Probabilistic Graphical Models

Lecture 12: Continuous-Variable Models

Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

25/11/2011

Outline

The World Is Not Discrete

- Approximate Bayesian inference?
  By far most activity for discrete variable models
  - Clean language of combinatorics on graphs. No numerical issues
  - Some relaxations are very fast (graph cuts)
  - Everything can be gridded, discretized, quantized in principle

The World Is Not Discrete

- Approximate Bayesian inference?
  By far most activity for discrete variable models
  - Clean language of combinatorics on graphs. No numerical issues
  - Some relaxations are very fast (graph cuts)
  - Everything can be gridded, discretized, quantized in principle
- Viewed at useful scales, many problems are continuous.
  Quantization destroys structure useful for efficient computation.
  Trajectory of projectile? Planetary motion? Natural image?

Continuous	Discrete
Newton mechanics	Quantum mechanics
Differential equations	Discretized finite differences
Integrals	Ever larger sums

A Different World

Continuous inference needs more

Discrete inference

- Boils down to size of sums, hardness of graph (treewidth)
- True marginals easy to represent, “just” hard to compute

Continuous inference

- Distribution representation at least as important as graph
- Even local computations (often) not exact ($\int$ for $\sum$)
- Numerical errors have to be controlled
- No ground truth even for smallish problems.
  Local true marginals cannot be represented exactly

A Different World

Continuous inference: More flexibility, sometimes simpler

Discrete inference

- Most approaches today:
  - Recursive hyper-tree computations (smaller $\sum$)
  - Tractable combinatorial graph algorithms
- Smoothness? Non-local search directions? Global correlations?

Continuous inference

- Continuous optimization: Host of different approaches
- Global information from local computations (gradient, Hessian)
- Global correlation information over all variables (PCA)
- Continuous scientific computing well developed
  - Least squares estimation
  - Signal processing (Fourier transforms, ...)
  - PDEs

Sparse Linear Model

- Continuous variable inference:
  - Many different models for many different applications
  - Many (more or less) generic concepts
  - This lecture: Little time remaining . . .

Sparse Linear Model

- Continuous variable inference:
  - Many different models for many different applications
  - Many (more or less) generic concepts
  - This lecture: Little time remaining . . .
- Fortunately: One model
  - Surprisingly many applications (and growing)
  - Surprisingly many generic concepts can be demonstrated

Sparse Linear Model

- Continuous variable inference:
  - Many different models for many different applications
  - Many (more or less) generic concepts
  - This lecture: Little time remaining . . .
- Fortunately: One model
  - Surprisingly many applications (and growing)
  - Surprisingly many generic concepts can be demonstrated
- Workhorse for much of remaining lectures:
  Sparse linear model
- Some important points we will skip
  - Multimodality of posteriors
  - Models with continuous and discrete variables

Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

Uniform shrinkage:	All weights are smallish $\Longrightarrow$ Gaussian F6
Selective shrinkage:	Most weights are tiny, but some can be tall $\Longrightarrow$ Sparsity

Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

Uniform shrinkage:	All weights are smallish $\Longrightarrow$ Gaussian
Selective shrinkage:	Most weights are tiny, but some can be tall $\Longrightarrow$ Sparsity

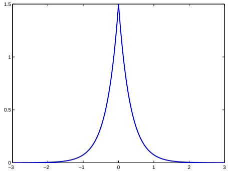
- How to implement sparsity?
  - Combinatorial search [:-(]

Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

  Uniform shrinkage: All weights are smallish $\Longrightarrow$ Gaussian
  Selective shrinkage: Most weights are tiny, but some can be tall $\Longrightarrow$ Sparsity

- How to implement sparsity?
  - Combinatorial search [:-(]
  - Super-Gaussian distributions [:-)]

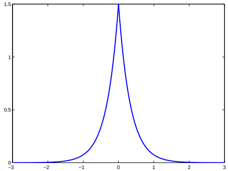


Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

  Uniform shrinkage: All weights are smallish $\Longrightarrow$ Gaussian
  Selective shrinkage: Most weights are tiny, but some can be tall $\Longrightarrow$ Sparsity

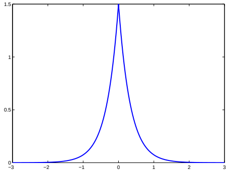
- How to implement sparsity?
  - Combinatorial search [:-(]
  - Super-Gaussian distributions [:-)]
- We know sparsity for efficiency (SVMs, sparse matrices, ...)
  Here: Sparsity captures signals better
  (would be faster without)



Sparsity

- Statistics needs regularization: notions of simplicity
- Linear functions are simple if their weights are small

  Uniform shrinkage: All weights are smallish $\Longrightarrow$ Gaussian
  Selective shrinkage: Most weights are tiny, but some can be tall $\Longrightarrow$ Sparsity



- How to implement sparsity?
  - Combinatorial search [:-(]
  - Super-Gaussian distributions [:-)]
- We know sparsity for efficiency (SVMs, sparse matrices, . . .)
  Here: Sparsity captures signals better
  (would be faster without)
- We know sparse estimation (Lasso, basis pursuit, . . .)
  Here: Bayesian inference with sparsity distributions

Sparse Linear Model

Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

Sparse Linear Model

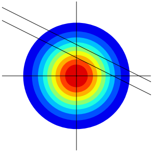
Linear Model

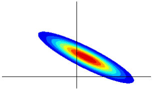
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

Gaussian Prior $P(\boldsymbol{u})$

- Renders inference simple
  Chosen often only for that
- Does not enforce $\boldsymbol{b}_j^T \boldsymbol{u} \approx 0$ strongly
- Does not allow any large $\boldsymbol{b}_j^T \boldsymbol{u}$





Sparse Linear Model

Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

- Whatever images are, they are not Gaussian!

Sparse Linear Model

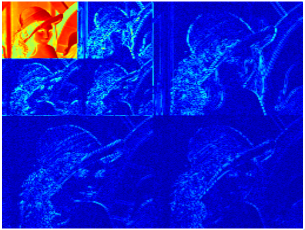
Linear Model

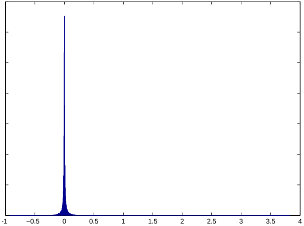
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

- Whatever images are, they are not Gaussian!
  - Wavelet transform coefficients super-Gaussian

Simoncelli, SPIE 99





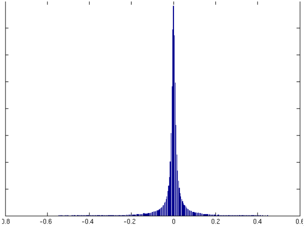
Sparse Linear Model

Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

- Whatever images are, they are not Gaussian!
  - Wavelet transform coefficients super-Gaussian — Simoncelli, SPIE 99
  - Spatial smoothness: Image gradient super-Gaussian



Sparse Linear Model

Linear Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

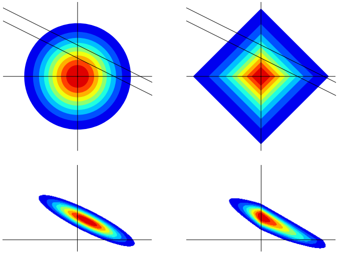
$\boldsymbol{X}$	Design matrix
$\boldsymbol{u} \in \mathbb{R}^n$	Latent variables
$\boldsymbol{y} \in \mathbb{R}^m$	Responses

Laplace (Sparsity) Prior $P(\boldsymbol{u})$

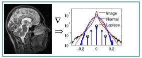
$\boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$ linear statistics

- Allows few s_j to be large
- Forces most $s_j \approx 0$

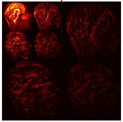
$$P(s_j) = \frac{\tau}{2} e^{-\tau |s_j|}, \ \tau > 0$$



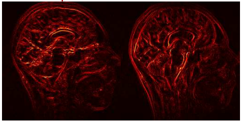
Sparse Linear Model



$$P(\mathbf{u}) \propto \prod_{i=1}^{q} t_i(s_i) = e^{-\tau_w \|\mathbf{B}_w \mathbf{u}\|_1} \times e^{-\tau_{tv} \|\mathbf{B}_{tv} \mathbf{u}\|_1}, \qquad \mathbf{s} = \mathbf{B}\mathbf{u}$$

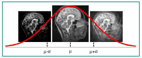


wavelet



gradient

$$P(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{u}, \sigma^2\mathbf{I})$$



$$P(\mathbf{u}|\mathbf{y}) \propto P(\mathbf{u})P(\mathbf{y}|\mathbf{u})$$

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i)$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i)$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?
  If all $t_i(s_i)$ were Gaussian . . .

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?
  If all $t_i(s_i)$ were Gaussian ...
- Approximate $P(\boldsymbol{u}|\boldsymbol{y})$ by Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{b}, \gamma)$

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?
  If all $t_i(s_i)$ were Gaussian ...
- Approximate $P(\boldsymbol{u}|\boldsymbol{y})$ by Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{b}, \gamma)$
- Bad idea: $t_j(s_j)$ does not look like any Gaussian!

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?
  If all $t_i(s_i)$ were Gaussian ...
- Approximate $P(\boldsymbol{u}|\boldsymbol{y})$ by Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{b}, \gamma)$
- Bad idea: $t_j(s_j)$ does not look like any Gaussian!
- Replace $t_j(s_j) \to e^{b_i s_i - s_i^2/(2\gamma_i)}$,
  then adjust $\boldsymbol{b}$, γ to fit joint posterior, not single $t_j(s_j)$!
  $\Rightarrow$ Done by most algorithms: Good idea

Gaussian Approximations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable. Why?
  If all $t_i(s_i)$ were Gaussian ...
- Approximate $P(\boldsymbol{u}|\boldsymbol{y})$ by Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}; \boldsymbol{b}, \gamma)$
- Bad idea: $t_j(s_j)$ does not look like any Gaussian!
- Replace $t_j(s_j) \to e^{b_i s_i - s_i^2/(2\gamma_i)}$,
  then adjust $\boldsymbol{b}$, γ to fit joint posterior, not single $t_j(s_j)$!
  $\Rightarrow$ Done by most algorithms: Good idea
- Criterion to minimize? Divergence $P(\boldsymbol{u}|\boldsymbol{y}) \leftrightarrow Q(\boldsymbol{u}|\boldsymbol{y})$?
  $\Rightarrow$ Closer look at sparsity potentials $t_i(s_i)$

Selective Shrinkage and γ

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

Uniform shrinkage $\Leftrightarrow$ Gaussian prior
Selective shrinkage $\Leftrightarrow$ Sparsity prior (super-Gaussian)

$Q(\boldsymbol{u}|\boldsymbol{y})$ is Gaussian. Where is selective shrinkage?

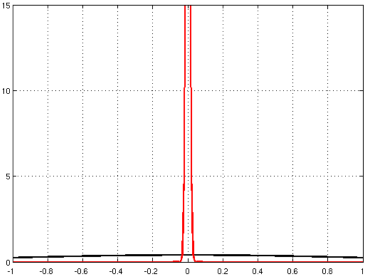
Selective Shrinkage and γ

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

Uniform shrinkage $\Leftrightarrow$ Gaussian prior
Selective shrinkage $\Leftrightarrow$ Sparsity prior (super-Gaussian)

$Q(\boldsymbol{u}|\boldsymbol{y})$ is Gaussian. Where is selective shrinkage?

- The γ_j allow for selective shrinkage
  - γ_j small: $|s_j|$ constrained to be small



Selective Shrinkage and γ

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

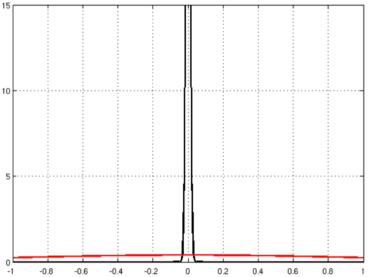
Uniform shrinkage $\Leftrightarrow$ Gaussian prior
Selective shrinkage $\Leftrightarrow$ Sparsity prior (super-Gaussian)

$Q(\boldsymbol{u}|\boldsymbol{y})$ is Gaussian. Where is selective shrinkage?

- The γ_j allow for selective shrinkage
    - γ_j small: $|s_j|$ constrained to be small
    - γ_j large: s_j rather unconstrained

  Your exercise sheet:

$$\mathrm{Var}_Q[s_j|\boldsymbol{y}] \leq \gamma_j$$



Selective Shrinkage and γ

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

Uniform shrinkage $\Leftrightarrow$ Gaussian prior
Selective shrinkage $\Leftrightarrow$ Sparsity prior (super-Gaussian)

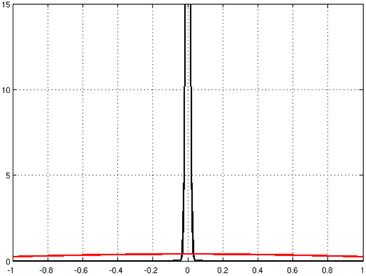
$Q(\boldsymbol{u}|\boldsymbol{y})$ is Gaussian. Where is selective shrinkage?

- The γ_j allow for selective shrinkage
  - γ_j small: $|s_j|$ constrained to be small
  - γ_j large: s_j rather unconstrained

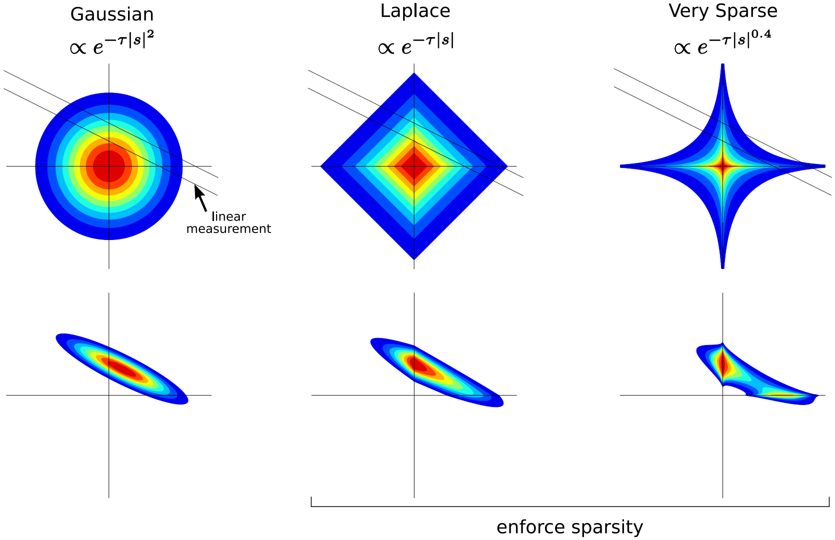
  Your exercise sheet:

$$\mathrm{Var}_Q[s_j|\boldsymbol{y}] \leq \gamma_j$$

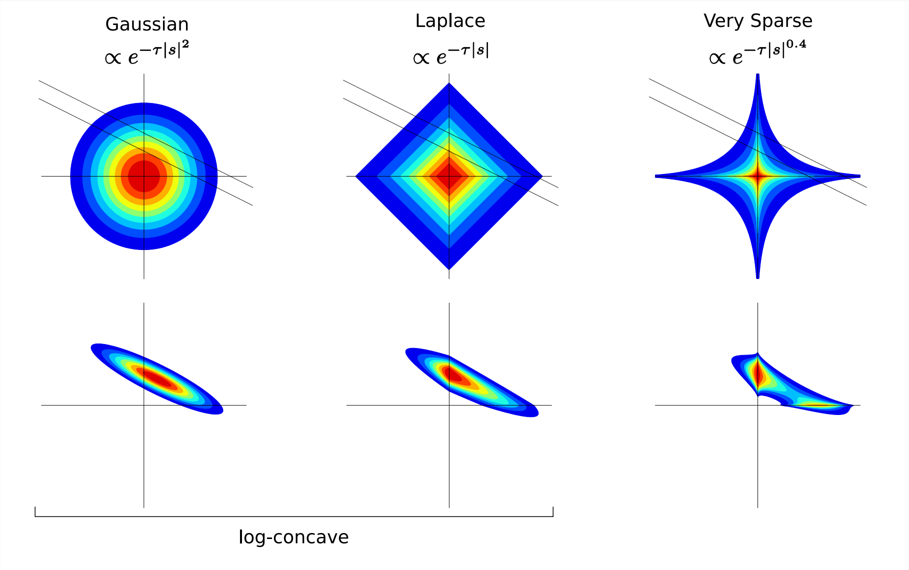
- Variational inference relaxation: Update γ_j to implement selective shrinkage



Sparsity Priors



Sparsity Priors



Gaussian-Form Representations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Statistically, it's crucial that $t_i(s_i)$ are not Gaussian
- Computationally, we can only deal with Gaussian inference

What to do when you're stuck?

Gaussian-Form Representations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1}P(\boldsymbol{y}|\boldsymbol{u})\prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u})\prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Statistically, it's crucial that $t_i(s_i)$ are not Gaussian
- Computationally, we can only deal with Gaussian inference

What to do when you're stuck? Add new variables!

Represent $t_i(s_i)$ as latent Gaussian

Gaussian-Form Representations

$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Statistically, it's crucial that $t_i(s_i)$ are not Gaussian
- Computationally, we can only deal with Gaussian inference

What to do when you're stuck? Add new variables!

Represent $t_i(s_i)$ as latent Gaussian

- Gaussian scale mixtures $\quad t_i(s_i) = \int_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i)} f_i(\gamma_i)\, d\gamma_i$
- Super-Gaussian potentials $\quad t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i)} g_i(\gamma_i)$

F8

Gaussian Scale Mixtures

- Mixture of Gaussians: Typically over means

$$P(X) = \sum_{j=1}^{k} \pi_k N(X|\mu_k, \sigma^2)$$

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

Gaussian Scale Mixtures

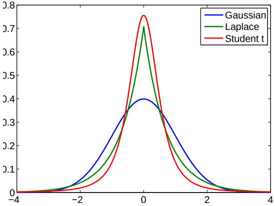
- Mixture of Gaussians: Typically over means

$$P(X) = \sum_{j=1}^{k} \pi_k N(X|\mu_k, \sigma^2)$$

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

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

$\Rightarrow$ Mass at different scales



Gaussian Scale Mixtures

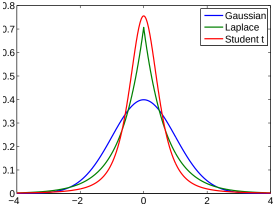
- Mixture of Gaussians: Typically over means

$$P(X) = \sum_{j=1}^{k} \pi_k N(X|\mu_k, \sigma^2)$$

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

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

  $\Rightarrow$ Mass at different scales
- Why not mix over the scales?

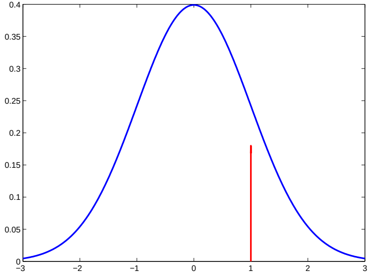


Gaussian Scale Mixtures

- $X = \rho Y$, $Y \sim N(0, 1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$

Gaussian Scale Mixtures

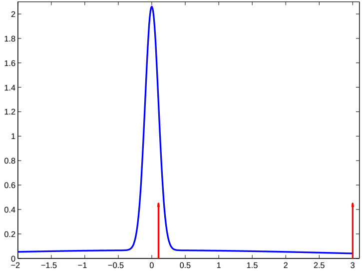
- $X = \rho Y$, $Y \sim N(0,1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$
- Many distributions you know are scale mixtures
  - Gaussian [:-)].



$$P(X) = N(X|0, \rho^2)$$

Gaussian Scale Mixtures

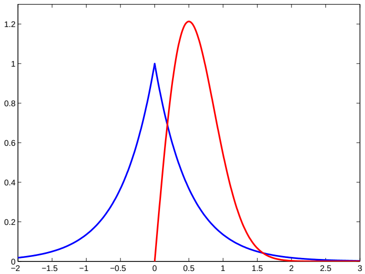
- $X = \rho Y$, $Y \sim N(0,1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$
- Many distributions you know are scale mixtures
  - Gaussian [:-)]. Spike and slab



$$P(X) = \pi N(X|0,\rho_1^2) + (1-\pi)N(X|0,\rho_2^2), \quad \rho_1 \ll \rho_2$$

Gaussian Scale Mixtures

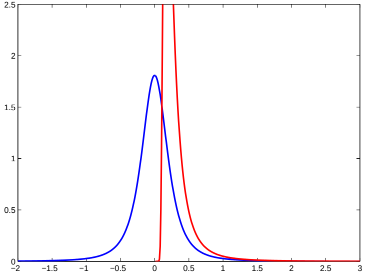
- $X = \rho Y$, $Y \sim N(0,1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$
- Many distributions you know are scale mixtures
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)



$$P(X) \propto e^{-\tau|X|^{\alpha}}, \quad \alpha \in (0,2],\ \tau > 0$$

Gaussian Scale Mixtures

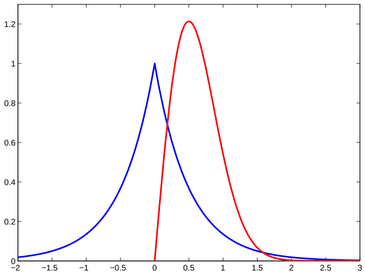
- $X = \rho Y$, $Y \sim N(0,1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$
- Many distributions you know are scale mixtures
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)
  - Student's t



$$P(X) \propto (1 + (\tau/\nu)s^2)^{-(\nu+1)/2}, \quad \tau, \nu > 0$$

Gaussian Scale Mixtures

- $X = \rho Y$, $Y \sim N(0,1)$, $\rho \sim P(\rho)\mathrm{I}_{\{\rho>0\}}$
- Many distributions you know are scale mixtures
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)
  - Student's t



West, Biom. 87

- Duality between $P(X)$ and $P(\rho)$
- For the Laplace:

$$\frac{\tau}{2} e^{-\tau|s|} = \mathrm{E}[N(s|0,\gamma)], \quad \gamma \sim (\tau^2/2)e^{-(\tau^2/2)\gamma}$$

$$= \int N(s|0,\gamma)P(\gamma)\,d\gamma \quad [\texttt{scale\_mix\_plot}]$$

Super-Gaussian Potentials

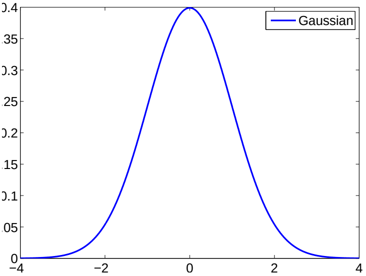
$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

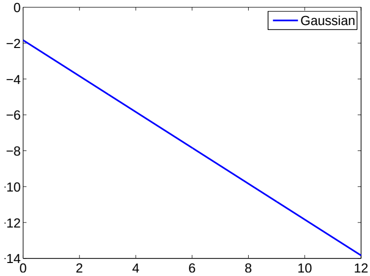
- $t_i(s_i)$ is even: Let's look at $s_i^2 \mapsto t_i(s_i)$
  $t_i(s_i)$ is positive: Let's look at $s_i^2 \mapsto 2 \log t_i(s_i)$ F11
- What's that for a Gaussian $t_i(s_i) = N(s_i|0, \sigma_i^2)$?

Super-Gaussian Potentials

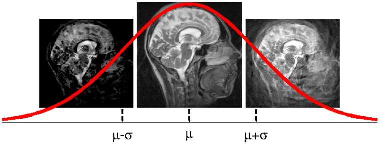
$$P(\boldsymbol{u}|\boldsymbol{y}) = Z^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i) \approx Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- $t_i(s_i)$ is even: Let's look at $s_i^2 \mapsto t_i(s_i)$
  $t_i(s_i)$ is positive: Let's look at $s_i^2 \mapsto 2 \log t_i(s_i)$
- What's that for a Gaussian $t_i(s_i) = N(s_i|0, \sigma_i^2)$?
  A linear (affine) function





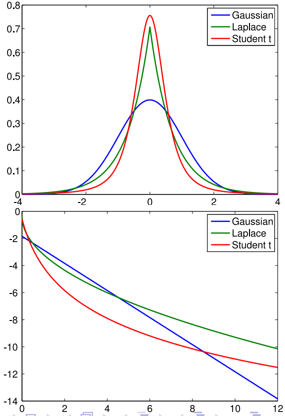
Super-Gaussian Potentials



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

Sparsity potentials are super-Gaussian F12

$$s^2 \mapsto 2 \log t(s) \quad \text{convex}$$



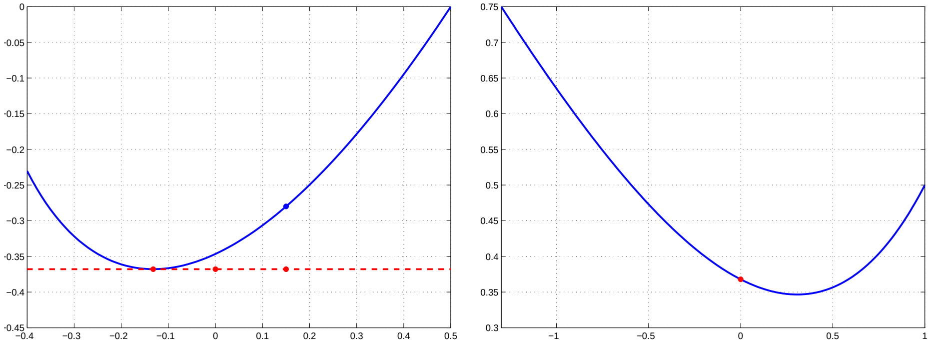
Convex (Legendre) Duality

Super-Gaussian:
$t(s)$ even, $s^2 \mapsto \log t(s)$ convex. Remember Jensen's inequality?

Convex (Legendre) Duality

Super-Gaussian:

$t(s)$ even, $s^2 \mapsto \log t(s)$ convex. Remember Jensen's inequality? F13

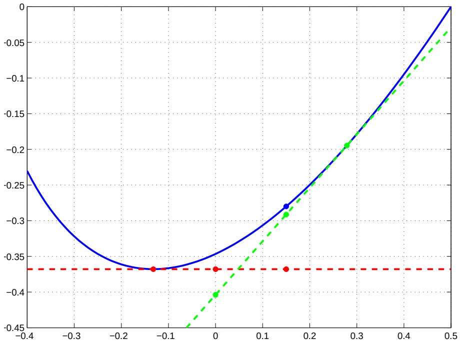


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

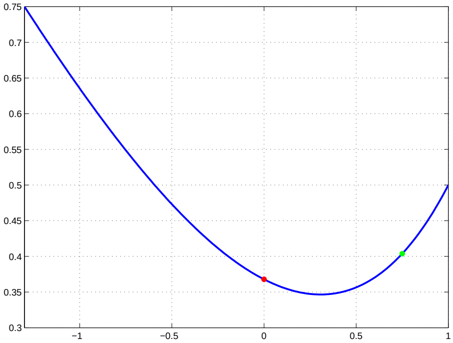
$$t(s) = \max_\gamma e^{(-s^2/\gamma - h(\gamma))/2} \qquad h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

Convex (Legendre) Duality

Super-Gaussian:
$t(s)$ even, $s^2 \mapsto \log t(s)$ convex.



Remember Jensen's inequality?



$$f(x) = \max_\pi x\pi - f^*(\pi)$$

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

$$f^*(\pi) = \max_x \pi x - f(x)$$

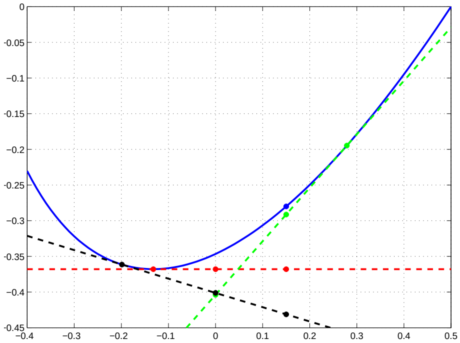
$$h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

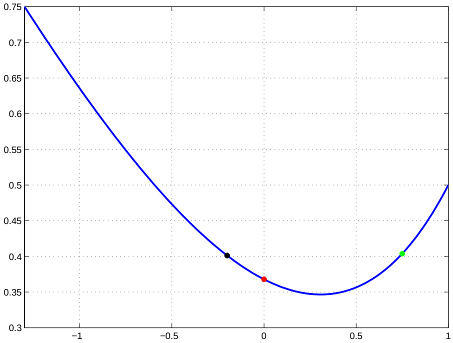
Convex (Legendre) Duality

Super-Gaussian:

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

Remember Jensen's inequality?





$$f(x) = \max_\pi x\pi - f^*(\pi)$$

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

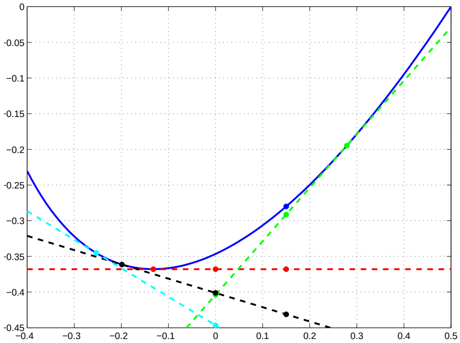
$$f^*(\pi) = \max_x \pi x - f(x)$$

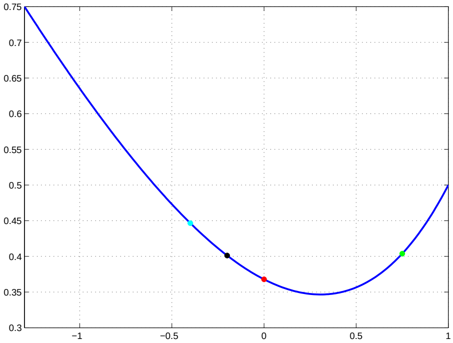
$$h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

Convex (Legendre) Duality

Super-Gaussian:
$t(s)$ even, $s^2 \mapsto \log t(s)$ convex.

Remember Jensen's inequality?





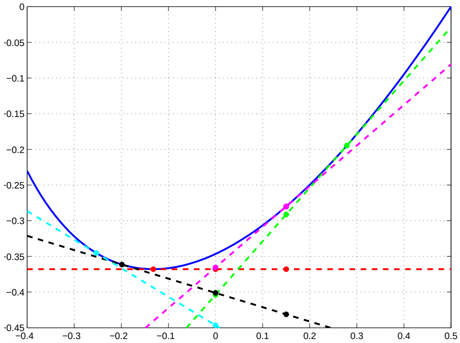
$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

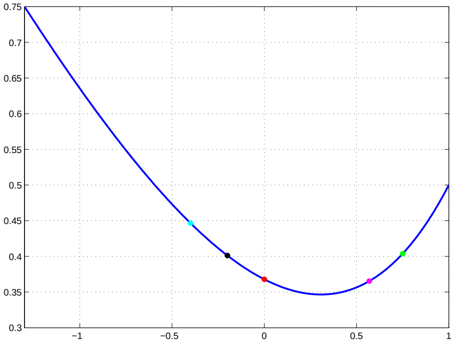
$$t(s) = \max_\gamma e^{(-s^2/\gamma - h(\gamma))/2} \qquad h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

Convex (Legendre) Duality

Super-Gaussian:
$t(s)$ even, $s^2 \mapsto \log t(s)$ convex.

Remember Jensen's inequality?





$$f(x) = \max_\pi x\pi - f^*(\pi)$$

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

$$f^*(\pi) = \max_x \pi x - f(x)$$

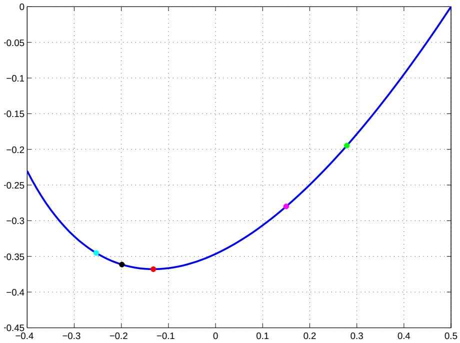
$$h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

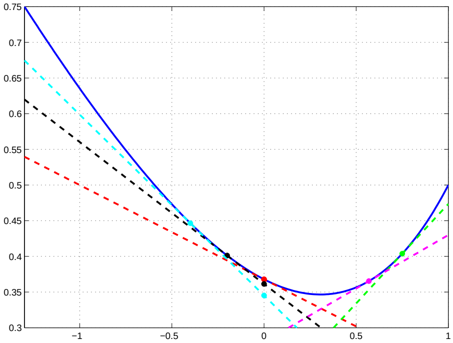
Convex (Legendre) Duality

Super-Gaussian:

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

Remember Jensen's inequality?

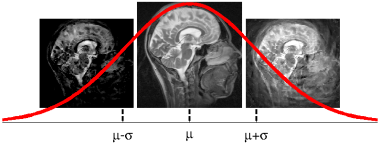




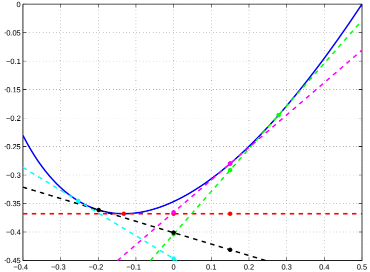
$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

$$t(s) = \max_\gamma e^{(-s^2/\gamma - h(\gamma))/2} \qquad h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

Super-Gaussian Potentials



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



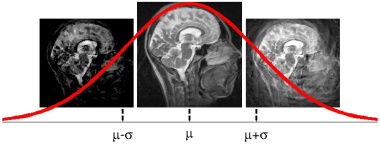
Sparsity potentials are super-Gaussian

$$s_i^2 \mapsto 2\log t_i(s_i) \quad \text{convex}$$

Convex (Legendre) duality

$$2\log t_i(s_i) = \max_{\pi_i}(s_i^2)\pi_i - f^*(\pi_i)$$

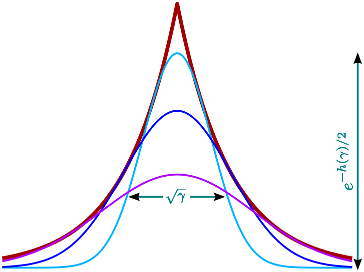
Super-Gaussian Potentials



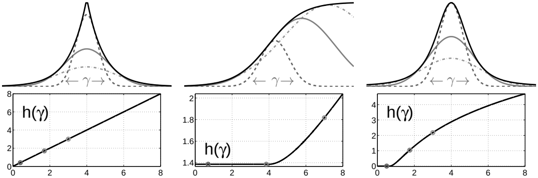
$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Sparsity potentials are super-Gaussian

$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

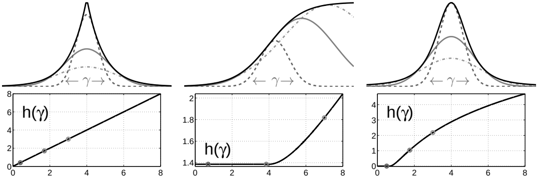


Super-Gaussian Potentials



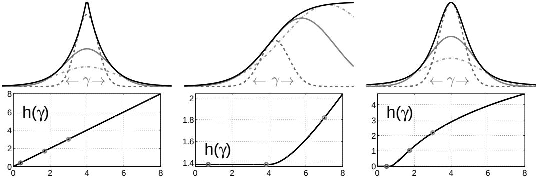
- $t(s) = \hat{t}(s)e^{\kappa s}$ super-Gaussian iff
  - $\hat{t}(s)$ even function (for some κ; $\kappa = 0$ if $t(s)$ itself even)
  - $s^2 \mapsto \log \hat{t}(s)$ convex, decreasing

Super-Gaussian Potentials



- $t(s) = \hat{t}(s)e^{\kappa s}$ super-Gaussian iff
  - $\hat{t}(s)$ even function (for some κ; $\kappa = 0$ if $t(s)$ itself even)
  - $s^2 \mapsto \log \hat{t}(s)$ convex, decreasing
- Bernoulli (logistic) $t(s) = (1 + e^{-y\tau s})^{-1}$, $y \in \{\pm 1\}$? F15

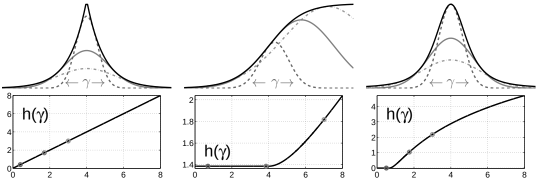
Super-Gaussian Potentials



- $t(s) = \hat{t}(s)e^{\kappa s}$ super-Gaussian iff
  - $\hat{t}(s)$ even function (for some κ; $\kappa = 0$ if $t(s)$ itself even)
  - $s^2 \mapsto \log \hat{t}(s)$ convex, decreasing
- Bernoulli (logistic) $t(s) = (1 + e^{-y\tau s})^{-1}$, $y \in \{\pm 1\}$?
  $t(s) = \hat{t}(s)e^{(y\tau/2)s}$, $g(x) \doteq -\log\cosh((y\tau/2)x^{1/2})$
- All scale mixtures are super-Gaussian

Palmer *et.al.*, NIPS 2005

Super-Gaussian Potentials



- $t(s) = \hat{t}(s)e^{\kappa s}$ super-Gaussian iff
  - $\hat{t}(s)$ even function (for some κ; $\kappa = 0$ if $t(s)$ itself even)
  - $s^2 \mapsto \log \hat{t}(s)$ convex, decreasing
- Bernoulli (logistic) $t(s) = (1 + e^{-y\tau s})^{-1}$, $y \in \{\pm 1\}$?
  $t(s) = \hat{t}(s)e^{(y\tau/2)s}$, $g(x) \doteq -\log\cosh((y\tau/2)x^{1/2})$
- All scale mixtures are super-Gaussian

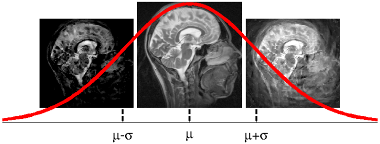
Palmer *et.al.*, NIPS 2005

F4b

- Some closure properties: $\{t_i(s_i)\}$ super-Gaussian, $\alpha_i > 0$
  - $\prod_i t_i(s_i)^{\alpha_i}$ super-Gaussian
  - $\sum_i \alpha_i t_i(s_i)$ super-Gaussian

F15b

Super-Gaussian Bounding

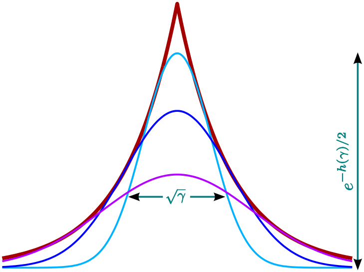


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

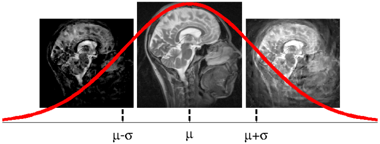
Sparsity potentials are super-Gaussian

$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2},$$

$$h(\gamma) := \sum\nolimits_i h_i(\gamma_i)$$



Super-Gaussian Bounding

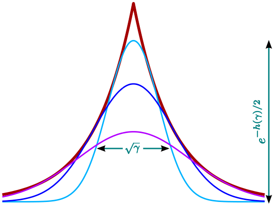


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Exact representation

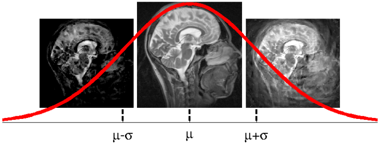
$$\log Z$$
$$= \log \int P(\boldsymbol{y}|\boldsymbol{u}) \max_{\gamma} e^{-(\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} + h(\gamma))/2} \, d\boldsymbol{u}$$

F16



$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

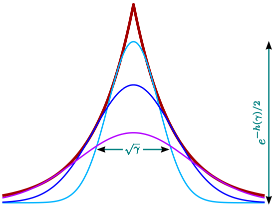
Super-Gaussian Bounding



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

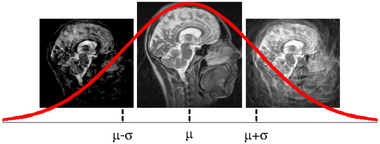
Lower bound

$$\begin{aligned} &\log Z \\ = &\log \int P(\boldsymbol{y}|\boldsymbol{u}) \max_{\gamma} e^{-(\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} + h(\gamma))/2} \, d\boldsymbol{u} \\ \geq &\max_{\gamma} \log \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-(\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} + h(\gamma))/2} \, d\boldsymbol{u} \end{aligned}$$



$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

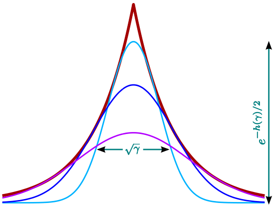
Super-Gaussian Bounding



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

Lower bound

$$\begin{aligned} &\log Z \\ \geq\ &\max_{\gamma} \log \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-(\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} + h(\gamma))/2}\, d\boldsymbol{u} \\ =\ &\max_{\gamma} \log Z_Q(\gamma) - h(\gamma)/2 \end{aligned}$$

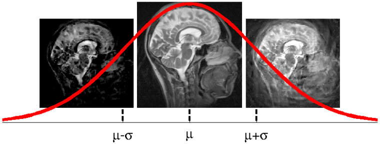


$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

Gaussian approximation

$$Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q{}^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}, \ \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$$

Super-Gaussian Bounding



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

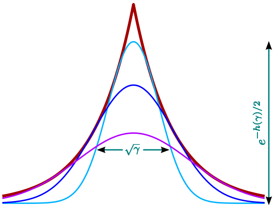
Variational problem: $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$

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

Gaussian approximation

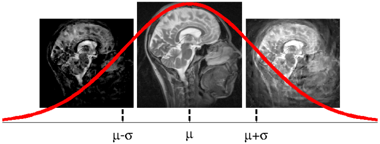
$$Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}, \ \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},$$

$$Z_Q = \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2} \, d\boldsymbol{u}$$



$$t_i(s_i) = \max_{\gamma_i > 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

Super-Gaussian Bounding

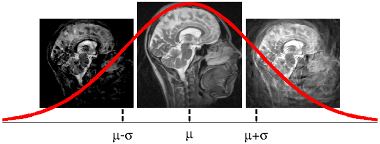


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i > 0} \dots$
  $\Rightarrow$ Auxiliary variables $\gamma \succ \mathbf{0}$

Super-Gaussian Bounding

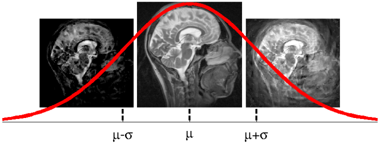


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i > 0} \dots$
  $\Rightarrow$ Auxiliary variables $\gamma \succ \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_{\gamma}$
  $\Rightarrow$ Global lower bound on $\log Z$ (not tight)

Super-Gaussian Bounding

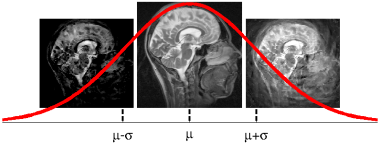


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i > 0} \dots$
  $\Rightarrow$ Auxiliary variables $\gamma \succ \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_\gamma$
  $\Rightarrow$ Global lower bound on $\log Z$ (not tight)
- Lower bounds are log partition functions of Gaussians $Q(\boldsymbol{u}|\boldsymbol{y})$
  $\Rightarrow$ Approximation family $\mathcal{Q} = \{Q(\boldsymbol{u}|\boldsymbol{y})\}$

Super-Gaussian Bounding



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

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i > 0} \dots$
  $\Rightarrow$ Auxiliary variables $\boldsymbol{\gamma} \succ \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_{\boldsymbol{\gamma}}$
  $\Rightarrow$ Global lower bound on $\log Z$ (not tight)
- Lower bounds are log partition functions of Gaussians $Q(\boldsymbol{u}|\boldsymbol{y})$
  $\Rightarrow$ Approximation family $\mathcal{Q} = \{Q(\boldsymbol{u}|\boldsymbol{y})\}$
- Divergence $Q(\boldsymbol{u}|\boldsymbol{y}) \leftrightarrow P(\boldsymbol{u}|\boldsymbol{y})$? Maximize lower bound!
  $\Rightarrow$ Divergence $\phi(\boldsymbol{\gamma}) = -2 \log Z_Q + h(\boldsymbol{\gamma})$

Coordinate Descent Algorithm

- Simple algorithm: Update single variables γ_j

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

Coordinate Descent Algorithm

- Simple algorithm: Update single variables γ_j

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

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

Coordinate Descent Algorithm

- Simple algorithm: Update single variables γ_j

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

- Representation of $Q(\boldsymbol{u}|\boldsymbol{y})$: Backbone for Gaussian propagation
  Moderate size problems: Cholesky representation
- Large scale problems?
  This algorithm is too slow (not scalable)

MAP Estimation and Variational Inference

MAP Estimation

$$\max_{\boldsymbol{u}} \log P(\boldsymbol{u}|\boldsymbol{y})Z$$

$$= \max_{\boldsymbol{u}} \log N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \max_{\boldsymbol{\gamma}} e^{-(\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}+h(\boldsymbol{\gamma}))/2}$$

$$||$$

$$\max_{\boldsymbol{\gamma}} \max_{\boldsymbol{u}} \log N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) e^{-(\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}+h(\boldsymbol{\gamma}))/2}$$

Bayesian Inference

$$\log Z$$

$$= \log \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \max_{\boldsymbol{\gamma}} e^{-(\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}+h(\boldsymbol{\gamma}))/2} \, d\boldsymbol{u}$$

$$\geq$$

$$\max_{\boldsymbol{\gamma}} \log \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) e^{-(\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}+h(\boldsymbol{\gamma}))/2} \, d\boldsymbol{u}$$

Wrap-Up

- Continuous-variable approximate inference: A different game
- Sparse linear model:
  Combinatorial properties with continuous variables
- Gaussian distributions (possibly graph-structured):
  Major backbone for continuous-variable inference
- Gaussian-form representations:
  - Scale mixtures
  - Super-Gaussian potentials
- Super-Gaussian bounding:
  From local potential bounds to global log partition function bound