) / 2
\]

Convex Variational Inference

\[ h(\gamma) \text{ convex iff log potentials } s_j \mapsto \log t_j(s_j) \text{ concave}. \]

Variational relaxation convex iff MAP estimation convex
A Convex Problem

\[ Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag } \gamma, \]
\[
\min_{\gamma > 0} - \log \int P(y|u)Q(u; \gamma) \, du + h(\gamma)/2
\]

Convex Variational Inference

\[ h(\gamma) \text{ convex iff log potentials } s_j \mapsto \log t_j(s_j) \text{ concave.} \]
Variational relaxation \textit{convex} iff MAP estimation \textit{convex}

Convex, but still really large . . .
Gaussian approximate posterior:

\[ Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag } \gamma \]

- Single site updating (coordinate descent):
  - Optimize w.r.t. single \( \gamma_j \) at a time [most previous algorithms]
  - Update requires \( Q(s_j|y) \): Solve system \( A v = c_j \)
  - At least 196096 linear systems \( \Rightarrow \) Out of the question!

- Gradient descent for all of \( \gamma \)
  - Needs all variances \( \text{Var}_Q[s_j|y] \) in every step
- Inference on full images requires new ideas
  - Decoupling of costly parts
Gaussian approximate posterior:

\[ Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag} \gamma \]

- Single site updating (coordinate descent):
  - Optimize w.r.t. single \( \gamma_j \) at a time [most previous algorithms]
    - Update requires \( Q(s_j|y) \): Solve system \( Av = c_j \)
    - At least 196096 linear systems ⇒ Out of the question!

- Gradient descent for all of \( \gamma \)
  ⇒ Needs all variances \( \text{Var}_Q[s_j|y] \) in every step

- Inference on full images requires new ideas
  ⇒ Decoupling of costly parts
Gaussian approximate posterior:

$$Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag } \gamma$$

- Single site updating (coordinate descent):
  - Optimize w.r.t. single $\gamma_j$ at a time [most previous algorithms]
    - Update requires $Q(s_j|y)$: Solve system $A v = c_j$
    - At least 196096 linear systems $\Rightarrow$ Out of the question!

- Gradient descent for all of $\gamma$
  - Needs all variances $\text{Var}_Q[s_j|y]$ in every step

  - Inference on full images requires new ideas
  - Decoupling of costly parts
Gaussian approximate posterior:

\[ Q(u|y) = N(u_*, \sigma^2 A^{-1}), \quad A = X^T X + B^T \Gamma^{-1} B, \quad \Gamma = \text{diag} \gamma \]

- Single site updating (coordinate descent):
  - Optimize w.r.t. single \( \gamma_j \) at a time [most previous algorithms]
  - Update requires \( Q(s_j|y) \): Solve system \( Av = c_j \)
  - At least 196096 linear systems \( \Rightarrow \text{Out of the question!} \)
- Gradient descent for all of \( \gamma \)
  \( \Rightarrow \) Needs all variances \( \text{Var}_Q[s_j|y] \) in every step
- Inference on full images requires new ideas
  \( \Rightarrow \) Decoupling of costly parts
Variational relaxation

1. Bound potentials by Gaussians
   ⇒ **Convex** criterion to minimize

2. Decouple criterion (Fenchel duality)
   ⇒ **Scalable** algorithm
Large Scale Sparse Inference

Variational relaxation

1. Bound potentials by Gaussians
   ⇒ **Convex** criterion to minimize

2. Decouple criterion (Fenchel duality)
   ⇒ **Scalable** algorithm
Large Scale Sparse Inference

Variational relaxation

1. Bound potentials by Gaussians
   ⇒ Convex criterion to minimize

2. Decouple criterion (Fenchel duality)
   ⇒ Scalable algorithm

Double loop (concave-convex, MM) inference algorithm

- Inner loops: IRLS, “smooth sparse estimation”
  ⇒ Gaussian means only
- Once per outer loop (usually 2–6): Gaussian variances
- Score computation needs variances as well
Variational relaxation

1. Bound potentials by Gaussians
   ⇒ **Convex** criterion to minimize

2. Decouple criterion (Fenchel duality)
   ⇒ **Scalable** algorithm

Double loop (concave-convex, MM) inference algorithm

- Inner loops: IRLS, “smooth sparse estimation”
  ⇒ Gaussian **means** only

- Once per outer loop (usually 2–6): Gaussian **variances**

- Score computation needs variances as well
Large Scale Sparse Inference

Variational relaxation

1. Bound potentials by Gaussians
   ⇒ **Convex** criterion to minimize

2. Decouple criterion (Fenchel duality)
   ⇒ **Scalable** algorithm

Double loop (concave-convex, MM) inference algorithm

- Inner loops: IRLS, “smooth sparse estimation”
  ⇒ Gaussian **means** only
- Once per outer loop (usually 2–6): Gaussian **variances**
- Score computation needs variances as well

How to get **variances** at large scale?
**PCA Approximations (Lanczos)**

- Variances $\mathbf{z} = \text{diag}^{-1}(\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^T)$ too much for us
- Project to $k$ dimensions, retain as much of $\mathbf{z}$ (variance) as possible.
  Sounds familiar?
- Scalable $k$-PCA ($k \ll n$): Lanczos algorithm (Matlab `eigs`)
PCA Approximations (Lanczos)

- Variances $z = \text{diag}^{-1}(BA^{-1}B^T)$ too much for us
- Project to $k$ dimensions, retain as much of $z$ (variance) as possible.
  Sounds familiar? Principal Components Analysis of $A^{-1}$!
  
  $$A \approx U \Lambda U^T \Rightarrow z \approx \text{diag}^{-1}(BU \Lambda^{-1} U^T B^T)$$

- Scalable $k$-PCA ($k \ll n$): Lanczos algorithm (Matlab `eigs`)
PCA Approximations (Lanczos)

- Variances $z = \text{diag}^{-1}(BA^{-1}B^T)$ too much for us
- Project to $k$ dimensions, retain as much of $z$ (variance) as possible.
  Sounds familiar? **Principal Components Analysis of $A^{-1}$!**

$$A \approx U \Lambda U^T \quad \Rightarrow \quad z \approx \text{diag}^{-1}(BU \Lambda^{-1}U^T B^T)$$

- Scalable $k$-PCA ($k \ll n$): Lanczos algorithm (Matlab `eigs`)
PCA Approximations (Lanczos)

- Variances $\mathbf{z} = \text{diag}^{-1}(\mathbf{B} \mathbf{A}^{-1} \mathbf{B}^T)$ too much for us
- Project to $k$ dimensions, retain as much of $\mathbf{z}$ (variance) as possible.
  Sounds familiar? Principal Components Analysis of $\mathbf{A}^{-1}$!

$$\mathbf{A} \approx \mathbf{U} \Lambda \mathbf{U}^T \Rightarrow \mathbf{z} \approx \text{diag}^{-1}(\mathbf{B} \mathbf{U} \Lambda^{-1} \mathbf{U}^T \mathbf{B}^T)$$

- Scalable $k$-PCA ($k \ll n$): Lanczos algorithm (Matlab `eigs`)

Variational Inference:
Factorization Assumptions

⊕ General
⊕ Simple to implement
⊕ Covariances gone before you even start

Variational Inference:
PCA Approximations

⊕ Not for all models
⊕ Data-dependent covariance approximation
Optimizing Cartesian Sequences

Bayes Optim.  VD Random  Low Pass
Experimental Results

Sparse reconstruction

Linear reconstruction
Experimental Results: Generalization

$s_0$ 100 120 140 160

$L_2$ reconstruction error

$N_{col}$, Number of columns

$s_0$ 100 120 140 160

$L_2$ reconstruction error

$N_{col}$, Number of columns

Seeger (MMCI)

Bayesian Design Optimization

13 September 2009 16 / 20
Experimental Results: Blow-ups (1/4 Nyquist)
Nothing here is specific to MRI. Could be used for computational photography:
- Blind deconvolution (work in progress)
- Coded aperture design
- Sensor placement in cameras

Beyond images: Sparsity just witnesses structure
- A.k.a. super-Gaussian, power law decay, robust, ...
- Temporal structure: Smooth, occasional jumps
- Combinatorial structures (graphs, networks), feature selection:
  Closest decomposable convex approximation
Nothing here is specific to MRI.
Could be used for computational photography:
- Blind deconvolution (work in progress)
- Coded aperture design
- Sensor placement in cameras

Beyond images: Sparsity just witnesses structure
- A.k.a. super-Gaussian, power law decay, robust, . . .
- Temporal structure: Smooth, occasional jumps
- Combinatorial structures (graphs, networks), feature selection:
  Closest decomposable convex approximation
Conclusions

Modern nonlinear image reconstruction:
Better images through robust low-level prior knowledge

Nonlinear design optimization makes the difference:
- Specific to reconstruction method
- Specific to signal class (natural/MR images)

Nonlinear Bayesian sampling optimization:
General, goal-directed alternative to trial-and-error

Driven by scalable variational inference
- Decoupling to speed up optimization
- Scalability through reductions to standard primitives
Conclusions

- Modern nonlinear image reconstruction:
  Better images through robust low-level prior knowledge
- Nonlinear design optimization makes the difference:
  - Specific to reconstruction method
  - Specific to signal class (natural/MR images)
- Nonlinear Bayesian sampling optimization:
  General, goal-directed alternative to trial-and-error
- Driven by scalable variational inference
  - Decoupling to speed up optimization
  - Scalability through reductions to standard primitives
Conclusions

- Modern nonlinear image reconstruction: Better images through robust low-level prior knowledge
- Nonlinear design optimization makes the difference:
  - Specific to reconstruction method
  - Specific to signal class (natural/MR images)
- Nonlinear Bayesian sampling optimization: General, goal-directed alternative to trial-and-error
  - Driven by scalable variational inference
    - Decoupling to speed up optimization
    - Scalability through reductions to standard primitives
Conclusions

- Modern nonlinear image reconstruction:
  Better images through robust low-level prior knowledge
- Nonlinear design optimization makes the difference:
  - Specific to reconstruction method
  - Specific to signal class (natural/MR images)
- Nonlinear Bayesian sampling optimization:
  General, goal-directed alternative to trial-and-error
- Driven by scalable variational inference
  - Decoupling to speed up optimization
  - Scalability through reductions to standard primitives
Outlook

Seeger, Nickisch, Pohmann, Schölkopf
Bayesian Experimental Design of Magnetic Resonance Imaging Sequences

Seeger, Nickisch, Pohmann, Schölkopf
Optimization of k-Space Trajectories for Compressed Sensing by Bayesian Experimental Design
Magnetic Resonance in Medicine (2009, in print)

Preliminary work. Much to be done:
- Speed-up through parallelization
- Theoretical characterization
- Known issues with non-Cartesian sequences
- Demonstrate benefits in real application

Future work
- 3D sequences with long scan time (Cartesian, radial)
- Design optimization over multiple neighbouring slices
- Design optimization over multiple receiver coils
- Computational photography
- Far away: Support of real-time MRI systems
Preliminary work. Much to be done:
- Speed-up through parallelization
- Theoretical characterization
- Known issues with non-Cartesian sequences
- Demonstrate benefits in real application

Future work
- 3D sequences with long scan time (Cartesian, radial)
- Design optimization over multiple neighbouring slices
- Design optimization over multiple receiver coils
- Computational photography
- Far away: Support of real-time MRI systems