Probabilistic Graphical Models

Inference for Continuous-Variable Models

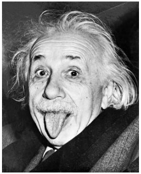
Volkan Cevher, Matthias Seeger
Ecole Polytechnique Fédérale de Lausanne

28/11/2011

Outline

Sparsity: A Fundamental Concept

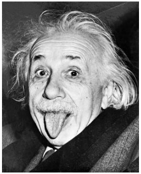
... as simple as possible, but not simpler.



What do you mean with simple?

Sparsity: A Fundamental Concept

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



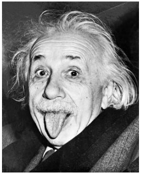
What do you mean with simple?

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?

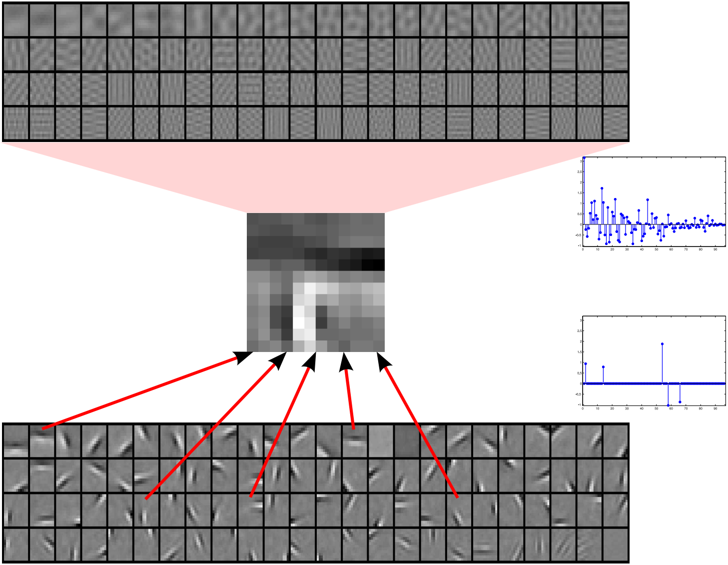
Classical (Gaussian)

- All specified elements
- Use each of them a little

Sparsity

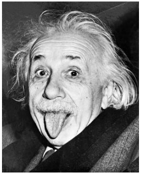
- As few elements as possible
- If at all, use them big

Sparsity: A Fundamental Concept



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

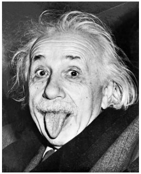
- As few elements as possible
- If at all, use them big

Classical linear framework: Shapes the way we think

- Nyquist/Shannon limit. Point spread function
- Aliasing. Ringing. Signal-to-noise ratio
- Linear measurements? Linear reconstruction!

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

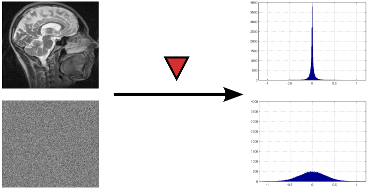
Sparsity

- 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

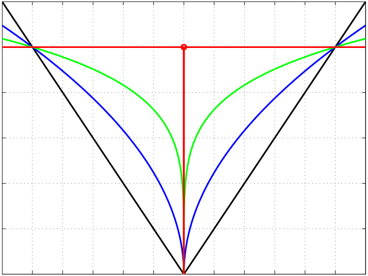
Many Faces of Sparsity

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



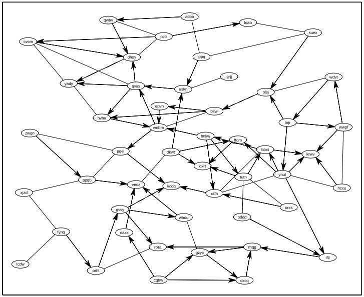
Many Faces of Sparsity

- 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

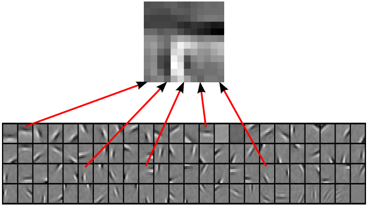
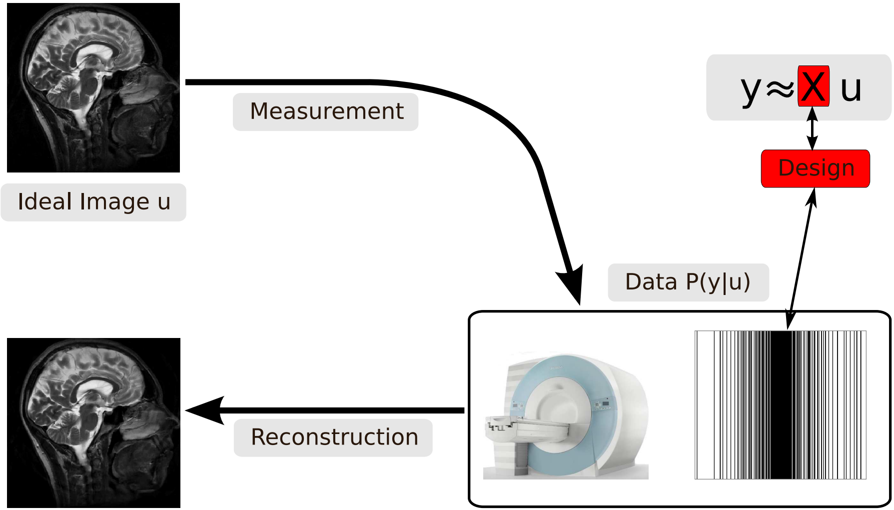
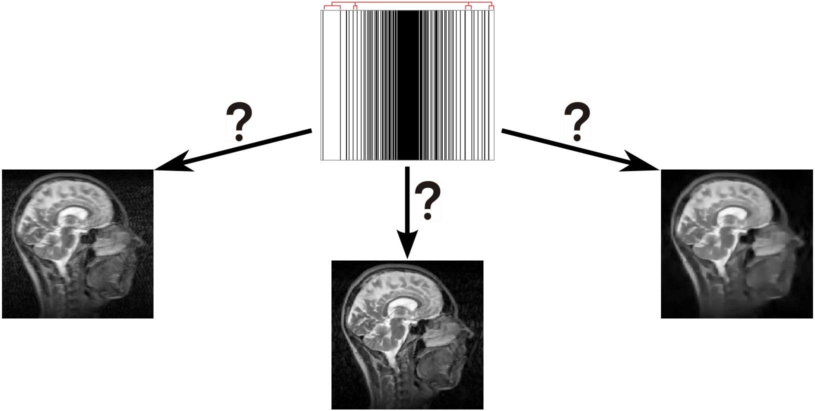


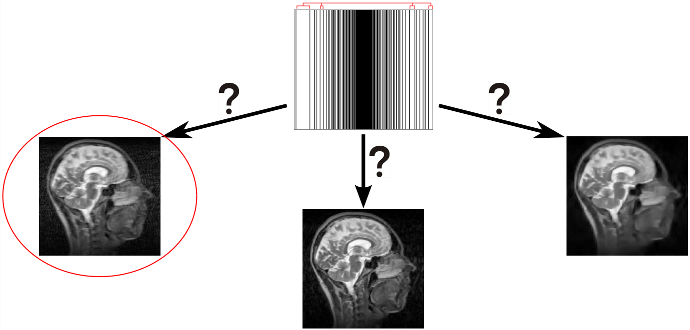
Image Reconstruction



Reconstruction is Ambiguous



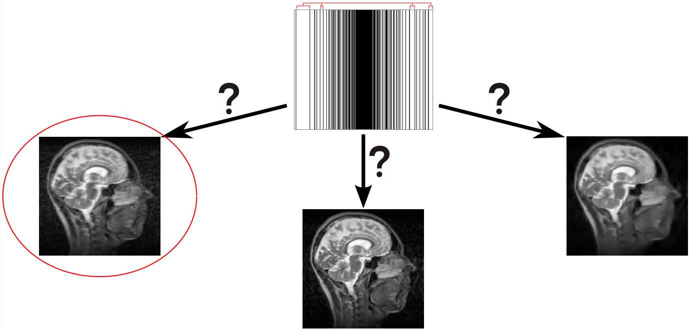
Least Squares Estimation



Least Squares Estimation (Linear Model)

$$\boldsymbol{u}_* = \operatorname{argmin}_{\boldsymbol{u}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 \quad \text{s.t. } \|\boldsymbol{u}\|^2 \text{ small}$$

Least Squares Estimation



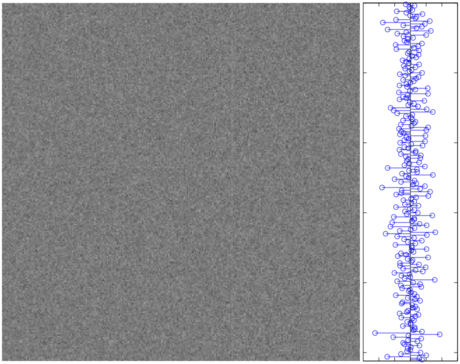
Least Squares Estimation (Linear Model)

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

Image Statistics

Whatever images are ...

they are not Gaussian!

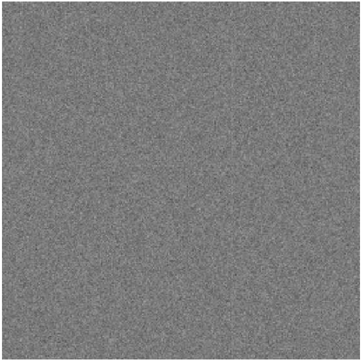


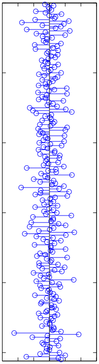
- Small noisy steps
- Gaussian random walker through pixel-land

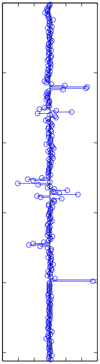
Image Statistics

Whatever images are . . .

they are not Gaussian!







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

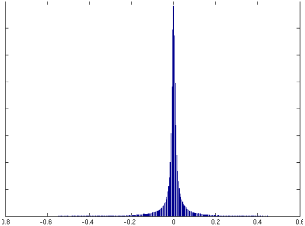
- Tiptoeing, edge jumping
- Gaussian won't do

Image Statistics

Whatever images are . . .

they are not Gaussian!

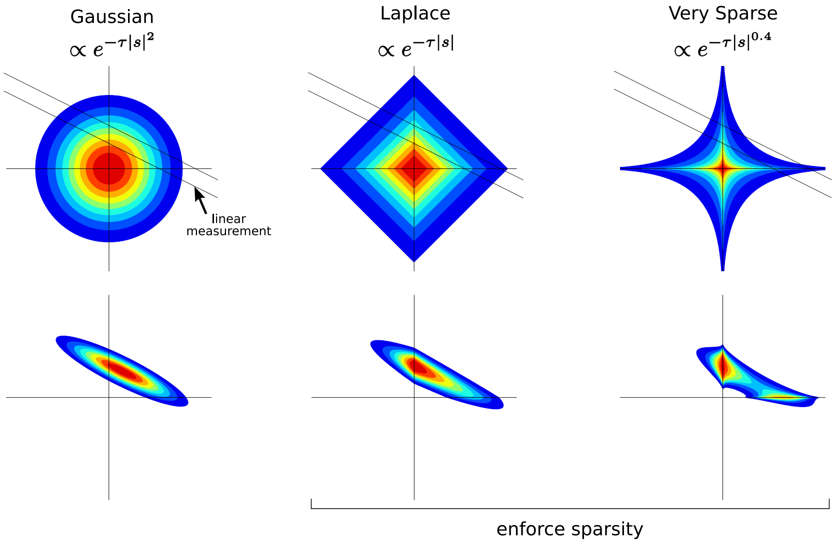
- Spatial smoothness: Image gradient super-Gaussian, sparse



Capture image properties in prior distribution $P(\boldsymbol{u})$

Sparsity Priors

courtesy Florian Steinke

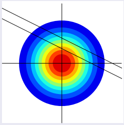


Best of Both Worlds

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

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

- Simple. Fast
- Well understood

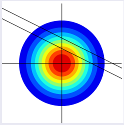


Best of Both Worlds

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

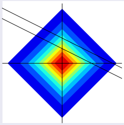
Gaussian Prior $P(\boldsymbol{u})$

- Simple. Fast
- Well understood



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

- Better prior for real-world signals (images)

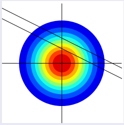


Best of Both Worlds

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

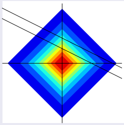
Gaussian Prior $P(\boldsymbol{u})$

- Simple. Fast
- Well understood



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

- Better prior for real-world signals (images)



Latent Gaussian Representations

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

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

Gaussian Scale Mixtures

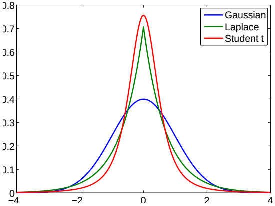
- Mixture of Gaussians? K-means, EM, ...

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

$t_i(s_i)$ 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, ...

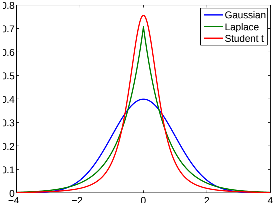
$$P(X) = \sum_{j=1}^{K} \pi_j N(X|\mu_j, \sigma^2)$$

$t_i(s_i)$ 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

$\Rightarrow$ Mass at different scales

- Why not mix over the scales?



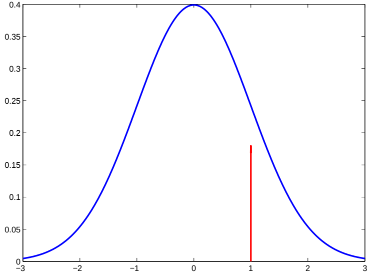
Gaussian Scale Mixtures

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

Gaussian Scale Mixtures

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

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

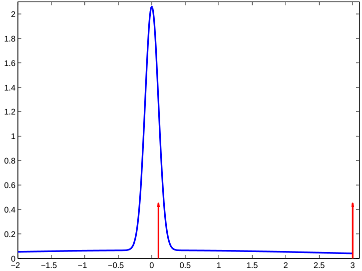


$$P(X) = N(X|0, \gamma)$$

Gaussian Scale Mixtures

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

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

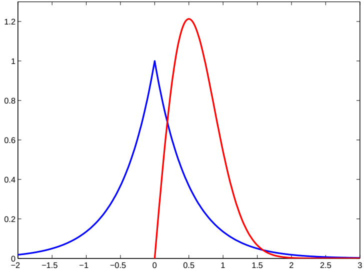


$$P(X) = \pi N(X|0,\gamma_1) + (1-\pi)N(X|0,\gamma_2), \quad \gamma_1 \ll \gamma_2$$

Gaussian Scale Mixtures

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

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

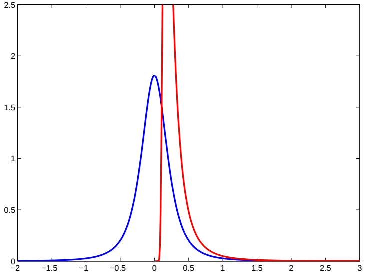


$$P(X) \propto e^{-\tau |X|^{\alpha}}, \quad \alpha \in (0,2],\ \tau > 0$$

Gaussian Scale Mixtures

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

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

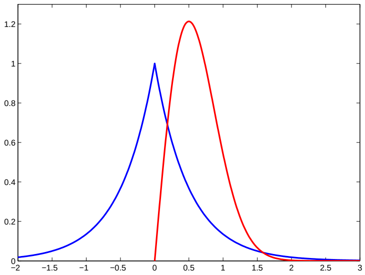


$$P(X) \propto (1 + (\tau/\nu)|X|^2)^{-(\nu+1)/2}, \quad \tau, \nu > 0$$

Gaussian Scale Mixtures

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

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



- Duality between $P(X)$ and $f(\gamma)$

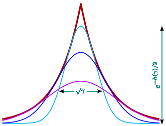
West, Biometrika 87

- For the Laplace:

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

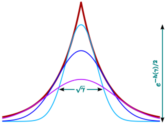
Super-Gaussian Potentials

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



Super-Gaussian Potentials

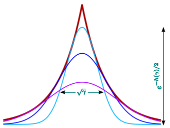
$$t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma)$$



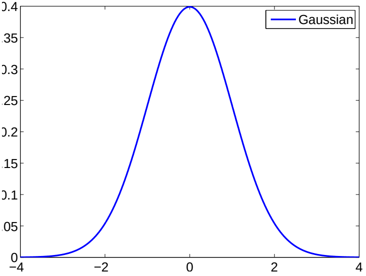
- $t(s)$ even and positive: Let's look at $|s|^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(|s||0, \sigma^2)$? F7

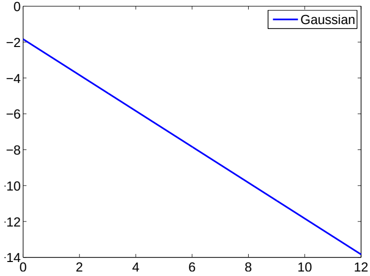
Super-Gaussian Potentials

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



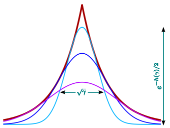
- $t(s)$ even and positive: Let's look at $|s|^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(|s||0, \sigma^2)$?
  A linear (affine) function





Super-Gaussian Potentials

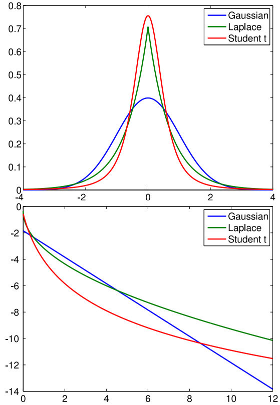
$$t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma)$$



Sparsity potentials are super-Gaussian

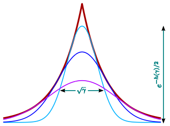
$$|s|^2 \mapsto 2 \log t(s) \quad \text{is convex}$$

- Affine $\rightarrow$ convex:
  Shift mass to center and tails



Super-Gaussian Potentials

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

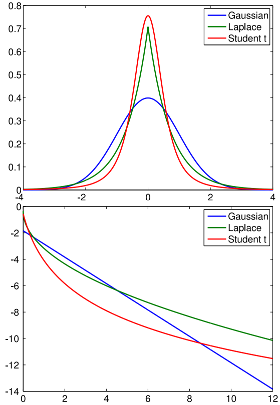


Sparsity potentials are super-Gaussian

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

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

Palmer *et.al.*, NIPS 2005



Scale Mixtures are Super-Gaussian

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

- $t(s)$ even and positive: F9
  $x := |s|^2 \quad \Rightarrow \quad t(s) = e^{g(x)}$

Scale Mixtures are Super-Gaussian

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

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

Scale Mixtures are Super-Gaussian

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

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

Boyd, Vandenberghe, 2002

$$\psi(x, \gamma) \ \text{convex} \ \forall \gamma \in \mathcal{C} \quad \Rightarrow \quad \log \int_{\mathcal{C}} e^{\psi(x,\gamma)}\, d\gamma \ \text{convex}$$

Scale Mixtures are Super-Gaussian

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

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

Boyd, Vandenberghe, 2002

$$\psi(x, \gamma) \;\text{convex}\; \forall \gamma \in \mathcal{C} \quad \Rightarrow \quad \log \int_{\mathcal{C}} e^{\psi(x,\gamma)}\, d\gamma \;\text{ convex}$$

- Apply to $g(x)$:

$$g(x) = \log \int_{\geq 0} e^{-x/(2\gamma)} f(\gamma)\, d\gamma = \log \int_{\geq 0} e^{-x/(2\gamma) + \log f(\gamma)}\, d\gamma$$

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 $\|\boldsymbol{s}_i\|$

Group Sparsity

$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$
- Useful for complex values: $|s_i| = \|(\Re s_i, \Im s_i)^T\|$

Group Sparsity

$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$
- 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

$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$
- 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
  $e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} \propto N(\boldsymbol{s}_i|\boldsymbol{0}, \gamma_i \boldsymbol{1})$

Group Sparsity

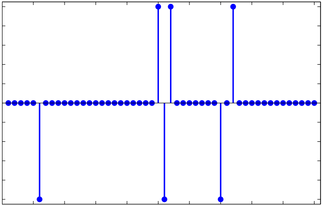
$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$
- 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 $e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} \propto N(\boldsymbol{s}_i|\boldsymbol{0}, \gamma_i \boldsymbol{1})$

Sparsity vs. Super-Gaussianity

Sparse $\boldsymbol{s}$

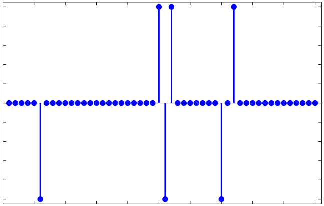
- Many/most $s_i = 0$



Sparsity vs. Super-Gaussianity

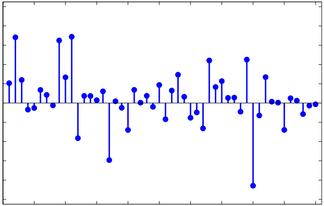
Sparse $\boldsymbol{s}$

- Many/most $s_i = 0$



Super-Gaussian $\boldsymbol{s}$

- Super-Gaussian statistics
- Soft sparsity, statistical sparsity, power law decay, ...



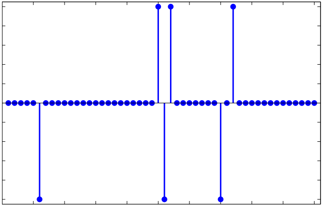
Sparsity vs. Super-Gaussianity

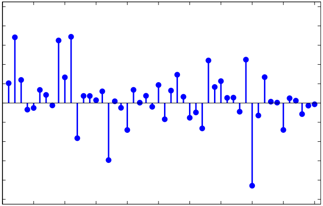
Sparse $\boldsymbol{s}$

- Many/most $s_i = 0$

Super-Gaussian $\boldsymbol{s}$

- Super-Gaussian statistics
- Soft sparsity, statistical sparsity, power law decay, ...





- $P(\boldsymbol{s})$ super-Gaussian: $\boldsymbol{s} \sim P(\boldsymbol{s})$ no zeros in general (only if $P(\boldsymbol{s})$ degenerate)

Where Are We?

- 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

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

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

Variational Approximations

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

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

Variational Approximations

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

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

Variational approximation

Apply variational principle to fit master function $\log Z$

Variational Approximations

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

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

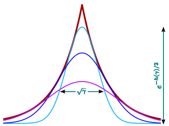
Variational approximation

Apply variational principle to fit master function $\log Z$

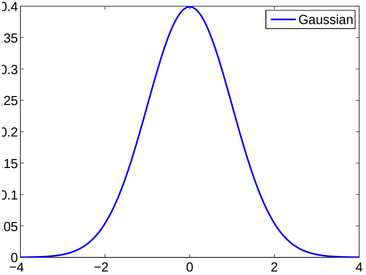
- Super-Gaussian bounding
- Expectation propagation
- Variational mean field Bayes [not here]
- Gaussian KL minimization [not here]

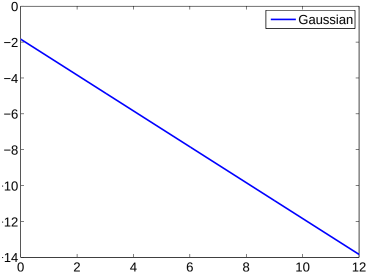
Super-Gaussian Potentials

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



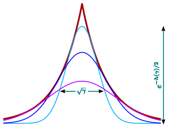
- $t(s)$ even and positive: Let's look at $|s|^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(|s||0, \sigma^2)$?
  A linear (affine) function





Super-Gaussian Potentials

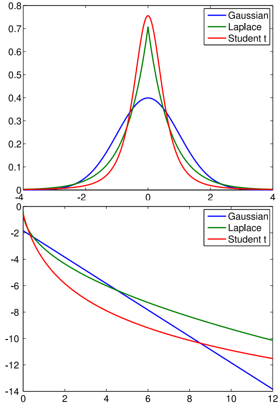
$$t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} e^{-h(\gamma)/2}$$



Sparsity potentials are super-Gaussian

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

- $t(s) = \max_{\gamma \geq 0} \ldots$. Why?



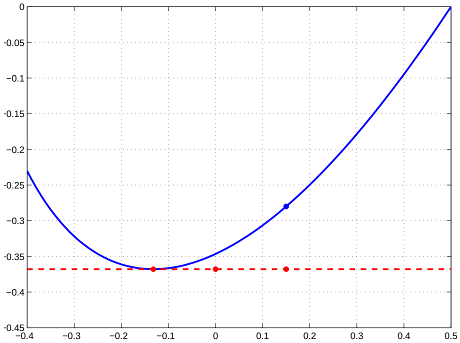
Convex (Fenchel) Duality

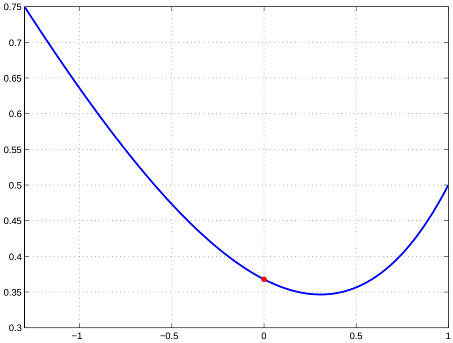
Super-Gaussian:
$t(s)$ even, $|s|^2 \mapsto \log t(s)$ convex.

Convex function:	Maximum of its affine lower bounds
Super-Gaussian function:	Maximum of its Gaussian lower bounds

Convex (Fenchel) Duality

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



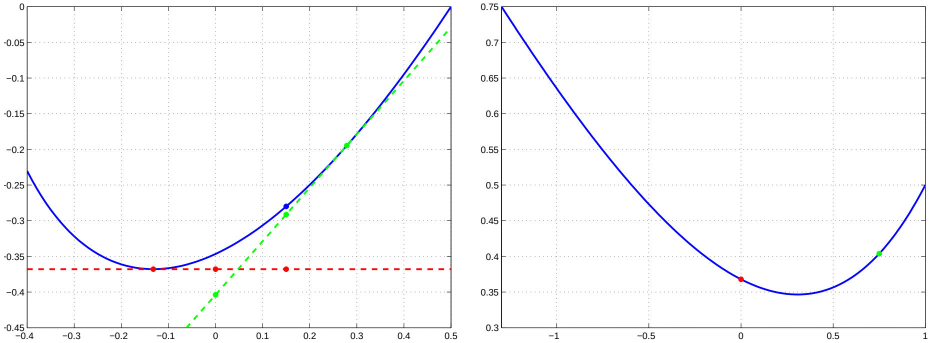


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Convex (Fenchel) Duality

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

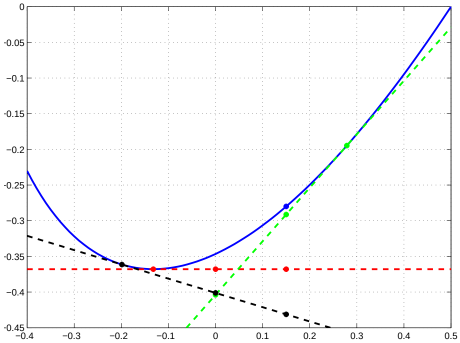


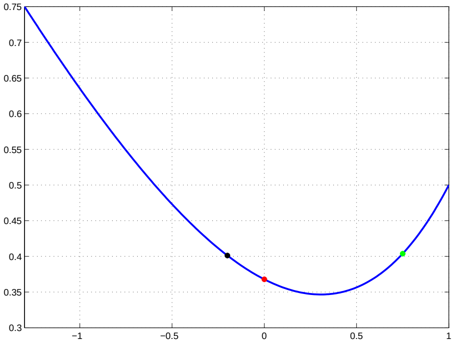
$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Convex (Fenchel) Duality

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



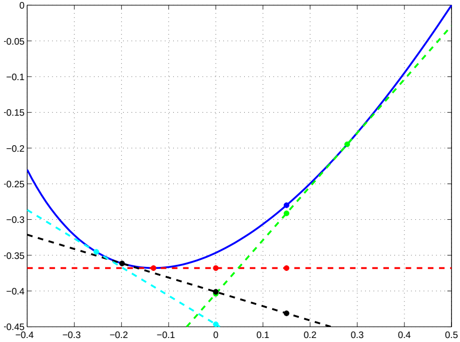


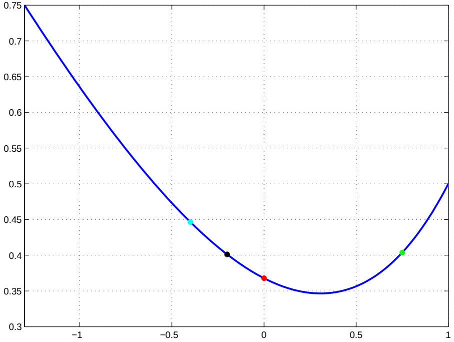
$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Convex (Fenchel) Duality

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



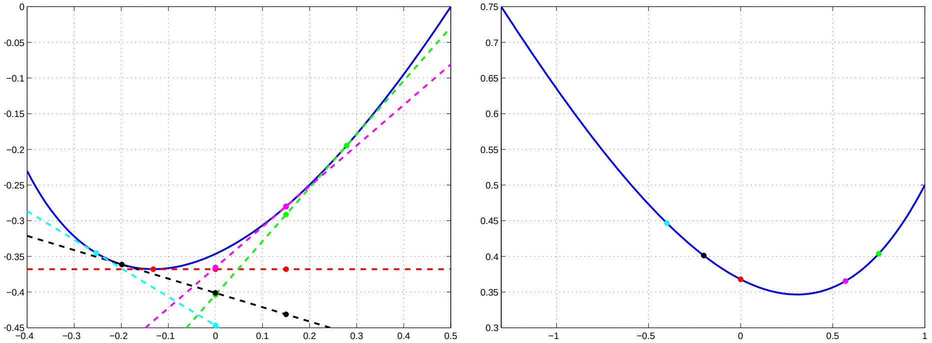


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Convex (Fenchel) Duality

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

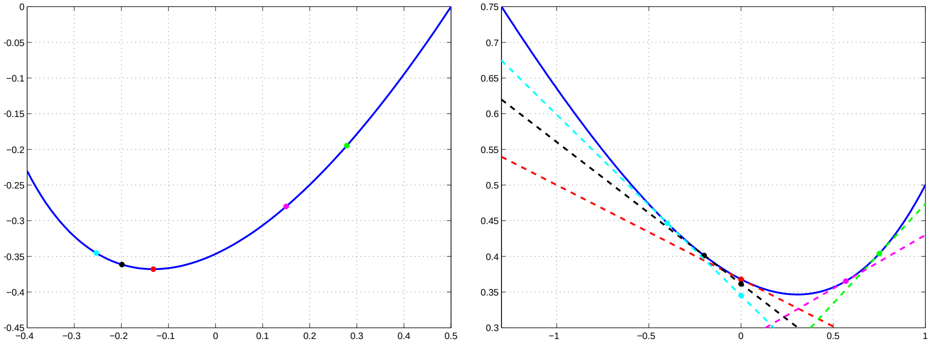


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Convex (Fenchel) Duality

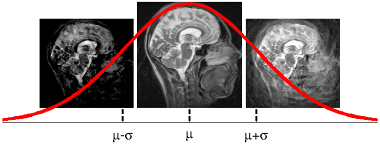
Super-Gaussian:
$t(s)$ even, $|s|^2 \mapsto \log t(s)$ convex.



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

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

Super-Gaussian Potentials



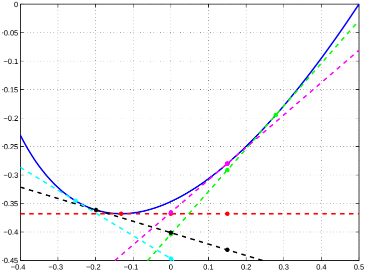
$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Sparsity potentials are super-Gaussian

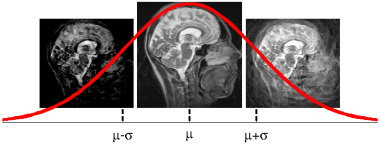
$$|s_i|^2 \mapsto 2\log t_i(s_i) \quad \text{is convex}$$

Convex (Fenchel) duality

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



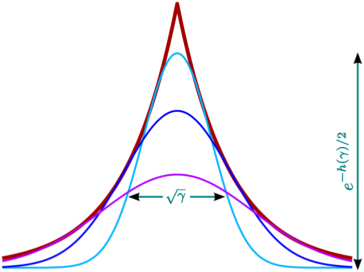
Super-Gaussian Potentials



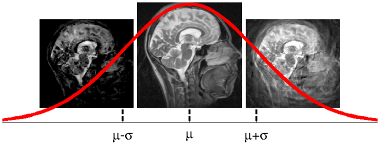
$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Sparsity potentials are super-Gaussian

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



Super-Gaussian Bounding

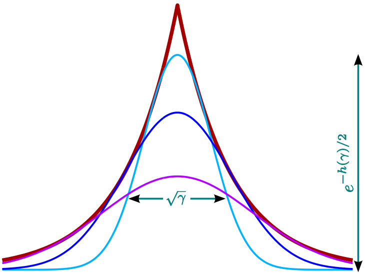


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

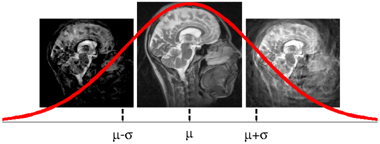
Sparsity potentials are super-Gaussian

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

$$h(\gamma) := \sum\nolimits_i h_i(\gamma_i), \quad \Gamma = \operatorname{diag} \gamma$$



Super-Gaussian Bounding

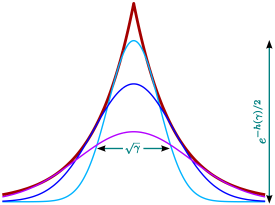


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Exact representation

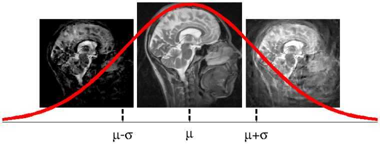
$$\log Z$$

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



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

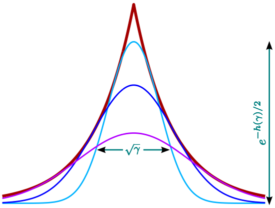
Super-Gaussian Bounding



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

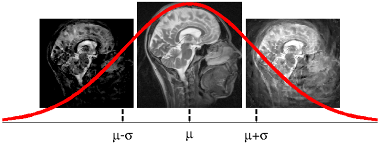
Lower bound

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



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

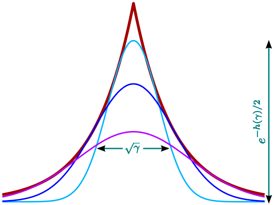
Super-Gaussian Bounding



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

Lower bound

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

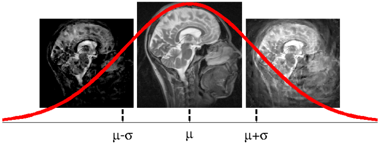


$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-|s_i|^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

Gaussian approximation

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

Super-Gaussian Bounding



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

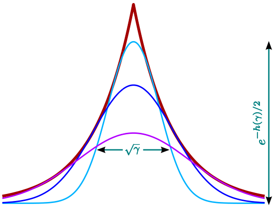
Variational problem: $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$

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

Gaussian approximation

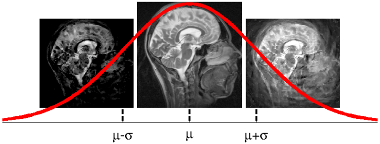
$$Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{-\boldsymbol{s}^H \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}, \ \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},$$

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



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

Super-Gaussian Bounding

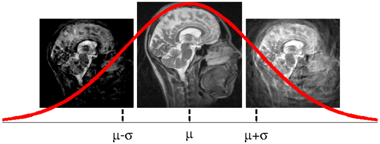


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

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

Super-Gaussian Bounding

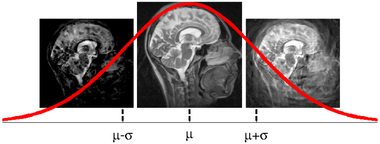


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i \geq 0} \dots$
  $\Rightarrow$ Auxiliary variables $\gamma \succeq \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_{\gamma}$
  $\Rightarrow$ Global lower bound on $\log Z$

Super-Gaussian Bounding

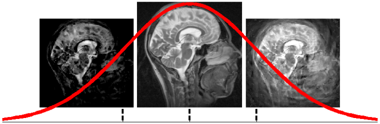


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

What did we do?

- Start with tight single potential bounds: $t_i(s_i) = \max_{\gamma_i \geq 0} \dots$
  $\Rightarrow$ Auxiliary variables $\boldsymbol{\gamma} \succeq \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_{\boldsymbol{\gamma}}$
  $\Rightarrow$ Global lower bound on $\log Z$
- 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: $t_i(s_i) = \max_{\gamma_i \geq 0} \dots$
  $\Rightarrow$ Auxiliary variables $\gamma \succeq \mathbf{0}$
- Plug into target function $\log Z$. Interchange $\int \dots d\boldsymbol{u} \leftrightarrow \max_\gamma$
  $\Rightarrow$ Global lower bound on $\log Z$
- 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!
  $\Rightarrow \phi(\gamma) = -2 \log Z_Q + h(\gamma)$

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

$$\log Z$$

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

$$\geq$$

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

Coordinate Descent Algorithm

- Simple algorithm: Update single variables γ_j

repeat
  for $j \in \{1, \dots, q\}$ **do**
    Update γ_j, based on marginal $Q(s_j|\boldsymbol{y})$
    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 $Q(s_j|\boldsymbol{y})$
    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 Seeger, JMLR 2008

Coordinate Descent Algorithm

- Simple algorithm: Update single variables γ_j

repeat
  for $j \in \{1, \dots, q\}$ **do**
    Update γ_j, based on marginal $Q(s_j|\boldsymbol{y})$
    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 Seeger, JMLR 2008
- Large scale problems?
  This algorithm is not scalable. Can do much better ...