Probabilistic Graphical Models

Inference for Continuous-Variable Models

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

28/11/2011
Outline

1. Introduction
2. Sparsity. Super-Gaussianity
3. Variational Bayesian Inference Relaxations
... as simple as possible, but not simpler.

What do you mean with **simple**?
Sparsity: A Fundamental Concept

... as simple as possible, but not simpler.

What do you mean with simple?

Classical (Gaussian)

- All specified elements
- Use each of them a little
Sparsity: A Fundamental Concept

... as simple as possible, but not simpler.

What do you mean with simple?

Classical (Gaussian)
- All specified elements
- Use each of them a little

Sparsity
- As few elements as possible
- If at all, use them big
Introduction

Sparsity: A Fundamental Concept

Graphical Models

28/11/2011
Introduction

Sparsity: A Fundamental Concept

... as simple as possible, but not simpler.

What do you mean with simple?

Classical (Gaussian)
- All specified elements
- Use each of them very little

Sparsity
- As few elements as possible
- If at all, use them big

Classical linear framework: Shapes the way we think
- Nyquist/Shannon limit. Point spread function
- Aliasing. Ringing. Signal-to-noise ratio
- Linear measurements? Linear reconstruction!
Introduction

Sparsity: A Fundamental Concept

... as simple as possible, but not simpler.

What do you mean with simple?

Classical (Gaussian)
- All specified elements
- Use each of them very little

Sparsity
- As few elements as possible
- If at all, use them big

Sparsity: A concept as basic as classical linear reconstruction
Profound implications for how we (should) think about modelling, reconstruction, acquisition of real-world signals
Many Faces of Sparsity

- Image modelling
  - Processing
  - Reconstruction
  - Acquisition (sampling)
  - Computational neuroscience

Relaxation of combinatorial optimization
Maximally sparse reconstruction
Learning dependency structure
Meinshausen, Bühlmann
Graphical Lasso
Sparse coding
Olshausen, Field
Learning image priors
Many Faces of Sparsity

- Image modelling
  - Processing
  - Reconstruction
  - Acquisition (sampling)
  - Computational neuroscience
- Relaxation of combinatorial optimization
  - Maximally sparse reconstruction
Many Faces of Sparsity

- Image modelling
  - Processing
  - Reconstruction
  - Acquisition (sampling)
  - Computational neuroscience
- Relaxation of combinatorial optimization
  - Maximally sparse reconstruction
- Learning dependency structure
  - Meinshausen, Buehlmann
  - Graphical Lasso
Many Faces of Sparsity

- Image modelling
  - Processing
  - Reconstruction
  - Acquisition (sampling)
  - Computational neuroscience
- Relaxation of combinatorial optimization
  - Maximally sparse reconstruction
- Learning dependency structure
  - Meinshausen, Buehlmann
  - Graphical Lasso
- Sparse coding
  - Olshausen, Field
  - Learning image priors
Image Reconstruction

$y \approx X u$

Design

Data $P(y|u)$

Reconstruction

Measurement

Ideal Image $u$
Reconstruction is Ambiguous
Least Squares Estimation (Linear Model)

\[ u_\ast = \arg\min_u \| y - Xu \|^2 \quad \text{s.t. } \| u \|^2 \text{ small} \]
Least Squares Estimation (Linear Model)

- Simple. Fast. Well understood
- Arbitrary decision (why squares?)
Whatever images are . . .

d they are not Gaussian!

- Small noisy steps
- Gaussian random walker through pixel-land
Whatever images are …

they are not Gaussian!

- Small noisy steps
- Gaussian random walker through pixel-land

- Tiptoeing, edge jumping
- Gaussian won’t do
Whatever images are ... they are not Gaussian!

- Spatial smoothness: Image gradient super-Gaussian, sparse

Capture image properties in prior distribution $P(u)$
Sparsity Priors

Gaussian
\( \propto e^{-\tau |s|^2} \)

Laplace
\( \propto e^{-\tau |s|} \)

Very Sparse
\( \propto e^{-\tau |s|^{0.4}} \)

enforce sparsity
Best of Both Worlds

$$P(u) \propto \prod_{i=1}^{q} t_i(s_i), \quad s = Bu, \quad t_i(s_i) = e^{-\tau_i |s_i|^2 / 2}$$

Gaussian Prior $P(u)$

- Simple. Fast
- Well understood
### Best of Both Worlds

$$P(u) \propto \prod_{i=1}^{q} t_i(s_i), \quad s = Bu, \quad t_i(s_i) = e^{-\tau_i|s_i|}$$

<table>
<thead>
<tr>
<th>Gaussian Prior $P(u)$</th>
<th>Sparsity Prior $P(u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple. Fast</td>
<td>Better prior for real-world signals (images)</td>
</tr>
<tr>
<td>Well understood</td>
<td></td>
</tr>
</tbody>
</table>

### Graphical Models

- Gaussian Prior $P(u)$
- Sparsity Prior $P(u)$
Best of Both Worlds

\[ P(u) \propto \prod_{i=1}^{q} t_i(s_i), \quad s = Bu, \quad t_i(s_i) = e^{-\tau_i|s_i|} \]

### Gaussian Prior \( P(u) \)
- Simple. Fast
- Well understood

### Sparsity Prior \( P(u) \)
- Better prior for real-world signals (images)

#### Latent Gaussian Representations
- Gaussian scale mixtures
  \[ t_i(s_i) = \int_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i)} f_i(\gamma_i) \, d\gamma_i \]
- Super-Gaussian potentials
  \[ t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i)} g_i(\gamma_i) \]
Mixture of Gaussians? $K$-means, EM, ... 

\[ P(X) = \sum_{j=1}^{K} \pi_j N(X|\mu_j, \sigma^2) \]

ti(si) unimodal: Means are not the issue
Gaussian Scale Mixtures

- Mixture of Gaussians? $K$-means, EM, ...  

$$P(X) = \sum_{j=1}^{K} \pi_j N(X | \mu_j, \sigma^2)$$

$t_i(s_i)$ unimodal: Means are not the issue

- What makes $t_i(s_i)$ non-Gaussian:
  - More mass close to origin
  - More mass in tails (far from origin)
  - Less mass at moderate distance

$\Rightarrow$ Mass at different scales
Gaussian Scale Mixtures

- Mixture of Gaussians? $K$-means, EM, ... 

$$P(X) = \sum_{j=1}^{K} \pi_j N(X|\mu_j, \sigma^2)$$

$t_i(s_i)$ unimodal: Means are not the issue

- What makes $t_i(s_i)$ non-Gaussian:
  - More mass close to origin
  - More mass in tails (far from origin)
  - Less mass at moderate distance
  ⇒ Mass at different scales

- Why not mix over the scales?
Gaussian Scale Mixtures

\[ X = \sqrt{\gamma} Y : Y \sim N(0, 1), \gamma \sim f(\gamma)I_{\{\gamma \geq 0\}} \]
$X = \sqrt{\gamma} Y$: $Y \sim N(0, 1)$, $\gamma \sim f(\gamma) I\{\gamma \geq 0\}$

- Many distributions you know are scale mixtures
- Gaussian \([-:-]\).

$$P(X) = N(X|0, \gamma)$$
\[ X = \sqrt{\gamma} Y: \quad Y \sim N(0, 1), \quad \gamma \sim f(\gamma)I_{\{\gamma \geq 0\}} \]

- Many distributions you know are scale mixtures
  - Gaussian [:-)]. Spike and slab

\[
P(X) = \pi N(X|0, \gamma_1) + (1 - \pi) N(X|0, \gamma_2), \quad \gamma_1 \ll \gamma_2
\]
Gaussian Scale Mixtures

\[ X = \sqrt{\gamma} Y: \ Y \sim N(0, 1), \ \gamma \sim f(\gamma)I\{\gamma \geq 0\} \]

- Many distributions you know are scale mixtures
  - Gaussian [:‐‐:]. Spike and slab
  - Exponential power (\(\alpha \leq 2\))

\[ P(X) \propto e^{-\tau|X|^\alpha}, \ \alpha \in (0, 2], \ \tau > 0 \]
Gaussian Scale Mixtures

\[ X = \sqrt{\gamma} Y : Y \sim N(0, 1), \gamma \sim f(\gamma)I_{\gamma \geq 0} \]

- Many distributions you know are scale mixtures
  - Gaussian [:‐‐]. Spike and slab
  - Exponential power (\( \alpha \leq 2 \))
  - Student’s t

\[ P(X) \propto (1 + (\tau/\nu)|X|^2)^{-(\nu+1)/2}, \quad \tau, \nu > 0 \]
\[ X = \sqrt{\gamma} Y: \ Y \sim N(0, 1), \ \gamma \sim f(\gamma) \mathbb{I}_{\{\gamma \geq 0\}} \]

- Many distributions you know are scale mixtures
  - Gaussian \([-:-])\]. Spike and slab
  - Exponential power \((\alpha \leq 2)\)
  - Student’s t

- Duality between \(P(X)\) and \(f(\gamma)\)
- For the Laplace:
  \[
  \frac{\tau}{2} e^{-\tau |s|} = \mathbb{E}[N(|s|; 0, \gamma)], \quad \gamma \sim (\tau^2/2) e^{-(\tau^2/2)\gamma}
  \]
  \[
  = \int_{\gamma \geq 0} N(s|0, \gamma) f(\gamma) \, d\gamma
  \]

West, Biometrika 87
Super-Gaussian Potentials

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma) \]
Super-Gaussian Potentials

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma) \]

- \( t(s) \) even and positive: Let’s look at \( |s|^2 \mapsto 2 \log t(s) \)
- What’s that for a Gaussian \( t(s) = \mathcal{N}(|s||0, \sigma^2) \)?
Super-Gaussian Potential

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma) \]

- \( t(s) \) even and positive: Let's look at \(|s|^2 \mapsto 2 \log t(s)\)
- What's that for a Gaussian \( t(s) = \mathcal{N}(|s||0, \sigma^2) \)?

A linear (affine) function
Sparsity potentials are super-Gaussian

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma) \]

\[ |s|^2 \mapsto 2 \log t(s) \] is convex

- **Affine \rightarrow** convex:
  - Shift mass to center and tails
Super-Gaussian Potentials

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma) \]

Sparsity potentials are super-Gaussian

\[ |s|^2 \mapsto 2 \log t(s) \quad \text{is convex} \]

- Affine \rightarrow\text{ convex:}
  Shift mass to center and tails
- Scale mixtures are super-Gaussian

Palmer et al., NIPS 2005
Scale Mixtures are Super-Gaussian

Gaussian scale mixture: \[ t(s) = \int_{\geq 0} e^{-|s|^2/(2\gamma)} f(\gamma) \, d\gamma \]

- \( t(s) \) even and positive:
  \[ x := |s|^2 \quad \Rightarrow \quad t(s) = e^{g(x)} \]
Gaussian scale mixture: \( t(s) = \int_{\geq 0} e^{-|s|^2/(2\gamma)} f(\gamma) \, d\gamma \)

- \( t(s) \) even and positive:
  \( x := |s|^2 \Rightarrow t(s) = e^{g(x)} \)
- Super-Gaussian? \(|s|^2 \mapsto 2 \log t(s)\) is convex.
  Show that \( g(x) \) is convex
Scale Mixtures are Super-Gaussian

Gaussian scale mixture: \( t(s) = \int_{\geq 0} e^{-|s|^2/(2\gamma)} f(\gamma) \, d\gamma \)

- \( t(s) \) even and positive:
  \( x := |s|^2 \Rightarrow t(s) = e^{g(x)} \)

- Super-Gaussian? \( |s|^2 \mapsto 2 \log t(s) \) is convex.
  Show that \( g(x) \) is convex

- Log-convexity: Closed under summation

\( \psi(x, \gamma) \) convex \( \forall \gamma \in C \) \( \Rightarrow \) \( \log \int_C e^{\psi(x, \gamma)} \, d\gamma \) convex

Boyd, Vandenberghe, 2002
Scale Mixtures are Super-Gaussian

Gaussian scale mixture: \[ t(s) = \int_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} f(\gamma) \, d\gamma \]

- \( t(s) \) even and positive:
  \[ x := |s|^2 \Rightarrow t(s) = e^{g(x)} \]
- Super-Gaussian? \(|s|^2 \mapsto 2 \log t(s)\) is convex.
  Show that \( g(x) \) is convex
- Log-convexity: Closed under summation

\[ \psi(x, \gamma) \mbox{ convex } \forall \gamma \in C \Rightarrow \log \int_C e^{\psi(x, \gamma)} \, d\gamma \mbox{ convex} \]

- Apply to \( g(x) \):
  \[ g(x) = \log \int_{\gamma \geq 0} e^{-x/(2\gamma)} f(\gamma) \, d\gamma = \log \int_{\gamma \geq 0} e^{-x/(2\gamma)+\log f(\gamma)} \, d\gamma \]
Group Sparsity

\[ t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i)} g_i(\gamma_i) \]

- \( t_i(s_i) \) depends on absolute value \(|s_i|\) only
- Can just as well plug in vector norm \(\|s_i\|\)
Group Sparsity

\[ t(s_i) = \max_{\gamma_i \geq 0} \exp\left( -\frac{\|s_i\|^2}{2\gamma_i} \right) g_i(\gamma_i) \]

- \( t_i(s_i) \) depends on absolute value \(|s_i|\) only
- Can just as well plug in vector norm \( \|s_i\| \)
- Useful for complex values: \(|s_i| = \|(\Re s_i, \Im s_i)^T\|\)
Group Sparsity

\[ t(s_i) = \max_{\gamma_i \geq 0} e^{-\|s_i\|^2/(2\gamma_i)} g_i(\gamma_i) \]

- \( t_i(s_i) \) depends on absolute value \( |s_i| \) only
- Can just as well plug in vector norm \( \|s_i\| \)
- Useful for complex values: \( |s_i| = \|(\Re s_i, \Im s_i)^T\| \)
- Useful to structure sparsity: Joint penalization of groups
  \( \Rightarrow \ell_1 - \ell_2 \) norms, group Lasso, and all that . . .
Group Sparsity

\[ t(s_i) = \max_{\gamma_i \geq 0} e^{-\|s_i\|^2/(2\gamma_i)} g_i(\gamma_i) \]

- \( t_i(s_i) \) depends on absolute value \(|s_i|\) only
- Can just as well plug in vector norm \(\|s_i\|\)
- Useful for complex values: \(|s_i| = \|(\Re s_i, \Im s_i)^T\|\)
- Useful to "structure" sparsity: Joint penalization of groups
  \( \Rightarrow \ell_1 - \ell_2 \) norms, group Lasso, and all that . . .
- Latent Gaussian representations: Just parameter tying
  \[ e^{-\|s_i\|^2/(2\gamma_i)} \propto N(s_i|0, \gamma_i 1) \]
Group Sparsity

\[ t(s_i) = \max_{\gamma_i \geq 0} e^{-\|s_i\|^2/(2\gamma_i)} g_i(\gamma_i) \]

- \( t_i(s_i) \) depends on absolute value \(|s_i|\) only
- Can just as well plug in vector norm \( \|s_i\| \)
- Useful for complex values: \(|s_i| = \|(\Re s_i, \Im s_i)^T\|
- Useful to structure sparsity: Joint penalization of groups
  \( \Rightarrow \ell_1 - \ell_2 \) norms, group Lasso, and all that...
- Latent Gaussian representations: Just parameter tying
  \( e^{-\|s_i\|^2/(2\gamma_i)} \propto N(s_i|0, \gamma_i 1) \)
Sparse $s$

- Many/most $s_i = 0$
Sparse $s$
- Many/most $s_i = 0$

Super-Gaussian $s$
- Super-Gaussian statistics
- Soft sparsity, statistical sparsity, power law decay, ...
### Sparse $s$
- Many/most $s_i = 0$

### Super-Gaussian $s$
- Super-Gaussian statistics
- Soft sparsity, statistical sparsity, power law decay, …

$P(s)$ super-Gaussian: $s \sim P(s)$ no zeros in general (only if $P(s)$ degenerate)
Real-world signals are not Gaussian. Gaussian assumptions made for convenience only.

Super-Gaussian distributions:
Trade-off between realistic and tractable/simple

Latent Gaussian representations:
  - Gaussian scale mixtures
  - Super-Gaussian potentials

Group potentials:
Simple way to structure sparsity

“Sparse” may mean super-Gaussian
Variational Approximations

\[ P(u|y) = Z^{-1} P(y|u) \prod_i t_i(s_i), \quad Z = \int P(y|u) \prod_i t_i(s_i) \, du \]

- Bayesian integration over \( P(u|y) \) intractable
Variational Approximations

\[ P(u|y) = Z^{-1} P(y|u) \prod_i t_i(s_i), \quad Z = \int P(y|u) \prod_i t_i(s_i) \, du \]

- Bayesian integration over \( P(u|y) \) intractable
- Integration tractable for Gaussians \( Q(u|y) \)
  \[ \Rightarrow \text{Approximate } P(u|y) \text{ by } Q(u|y)! \]
Variational Approximations

\[ P(u|y) = Z^{-1} P(y|u) \prod_i t_i(s_i), \quad Z = \int P(y|u) \prod_i t_i(s_i) \, du \]

- Bayesian integration over \( P(u|y) \) intractable
- Integration tractable for Gaussians \( Q(u|y) \)
  \( \Rightarrow \) Approximate \( P(u|y) \) by \( Q(u|y) \)!

Variational approximation

Apply variational principle to fit master function \( \log Z \)
Variational Approximations

\[ P(u|y) = Z^{-1} P(y|u) \prod_i t_i(s_i), \quad Z = \int P(y|u) \prod_i t_i(s_i) \, du \]

- Bayesian integration over \( P(u|y) \) intractable
- Integration tractable for \( \textbf{Gaussians} \) \( Q(u|y) \)
  \( \Rightarrow \) Approximate \( P(u|y) \) by \( Q(u|y) \)!

Variational approximation

Apply variational principle to fit master function \( \log Z \)

- Super-Gaussian bounding
- Expectation propagation
- Variational mean field Bayes [not here]
- Gaussian KL minimization [not here]
Super-Gaussian Potentials

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} e^{-h(\gamma)/2} \]

- **t(s) even and positive**: Let’s look at $|s|^2 \mapsto 2 \log t(s)$
- **What’s that for a Gaussian $t(s) = \mathcal{N}(|s|0, \sigma^2)$?**
  - A linear (affine) function

![Graphs showing the super-Gaussian potentials](image)
Super-Gaussian Potentials

\[ t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} e^{-h(\gamma)/2} \]

Sparsity potentials are super-Gaussian

\[ |s|^2 \mapsto 2 \log t(s) \] is convex

\[ t(s) = \max_{\gamma \geq 0} \ldots \text{Why?} \]
Convex (Fenchel) Duality

Super-Gaussian:
\[ t(s) \text{ even, } |s|^2 \rightarrow \log t(s) \text{ convex.} \]

Convex function: Maximum of its affine lower bounds
Super-Gaussian function: Maximum of its Gaussian lower bounds
Supergaussian:
\[ t(s) \text{ even, } |s|^2 \mapsto \log t(s) \text{ convex.} \]

\[
\begin{align*}
  f(x) &= \max_{\pi} x\pi - f^*(\pi) \\
  t(s) &= \max_{\gamma} e^{(-|s|^2/\gamma - h(\gamma))/2}
\end{align*}
\]

\[
\begin{align*}
  f^*(\pi) &= \max_x \pi x - f(x) \\
  h(\gamma) &= \max_s -|s|^2/\gamma - 2 \log t(s)
\end{align*}
\]
Super-Gaussian: 
\( t(s) \) even, \( |s|^2 \mapsto \log t(s) \) convex.

\[
f(x) = \max_{\pi} x_{\pi} - f^*(\pi)
\]
\[
t(s) = \max_{\gamma} e^{-|s|^2/\gamma - h(\gamma)}/2
\]
\[
f^*(\pi) = \max_{x} \pi x - f(x)
\]
\[
h(\gamma) = \max_{s} -|s|^2/\gamma - 2 \log t(s)
\]
Super-Gaussian:
\[ t(s) \text{ even, } |s|^2 \leftrightarrow \log t(s) \text{ convex.} \]

\[ f(x) = \max_{\pi} x_\pi - f^*(\pi) \]

\[ t(s) = \max_{\gamma} e^{-|s|^2/\gamma - h(\gamma)}/2 \]

\[ f^*(\pi) = \max_x \pi x - f(x) \]

\[ h(\gamma) = \max_s -|s|^2/\gamma - 2 \log t(s) \]
Super-Gaussian:
\( t(s) \) even, \( |s|^2 \mapsto \log t(s) \) convex.

\[
f(x) = \max_{\pi} x_{\pi} - f^*(\pi)
\]
\[
t(s) = \max_{\gamma} e^{-|s|^2/\gamma - h(\gamma)}/2
\]
\[
f^*(\pi) = \max_x \pi x - f(x)
\]
\[
h(\gamma) = \max_s -|s|^2/\gamma - 2 \log t(s)
\]
Super-Gaussian:

\( t(s) \) even, \( |s|^2 \mapsto \log t(s) \) convex.

\[
\begin{align*}
  f(x) &= \max_{\pi} x_\pi - f^*(\pi) \\
  t(s) &= \max_\gamma e^{-|s|^2/\gamma - h(\gamma)}/2
\end{align*}
\]

\[
\begin{align*}
  f^*(\pi) &= \max_x \pi x - f(x) \\
  h(\gamma) &= \max_s -|s|^2/\gamma - 2 \log t(s)
\end{align*}
\]
Super-Gaussian:
$t(s)$ even, $|s|^2 \mapsto \log t(s)$ convex.

\[ f(x) = \max_{\pi} x_{\pi} - f^*(\pi) \]
\[ t(s) = \max_{\gamma} e^{-|s|^2/\gamma - h(\gamma))/2} \]
\[ f^*(\pi) = \max_{x} \pi x - f(x) \]
\[ h(\gamma) = \max_{s} -|s|^2/\gamma - 2 \log t(s) \]
Super-Gaussian Potentials

\[ P(u|y) = \frac{P(y|u) \times P(u)}{P(y)} \]

Sparsity potentials are super-Gaussian

\[ |s_i|^2 \mapsto 2 \log t_i(s_i) \quad \text{is convex} \]

Convex (Fenchel) duality

\[ 2 \log t_i(s_i) = \max_{\pi_i} |s_i|^2 \pi_i - f^*(\pi_i) \]
**Super-Gaussian Potentials**

\[
P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}
\]

Sparsity potentials are **super-Gaussian**

\[
t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i) - h_i(\gamma_i)/2}
\]
Variational Bayesian Inference Relaxations

Super-Gaussian Bounding

\[ P(u|y) = \frac{P(y|u) \times P(u)}{P(y)} \]

Sparsity potentials are super-Gaussian

\[ t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i) - h_i(\gamma_i)/2} \]

\[ h(\gamma) := \sum_i h_i(\gamma_i), \quad \Gamma = \text{diag} \gamma \]
Super-Gaussian Bounding

Exact representation

\[
\log Z = \log \int P(y|u) \max_{\gamma} e^{-\left(s^H \Gamma^{-1} s + h(\gamma)\right)/2} \, du
\]

\[
P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}
\]

\[
t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i) - h(\gamma_i)/2}
\]
Variational Bayesian Inference Relaxations

Super-Gaussian Bounding

\[
P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}
\]

Lower bound

\[
\log Z = \log \int P(y|u) \max_{\gamma} e^{-(s^H \Gamma^{-1} s + h(\gamma))/2} \, du
\]

\[
\geq \max_{\gamma} \log \int P(y|u) e^{-(s^H \Gamma^{-1} s + h(\gamma))/2} \, du
\]

\[
t_i(s_i) = \max_{\gamma \geq 0} e^{-|s_i|^2/(2\gamma) - h(\gamma)/2}
\]
Super-Gaussian Bounding

\[ P(u|y) = \frac{P(y|u) \times P(u)}{P(y)} \]

Lower bound

\[
\log Z \geq \max_{\gamma} \log \int P(y|u) e^{-(s^H \Gamma^{-1} s + h(\gamma))}/2 \, du
\]

\[ = \max_{\gamma} \log Z_Q(\gamma) - h(\gamma)/2 \]

Gaussian approximation

\[ Q(u|y) = Z_Q^{-1} P(y|u) e^{-s^H \Gamma^{-1} s/2}, \quad s = B u \]
Variational Bayesian Inference Relaxations

Super-Gaussian Bounding

\[ P(u|y) = \frac{P(y|u) \times P(u)}{P(y)} \]

Variational problem: \( Q(u|y) \approx P(u|y) \)

\[ \min_{\gamma} \{ \phi(\gamma) = -2 \log Z_Q + h(\gamma) \} \]

Gaussian approximation

\[ Q(u|y) = Z_Q^{-1} P(y|u) e^{-s^H\Gamma^{-1}s/2}, \quad s = Bu, \]

\[ Z_Q = \int P(y|u) e^{-s^H\Gamma^{-1}s/2} \, du \]

\[ t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i) - h_i(\gamma_i)/2} \]
Super-Gaussian Bounding

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i \geq 0} \ldots$
- Auxiliary variables $\gamma \geq 0$

\[
P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}
\]
Variational Bayesian Inference Relaxations

Super-Gaussian Bounding

\[ P(u|y) = \frac{P(y|u) \times P(u)}{P(y)} \]

What did we do?

- Start with tight single potential bounds: \( t_i(s_i) = \max_{\gamma_i \geq 0} \ldots \Rightarrow \text{Auxiliary variables } \gamma \succeq 0 \)
- Plug into target function \( \log Z \). Interchange \( \int \ldots du \leftrightarrow \max_{\gamma} \Rightarrow \text{Global lower bound on } \log Z \)
Super-Gaussian Bounding

\[
P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}
\]

What did we do?

- Start with tight single potential bounds: \( t_i(s_i) = \max_{\gamma_i \geq 0} \ldots \)
  \( \Rightarrow \) Auxiliary variables \( \gamma \succeq 0 \)

- Plug into target function \( \log Z \). Interchange \( \int \ldots d\mathbf{u} \leftrightarrow \max_{\gamma} \)
  \( \Rightarrow \) Global lower bound on \( \log Z \)

- Lower bounds are log partition functions of Gaussians \( Q(u|y) \)
  \( \Rightarrow \) Approximation family \( Q = \{Q(u|y)\} \)
Super-Gaussian Bounding

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i \geq 0} \ldots$
  - Auxiliary variables $\gamma \succeq 0$
- Plug into target function $\log Z$. Interchange $\int \ldots du \leftrightarrow \max_{\gamma}$
  - Global lower bound on $\log Z$
- Lower bounds are log partition functions of Gaussians $Q(u|y)$
  - Approximation family $\mathcal{Q} = \{Q(u|y)\}$
- Divergence $Q(u|y) \leftrightarrow P(u|y)$? Maximize lower bound!
  - $\phi(\gamma) = -2 \log Z_Q + h(\gamma)$

$$P(u|y) = \frac{P(y|u) \times P(u)}{P(y)}$$
MAP Estimation

\[
\max_u \log \frac{P(u|y)Z}{\gamma} = \max_u \log N(y| Xu, \sigma^2 I) \max_\gamma e^{-(s^T \Gamma^{-1} s + h(\gamma))/2}
\]

Bayesian Inference

\[
\log Z = \log \int N(y| Xu, \sigma^2 I) \max_\gamma e^{-(s^T \Gamma^{-1} s + h(\gamma))/2} du
\]
Coordinate Descent Algorithm

- Simple algorithm: Update single variables $\gamma_j$

repeat
  for $j \in \{1, \ldots, q\}$ do
    Update $\gamma_j$, based on marginal $Q(s_j|y)$
    Gaussian propagation of pseudo-evidence change
  end for
  Refresh representation
until convergence
Coordinate Descent Algorithm

- Simple algorithm: Update single variables $\gamma_j$

repeat
for $j \in \{1, \ldots, q\}$ do
  Update $\gamma_j$, based on marginal $Q(s_j|y)$
  Gaussian propagation of pseudo-evidence change
end for
Refresh representation
until convergence

- Representation of $Q(u|y)$: Backbone for Gaussian propagation.
  Moderate size problems: Cholesky representation

Seeger, JMLR 2008
Simple algorithm: Update single variables $\gamma_j$

repeat
  for $j \in \{1, \ldots, q\}$ do
    Update $\gamma_j$, based on marginal $Q(s_j|y)$
    Gaussian propagation of pseudo-evidence change
  end for
end repeat

Refresh representation

until convergence

Representation of $Q(u|y)$: Backbone for Gaussian propagation.
Moderate size problems: Cholesky representation

Large scale problems?
This algorithm is not scalable. Can do much better . . .

Seeger, JMLR 2008