Sparse Linear Models: Estimation and Approximate Bayesian Inference

Matthias Seeger

Laboratory for Probabilistic Machine Learning
Ecole Polytechnique Fédérale de Lausanne

http://lapmal.epfl.ch/

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Buzzwords

- Denoising
- Natural image statistics
- Wavelet shrinkage
- Image coding

Buzzwords

- Denoising
- Natural image statistics
- Wavelet shrinkage
- Image coding
- Feature selection
- ℓ_1 relaxation
- Learning model structure
- Sparse covariance estimation

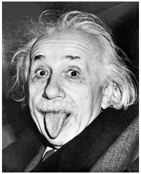
Buzzwords

- Denoising
- Natural image statistics
- Wavelet shrinkage
- Image coding
- Compressive sensing
- Below the Nyquist limit
- Sparse sampling
- Feature selection
- ℓ_1 relaxation
- Learning model structure
- Sparse covariance estimation

Buzzwords

- Denoising
- Natural image statistics
- Wavelet shrinkage
- Image coding
- Compressive sensing
- Below the Nyquist limit
- Sparse sampling
- Feature selection
- ℓ_1 relaxation
- Learning model structure
- Sparse covariance estimation
- Matching/basis pursuit
- Soft/hard thresholding
- {Group, graphical, adaptive} Lasso

Sparsity: A Fundamental Concept



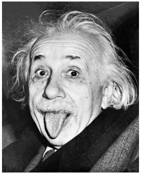
... as simple as possible, but not simpler.

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?

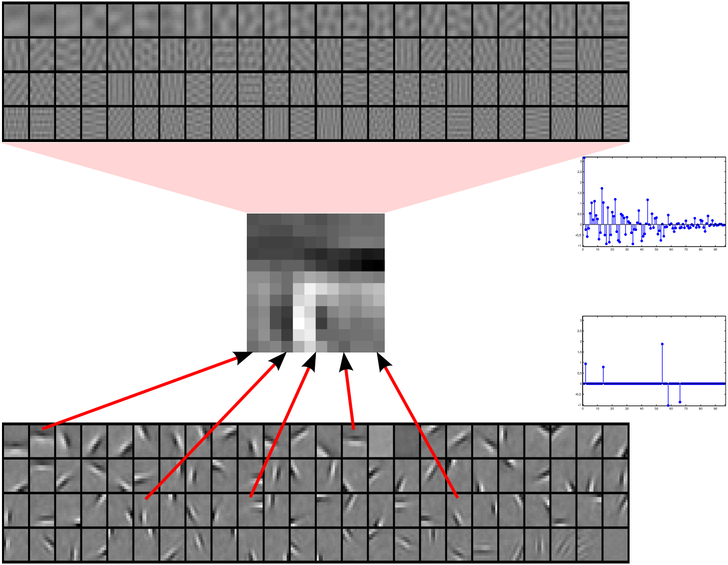
Classical (Gaussian)

- All specified elements
- Use each of them a little

Sparsity

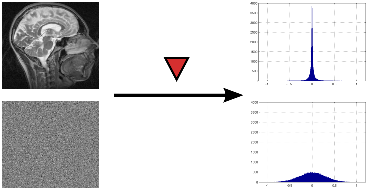
- As few elements as possible
- If at all, use them a lot

Sparsity: A Fundamental Concept



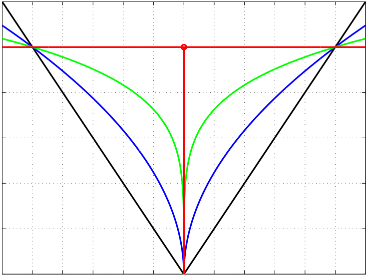
Many Faces of Sparsity

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



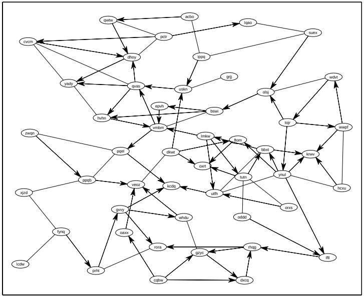
Many Faces of Sparsity

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

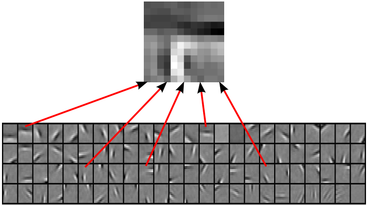


Many Faces of Sparsity

- 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

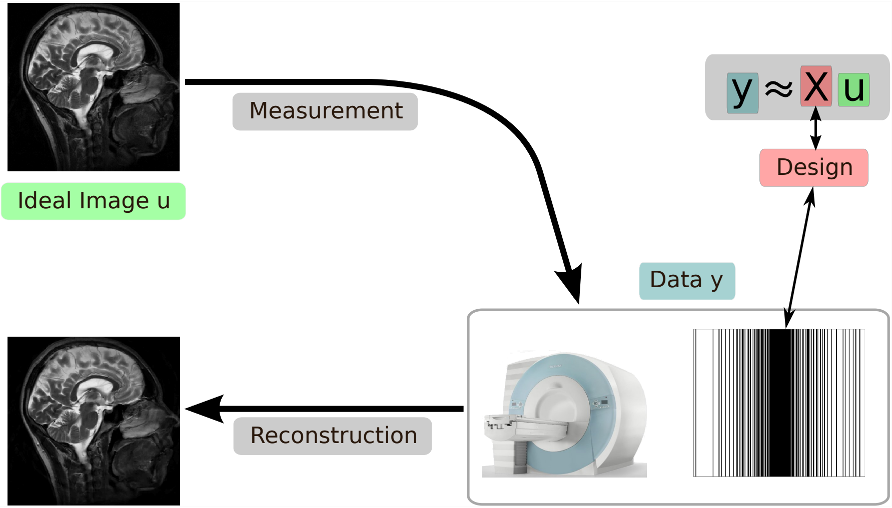
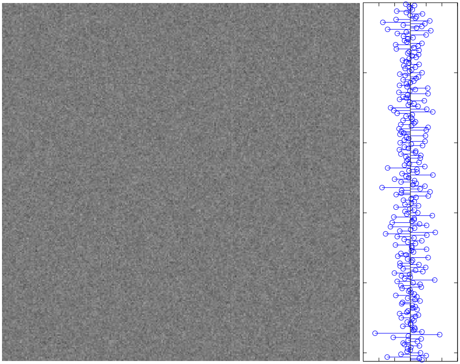
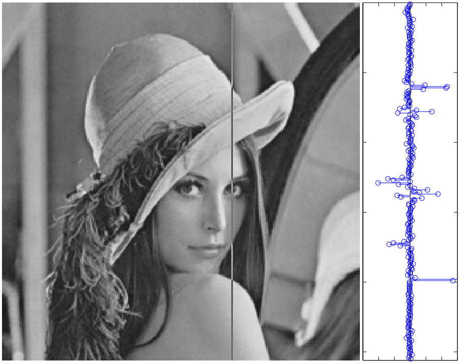


Image Statistics

Whatever images are . . .

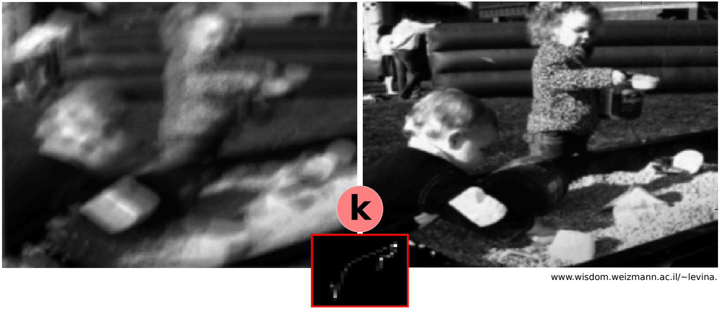
they are not Gaussian!





Bayesian Calibration

y



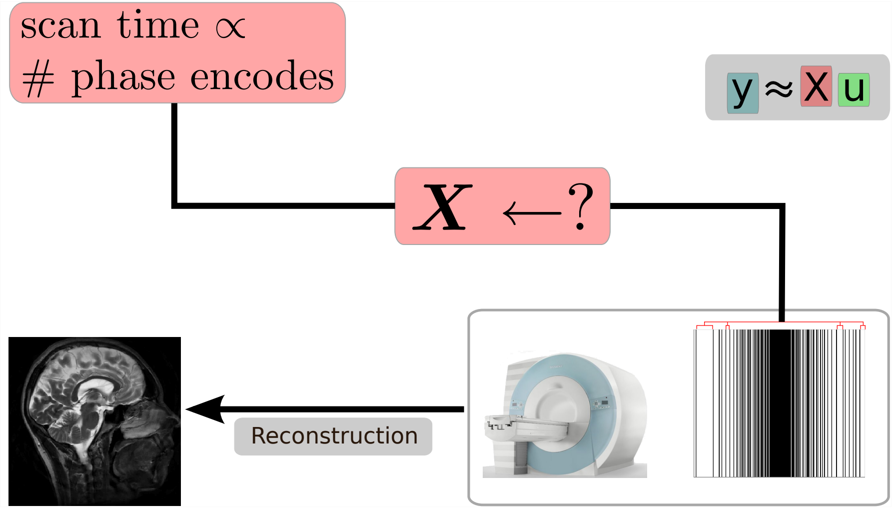
u

www.wisdom.weizmann.ac.il/~levina.

$$\mathbf{y} \approx \mathbf{k} \otimes \mathbf{u}$$

- Computer vision
  - Blind deconvolution
  - Calibrating camera parameters
- Magnetic resonance imaging
  - Autocalibrating parallel MRI

Bayesian Experimental Design

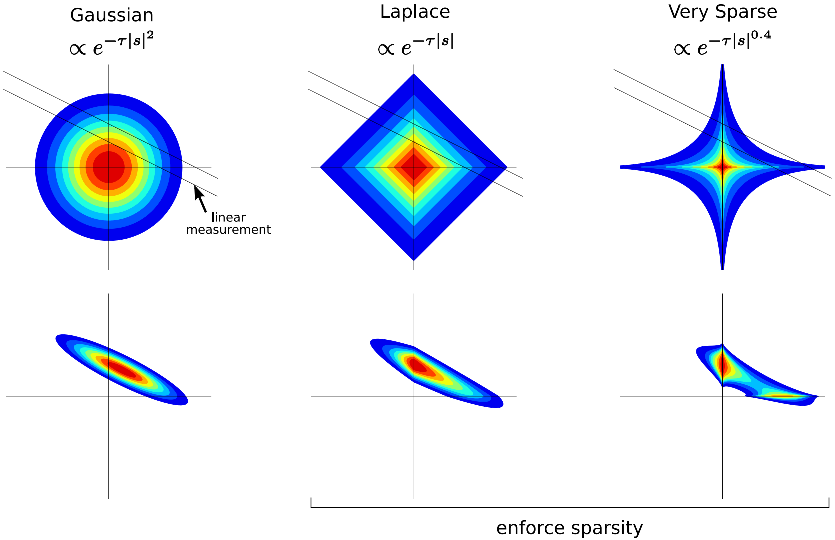


Outline

Outline

Sparsity Priors

courtesy Florian Steinke

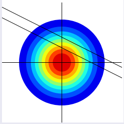


Best of Both Worlds

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

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

- Simple. Fast
- Well understood

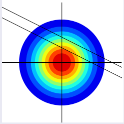


Best of Both Worlds

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

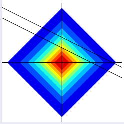
Gaussian Prior $P(\boldsymbol{u})$

- Simple. Fast
- Well understood



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

- Better prior for real-world signals (images)

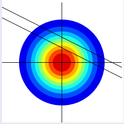


Best of Both Worlds

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

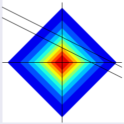
Gaussian Prior $P(\boldsymbol{u})$

- Simple. Fast
- Well understood



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

- Better prior for real-world signals (images)



Latent Gaussian Representations

- Gaussian scale mixtures $\quad t(s) = \int_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} f(\gamma)\, d\gamma$
- Super-Gaussian potentials $\quad t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma)$

Gaussian Scale Mixtures

- We know Gaussian mixtures over means (clustering, EM):

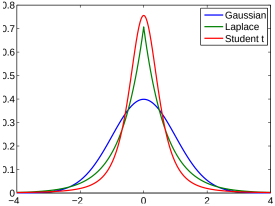
$$P(X) = \sum_{j=1}^{K} \pi_j N(X|\mu_j, \gamma)$$

Gaussian Scale Mixtures

- We know Gaussian mixtures over means (clustering, EM):

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

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



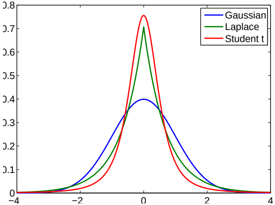
Gaussian Scale Mixtures

- We know Gaussian mixtures over means (clustering, EM):

$$P(X) = \sum_{j=1}^{K} \pi_j N(X|0, \gamma_j)$$

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

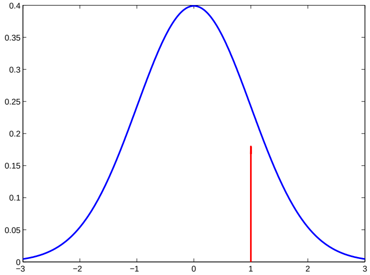
$\Rightarrow$ Need mixture over scales



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:
  - Gaussian [:-)].

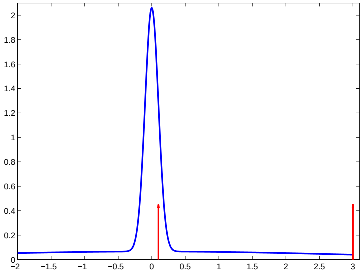


$$P(X) = N(X|0,\gamma)$$

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:
  - Gaussian [:-)]. Spike and slab

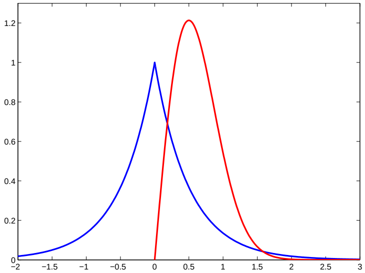


$$P(X) = \pi N(X|0,\gamma_1) + (1-\pi) N(X|0,\gamma_2), \quad \gamma_1 \ll \gamma_2$$

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:
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)

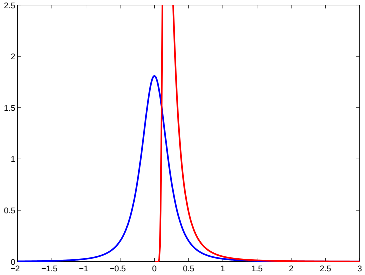


$$P(X) \propto e^{-\tau |X|^{\alpha}}, \quad \alpha \in (0,2],\ \tau > 0$$

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:
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)
  - Student's t

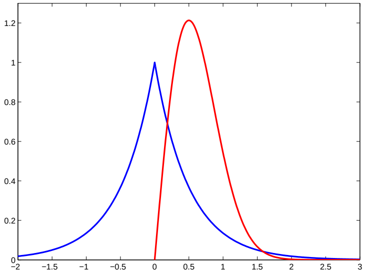


$$P(X) \propto \left(1 + \frac{\tau}{\nu}|X|^2\right)^{-(\nu+1)/2}, \quad \tau, \nu > 0$$

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:
  - Gaussian [:-)]. Spike and slab
  - Exponential power ($\alpha \leq 2$)
  - Student's t



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

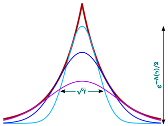
West, Biometrika 87

- For the Laplace:

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

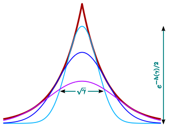
Super-Gaussian Potentials

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



Super-Gaussian Potentials

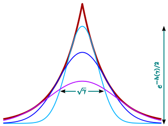
$$t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma)$$



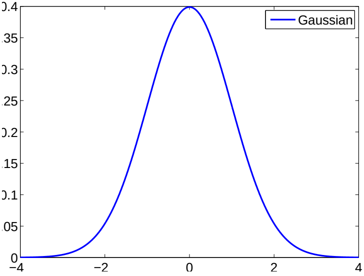
- $t(s)$ even and positive: Let's look at $|s|^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(s|0, \sigma^2)$?

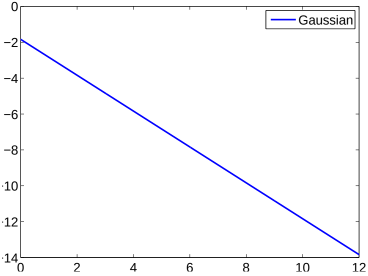
Super-Gaussian Potentials

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



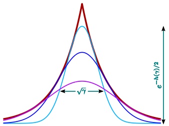
- $t(s)$ even and positive: Let's look at $|s|^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(s|0, \sigma^2)$?
  An affine function





Super-Gaussian Potentials

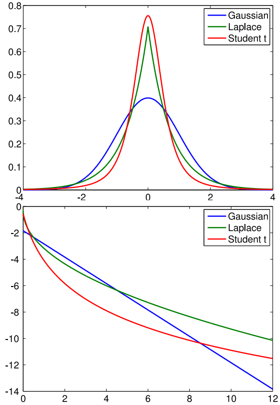
$$t(s) = \max_{\gamma \geq 0} e^{-|s|^2/(2\gamma)} g(\gamma)$$



Sparsity potentials are super-Gaussian

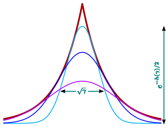
$$|s|^2 \mapsto 2 \log t(s) \quad \text{is convex}$$

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



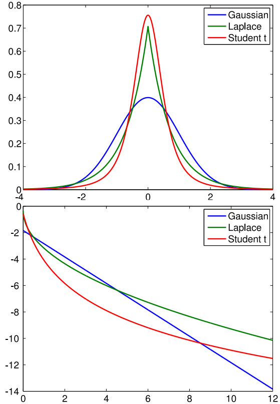
Super-Gaussian Potentials

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



Sparsity potentials are super-Gaussian

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



- Affine → convex:
  Shift mass to center and tails
- Scale mixtures are super-Gaussian

Palmer *et.al.*, NIPS 2005

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

Group Sparsity

$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} \underbrace{e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)}}_{\propto N(\boldsymbol{s}_i|\boldsymbol{0},\gamma_i \boldsymbol{I})} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$:
  Nothing but parameter tying

Group Sparsity

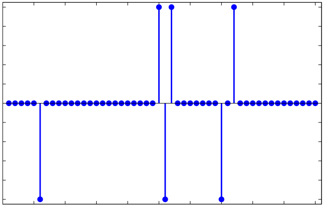
$$t(\boldsymbol{s}_i) = \max_{\gamma_i \geq 0} e^{-\|\boldsymbol{s}_i\|^2/(2\gamma_i)} g_i(\gamma_i)$$

- $t_i(s_i)$ depends on absolute value $|s_i|$ only
- Can just as well plug in vector norm $\|\boldsymbol{s}_i\|$:
  Nothing but parameter tying
- Useful to structure sparsity: Joint penalization of groups
  $\Rightarrow \ell_1 - \ell_2$ norms, group Lasso, ...

Sparsity vs. Super-Gaussianity

Sparse s

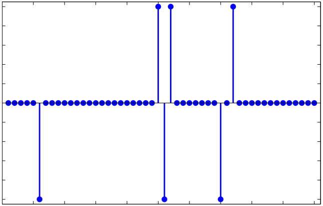
- Many/most $s_i = 0$



Sparsity vs. Super-Gaussianity

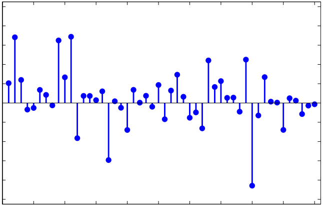
Sparse $\boldsymbol{s}$

- Many/most $s_i = 0$



Super-Gaussian $\boldsymbol{s}$

- Super-Gaussian statistics
- Soft sparsity, heavy tails, power law decay, ...



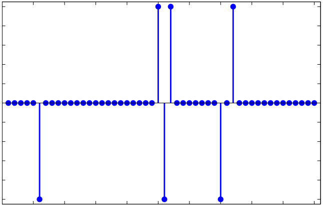
Sparsity vs. Super-Gaussianity

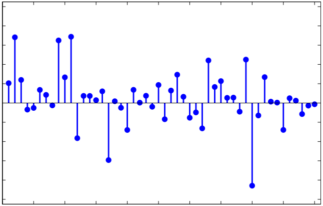
Sparse $\boldsymbol{s}$

- Many/most $s_i = 0$

Super-Gaussian $\boldsymbol{s}$

- Super-Gaussian statistics
- Soft sparsity, heavy tails, power law decay, . . .





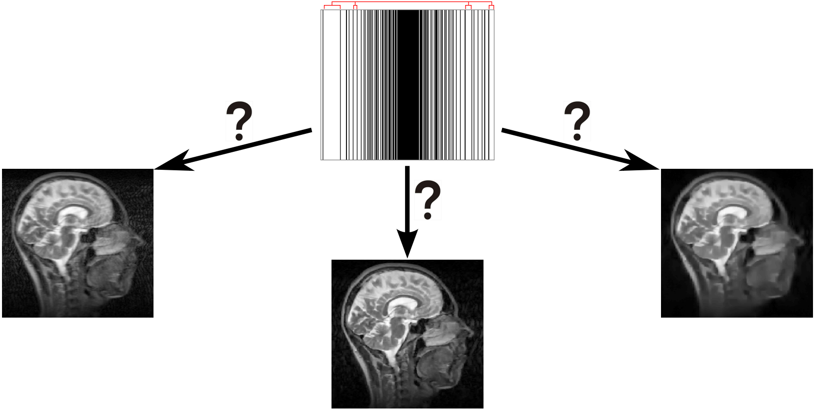
- Why call it sparse then?
  - “Super-Gaussian linear model”?
  - Wait until MAP estimation

Where Are We?

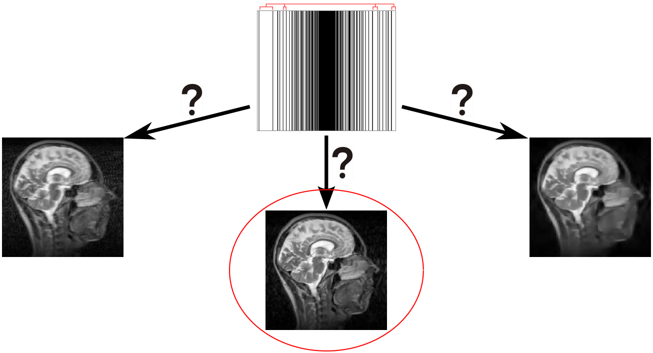
- Real-world signals are not Gaussian.
  Gaussian assumptions made for convenience
- Super-Gaussian distributions:
  Trade-off between realistic and tractable
- Latent Gaussian representations:
  - Gaussian scale mixtures
  - Super-Gaussian potentials (max representation)
- Group potentials: Structure your sparsity
- “Sparse” may mean super-Gaussian

Outline

Image Reconstruction



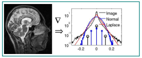
MAP Estimation



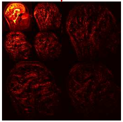
Maximum a Posteriori (MAP) Estimation

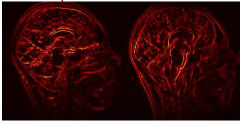
$$\boldsymbol{u}_* = \text{argmax}_{\boldsymbol{u}} P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})$$

Sparse Linear Model

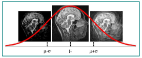


$$P(\mathbf{u}) \propto \prod_{i=1}^{q} t_i(s_i) = e^{-\tau_w \|\mathbf{B}_w \mathbf{u}\|_1} \times e^{-\tau_{tv} \|\mathbf{B}_{tv} \mathbf{u}\|_1}, \qquad \mathbf{s} = \mathbf{B}\mathbf{u}$$


wavelet


gradient

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



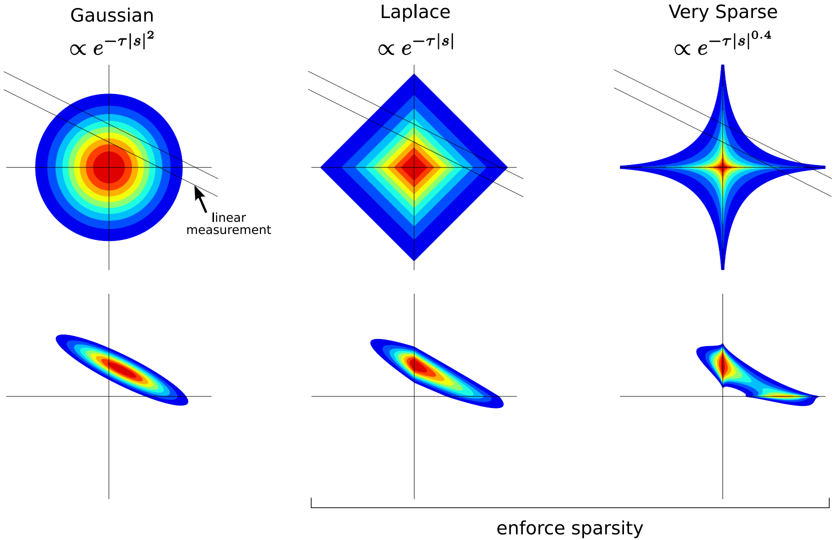
$$P(\mathbf{u}|\mathbf{y}) \propto P(\mathbf{u})P(\mathbf{y}|\mathbf{u})$$

MAP Estimation

$$\boldsymbol{u}_* = \operatorname{argmin}_{\boldsymbol{u}} \underbrace{\sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2}_{-2\log P(\boldsymbol{y}|\boldsymbol{u})} \underbrace{-2\sum_{i=1}^{q} \log t_i(s_i)}_{-\log P(\boldsymbol{u})}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}, \ \boldsymbol{y} \in \mathbb{C}^m$$

- MAP estimate is sparse.
  $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$: No more than m nonzero $s_{*,i}$ (if $|s_i| \mapsto -\log t_i(s_i)$ concave)

MAP Estimation

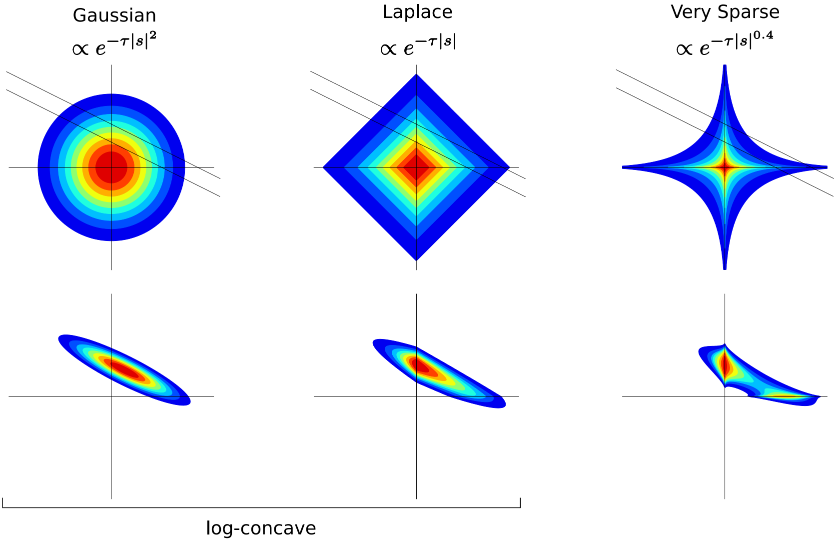


MAP Estimation

$$\boldsymbol{u}_* = \mathrm{argmin}_{\boldsymbol{u}} \underbrace{\sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2}_{-2\log P(\boldsymbol{y}|\boldsymbol{u})} \underbrace{-2\sum_{i=1}^{q} \log t_i(s_i)}_{-\log P(\boldsymbol{u})}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},\ \boldsymbol{y} \in \mathbb{C}^m$$

- MAP estimate is sparse.
  $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$: No more than m nonzero $s_{*,i}$ (if $|s_i| \mapsto -\log t_i(s_i)$ concave)
- MAP convex optimization problem $\Leftrightarrow$ $t_i(s_i)$ log-concave

Sparsity Priors



MAP Estimation

$$\boldsymbol{u}_* = \mathrm{argmin}_{\boldsymbol{u}} \underbrace{\sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2}_{-2\log P(\boldsymbol{y}|\boldsymbol{u})} \underbrace{-2\sum\nolimits_{i=1}^{q} \log t_i(s_i)}_{-\log P(\boldsymbol{u})}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},\ \boldsymbol{y} \in \mathbb{C}^m$$

- MAP estimate is sparse.
  $\boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$: No more than m nonzero $s_{*,i}$ (if $|s_i| \mapsto -\log t_i(s_i)$ concave)
- MAP convex optimization problem $\Leftrightarrow$ $t_i(s_i)$ log-concave
- Sparse and convex? Laplace potentials (Lasso) Tibshirani, JRSS-B 1996

Example: MAP Algorithm

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$

- Rewrite: Operator splitting.

Example: MAP Algorithm

$$\min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 \qquad \text{s.t. } \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$$

- Rewrite: Operator splitting.
  $\Rightarrow$ Update of each $\boldsymbol{u}$, $\boldsymbol{s}$ simple (ignoring constraint)

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \underbrace{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \lambda \boldsymbol{b}^T(\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s})}_{\text{Lagrangian}}$$

- Rewrite: Operator splitting.
  $\Rightarrow$ Update of each $\boldsymbol{u}$, $\boldsymbol{s}$ simple (ignoring constraint)
- Augmented Lagrangian technique ($\boldsymbol{b}$ Lagrange multipliers)

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \underbrace{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \lambda\boldsymbol{b}^T(\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s}) + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s}\|^2}_{\text{augmented Lagrangian}}$$

- Rewrite: Operator splitting.
  $\Rightarrow$ Update of each $\boldsymbol{u}$, $\boldsymbol{s}$ simple (ignoring constraint)
- Augmented Lagrangian technique ($\boldsymbol{b}$ Lagrange multipliers)

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \lambda \boldsymbol{b}^T(\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s}) + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s}\|^2$$

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2 - \frac{\lambda}{2}\|\boldsymbol{b}\|^2$$

Alternating Direction Methods of Multipliers (ADMM)

Iterate:

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2 - \frac{\lambda}{2}\|\boldsymbol{b}\|^2$$

Alternating Direction Methods of Multipliers (ADMM)

Iterate:

- Least squares projection (fixed $\boldsymbol{s}$, $\boldsymbol{b}$)

$$\boldsymbol{u} \leftarrow \operatorname{argmin} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2$$

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2 - \frac{\lambda}{2}\|\boldsymbol{b}\|^2$$

Alternating Direction Methods of Multipliers (ADMM)

Iterate:

- Least squares projection (fixed $\boldsymbol{s}$, $\boldsymbol{b}$)

$$\boldsymbol{u} \leftarrow \operatorname{argmin} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2$$

- Proximal map (fixed $\boldsymbol{u}$, $\boldsymbol{b}$)

$$\boldsymbol{s} \leftarrow \operatorname{argmin} \kappa\|\boldsymbol{s}\|_1 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - \boldsymbol{s} + \boldsymbol{b}\|^2$$

Example: MAP Algorithm

$$\max_{\boldsymbol{b}} \min_{\boldsymbol{u},\boldsymbol{s}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{s}\|_1 + \lambda\boldsymbol{b}^T(\boldsymbol{B}\boldsymbol{u}-\boldsymbol{s}) + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u}-\boldsymbol{s}\|^2$$

Alternating Direction Methods of Multipliers (ADMM)

Iterate:

- Least squares projection (fixed $\boldsymbol{s}$, $\boldsymbol{b}$)

$$\boldsymbol{u} \leftarrow \operatorname{argmin} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{u}\|^2 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u}-\boldsymbol{s}+\boldsymbol{b}\|^2$$

- Proximal map (fixed $\boldsymbol{u}$, $\boldsymbol{b}$)

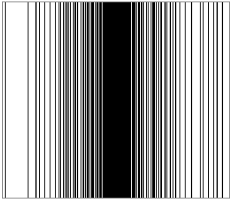
$$\boldsymbol{s} \leftarrow \operatorname{argmin} \kappa\|\boldsymbol{s}\|_1 + \frac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u}-\boldsymbol{s}+\boldsymbol{b}\|^2$$

- Lagrange multiplier update (fixed $\boldsymbol{u}$, $\boldsymbol{s}$)

$$\boldsymbol{b} \leftarrow \boldsymbol{b} + \boldsymbol{B}\boldsymbol{u} - \boldsymbol{s}$$

Example: MRI Reconstruction

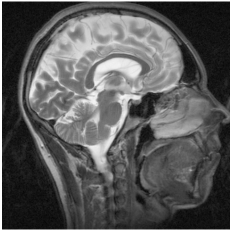
$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$



- $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT of size n, $J \subset \{1, \dots, n\}$
- Blocks of $\boldsymbol{B}$:
  Orthonormal (wavelets), FIR filters (Δ_x, Δ_y)

Example: MRI Reconstruction

$$\min_{\boldsymbol{u}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$

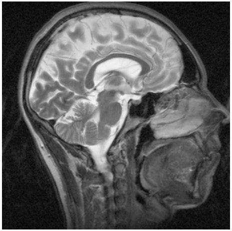


- $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT of size n, $J \subset \{1,\dots,n\}$
- Blocks of $\boldsymbol{B}$:
  Orthonormal (wavelets), FIR filters (Δ_x, Δ_y)
- Least squares projection:

$$\left(\boldsymbol{X}^H\boldsymbol{X} + \lambda\boldsymbol{B}^T\boldsymbol{B}\right)\boldsymbol{u} = \boldsymbol{r} := \boldsymbol{X}^H\boldsymbol{y} + \lambda\boldsymbol{B}^T(\boldsymbol{s} - \boldsymbol{b})$$

Example: MRI Reconstruction

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$

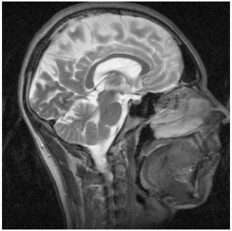


- $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT of size n, $J \subset \{1, \dots, n\}$
- Blocks of $\boldsymbol{B}$:
  Orthonormal (wavelets), FIR filters (Δ_x, Δ_y)
- Least squares projection:

$$\left(\boldsymbol{F}^T\boldsymbol{I}_{\cdot,J}\boldsymbol{I}_{J,\cdot}\boldsymbol{F} + \lambda\boldsymbol{B}^T\boldsymbol{B}\right)\boldsymbol{u} = \boldsymbol{r}$$

Example: MRI Reconstruction

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$

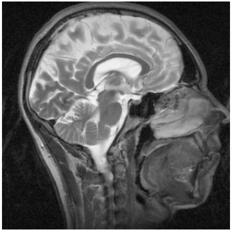


- $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT of size n, $J \subset \{1, \dots, n\}$
- Blocks of $\boldsymbol{B}$:
  Orthonormal (wavelets), FIR filters (Δ_x, Δ_y)
- Least squares projection:

$$\boldsymbol{F}^T\left(\boldsymbol{I}_{\cdot,J}\boldsymbol{I}_{J,\cdot} + \underbrace{\lambda\boldsymbol{F}\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{F}^T}_{\text{diagonal}}\right)\boldsymbol{F}\boldsymbol{u} = \boldsymbol{r}$$

Example: MRI Reconstruction

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$



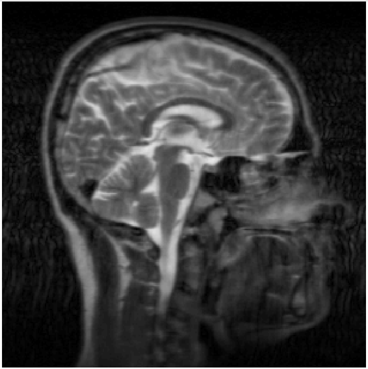
- $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT of size n, $J \subset \{1, \dots, n\}$
- Blocks of $\boldsymbol{B}$:
  Orthonormal (wavelets), FIR filters (Δ_x, Δ_y)
- Least squares projection:

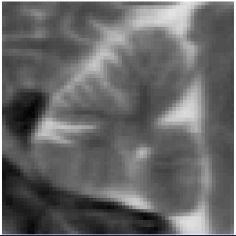
$$\underbrace{\left(\boldsymbol{I}_{\cdot,J}\boldsymbol{I}_{J,\cdot} + \boldsymbol{D}\right)}_{\text{diagonal}} \boldsymbol{F}\boldsymbol{u} = \boldsymbol{F}\boldsymbol{r}$$

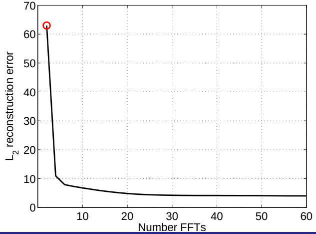
$\Rightarrow$ Two fast Fourier transforms only!

Example: MRI Reconstruction

courtesy Mateusz Malinowski

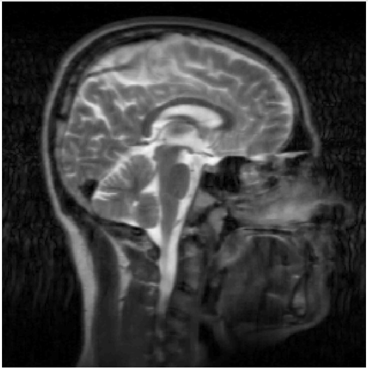


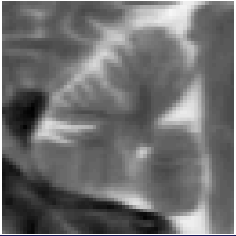


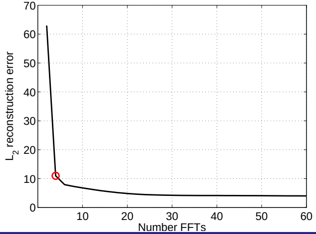


Example: MRI Reconstruction

courtesy Mateusz Malinowski

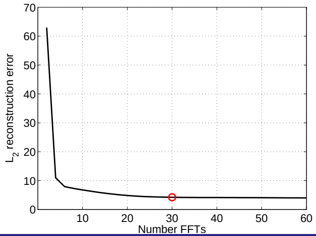






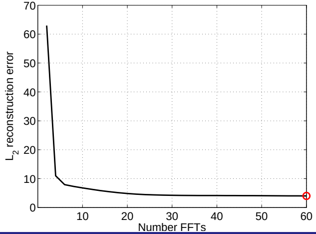
Example: MRI Reconstruction

courtesy Mateusz Malinowski



Example: MRI Reconstruction

courtesy Mateusz Malinowski

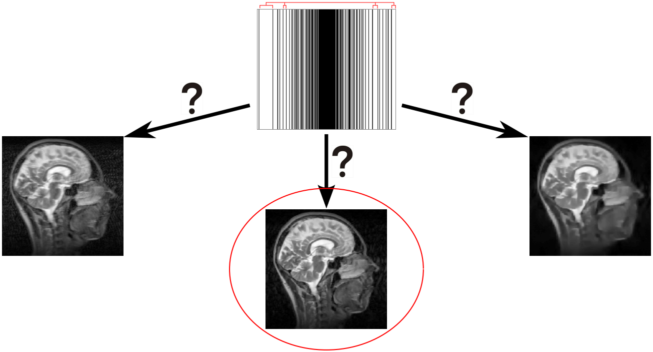


Where Are We?

- Sparse linear model:
  Linear couplings ($\boldsymbol{X}$, $\boldsymbol{B}$), super-Gaussian potentials
- MAP estimation:
  - Sparse solution if $|s_i| \mapsto -\log t_i(s_i)$ concave
  - Convex problem if $s_i \mapsto -\log t_i(s_i)$ convex (t_i log-concave)
  - Sparse and convex? Laplace potentials, ℓ_1
- Proximal splitting algorithms:
  Simple, efficient steps. Parallelizable

Outline

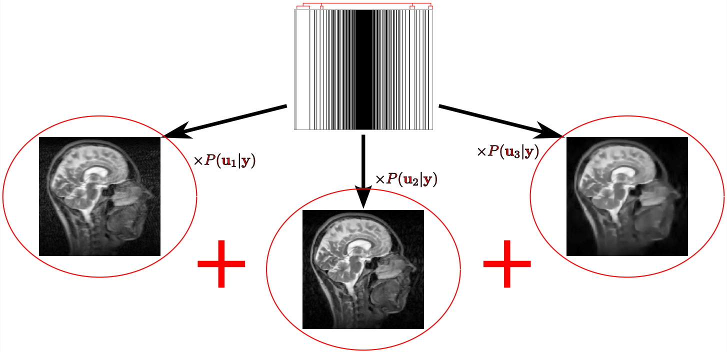
MAP Estimation



Maximum a Posteriori (MAP) Estimation

- There are many solutions. Why settle for any single one?

Integration, not Maximization



Use All Solutions

- Weight each solution by our uncertainty
- Average over them. Integrate, don't maximize

Robust Model Calibration

$$P(\boldsymbol{y}|\theta) = \int P(\boldsymbol{y}|\boldsymbol{u}, \theta) P(\boldsymbol{u}|\theta)\, d\boldsymbol{u}$$

Given raw data $\boldsymbol{y}$, no ground truth $\boldsymbol{u}$. Calibrate model parameters θ.

- Blind deconvolution (θ blur kernel)
- Multi-frame super-resolution (θ camera parameters, PSF)
- Image coding (θ codebook)
- Learning image priors ($P(\boldsymbol{u}) = P(\boldsymbol{u}|\theta)$)

Robust Model Calibration

$$P(\mathbf{y}|\boldsymbol{\theta}) = \int P(\mathbf{y}|\mathbf{u},\boldsymbol{\theta})P(\mathbf{u}|\boldsymbol{\theta})\,d\mathbf{u}$$

Given raw data $\mathbf{y}$, no ground truth $\mathbf{u}$. Calibrate model parameters $\boldsymbol{\theta}$.

MAP Estimation

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underbrace{\max_{\mathbf{u}} P(\mathbf{y}|\mathbf{u})P(\mathbf{u})}_{??}$$

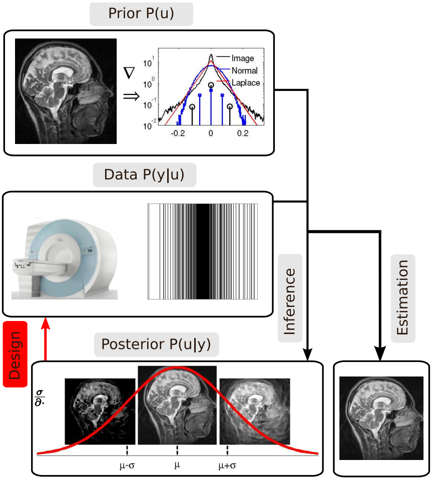
- All bets on one $\boldsymbol{\theta}$, all bets on one $\mathbf{u}$, ...
- Can work if
  - $\mathbf{u}$ much higher-D than $\boldsymbol{\theta}$
  - Additional engineering

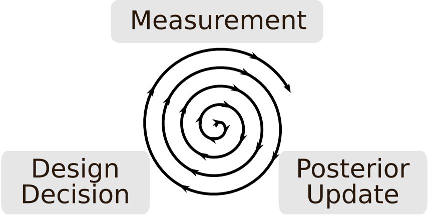
Bayesian Inference

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \underbrace{\int P(\mathbf{y}|\mathbf{u})P(\mathbf{u})\,d\mathbf{u}}_{\text{likelihood } P(\mathbf{y}|\boldsymbol{\theta})}$$

- Maximize true likelihood
- Account for uncertainty in $\mathbf{u}$: Cues for what $\boldsymbol{\theta}$ should be

Bayesian Experimental Design





Variational Bayesian Inference

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

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

Variational Bayesian Inference

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

- 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 Bayesian Inference

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

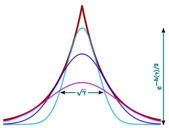
- Bayesian integration over $P(\boldsymbol{u}|\boldsymbol{y})$ intractable
- Integration tractable for Gaussians $Q(\boldsymbol{u}|\boldsymbol{y})$
  $\Rightarrow$ Approximate $P(\boldsymbol{u}|\boldsymbol{y})$ by $Q(\boldsymbol{u}|\boldsymbol{y})$!

Variational approximation

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

Super-Gaussian Priors

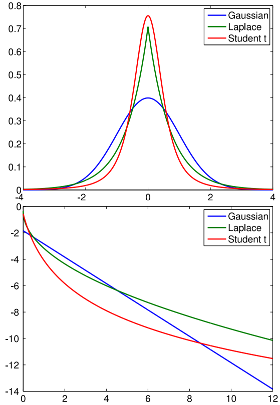
$$t(s) = \max_{\gamma \geq 0} e^{-\frac{1}{2}(s^2/\gamma + h(\gamma))}$$



Sparsity potentials are super-Gaussian

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

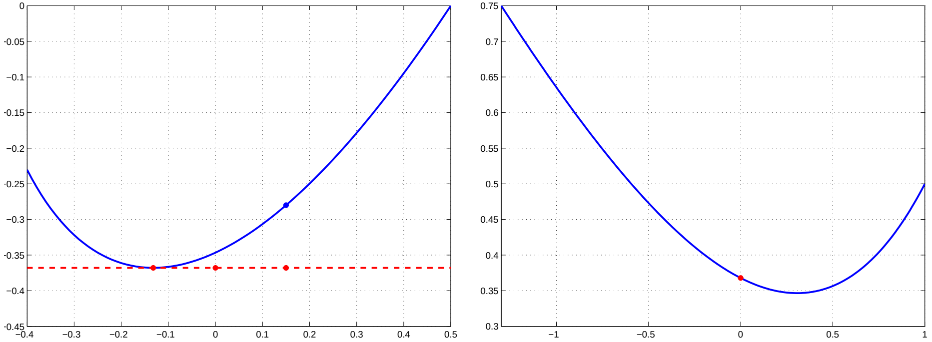
- Affine $\rightarrow$ convex:
  Shift mass to center and tails



Fenchel Duality

Super-Gaussian:

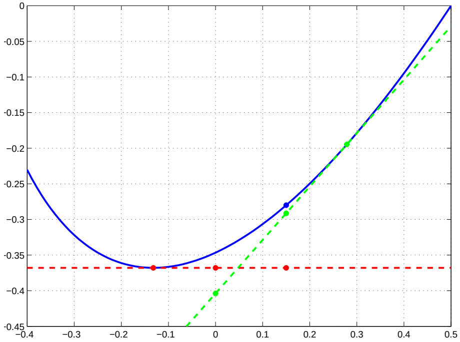
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2 \log t(s)\}$ convex.

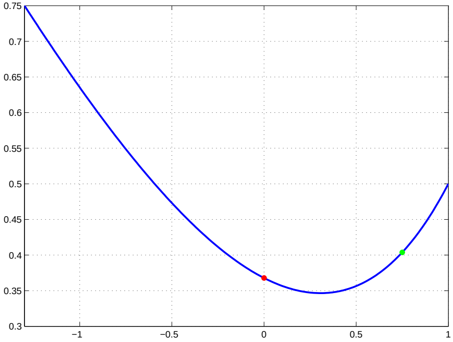


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

Fenchel Duality

Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.

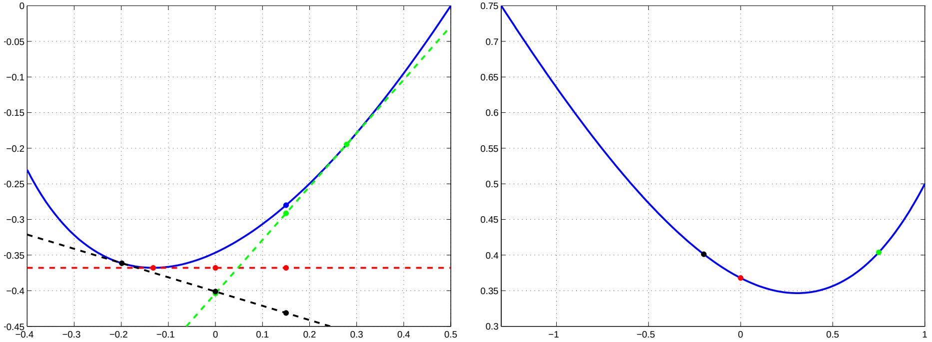




$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

Fenchel Duality

Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2 \log t(s)\}$ convex.

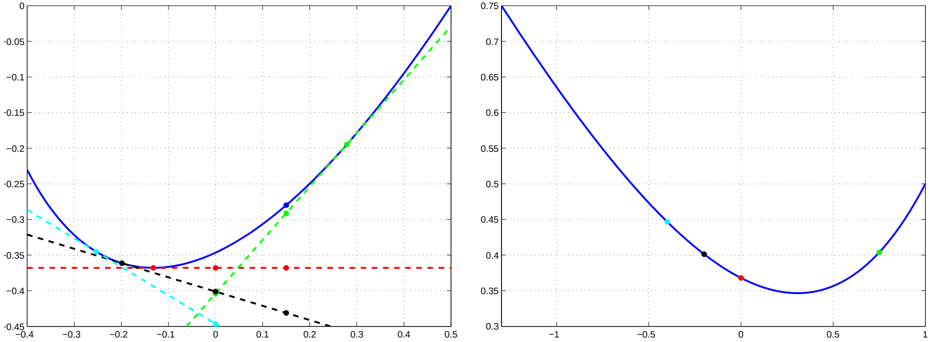


$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

Fenchel Duality

Super-Gaussian:

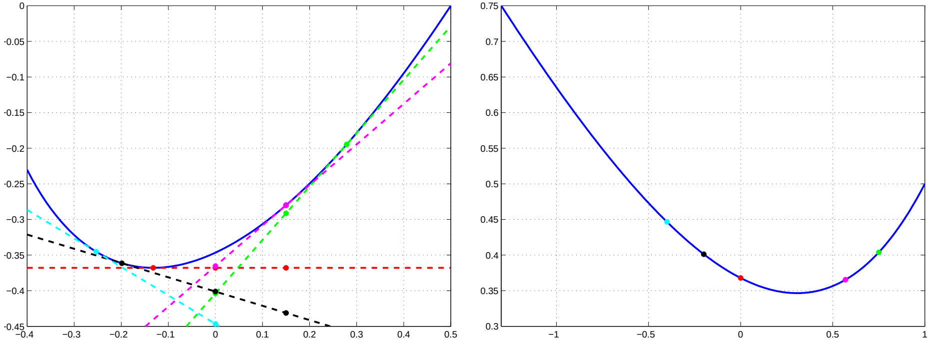
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.



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

Fenchel Duality

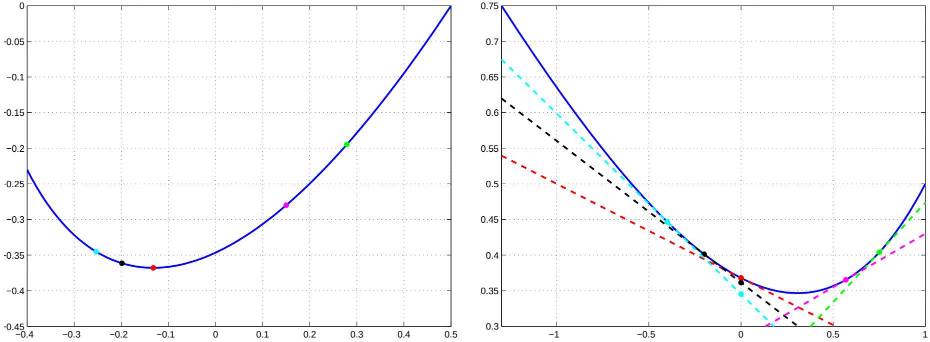
Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.



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

Fenchel Duality

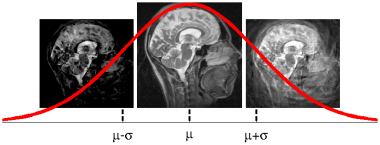
Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.



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

$$t(s) = \max_\gamma e^{-\frac{1}{2}(s^2/\gamma + h(\gamma))} \qquad h(\gamma) = \max_s -s^2/\gamma - 2\log t(s)$$

Super-Gaussian Bounding

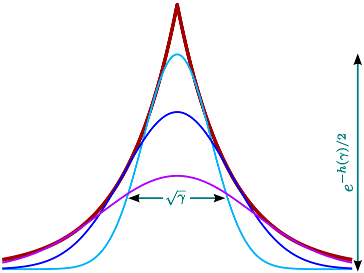


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

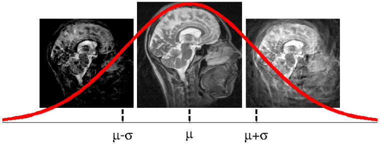
Sparsity potentials are super-Gaussian

$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-\frac{1}{2}(s_i^2/\gamma_i + h_i(\gamma_i))},$$

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



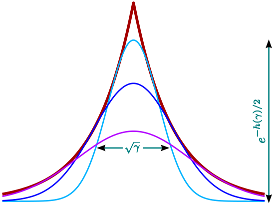
Super-Gaussian Bounding



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

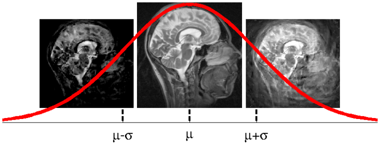
Exact representation

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



$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-\frac{1}{2}(s_i^2/\gamma_i + h_i(\gamma_i))}$$

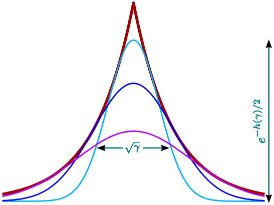
Super-Gaussian Bounding



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

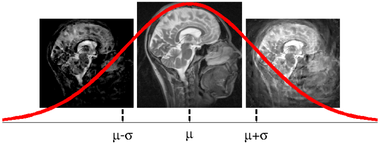
Lower bound

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



$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-\frac{1}{2}(s_i^2/\gamma_i + h_i(\gamma_i))}$$

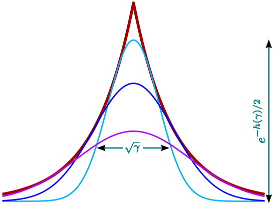
Super-Gaussian Bounding



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

Lower bound

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

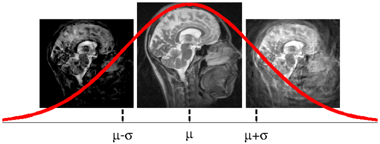


$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-\frac{1}{2}(s_i^2/\gamma_i + h_i(\gamma_i))}$$

Gaussian approximation

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

Super-Gaussian Bounding



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

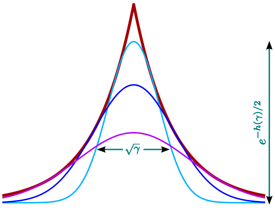
Variational problem: $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$

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

Gaussian approximation

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

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



$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-\frac{1}{2}(s_i^2/\gamma_i + h_i(\gamma_i))}$$

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

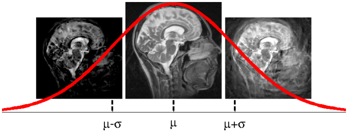
$$\log Z$$

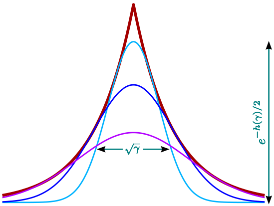
$$= \log \int N(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{u}, \sigma^2\boldsymbol{I}) \max_{\boldsymbol{\gamma}} e^{-(\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}+h(\boldsymbol{\gamma}))/2} \, d\boldsymbol{u}$$

$$\geq$$

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

Properties of Super-Gaussian Bounding





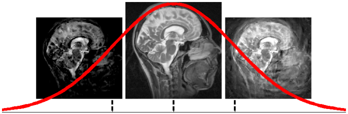
$$\min_{\gamma} -2\log \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u} + h(\gamma)$$

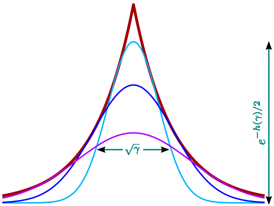
Super-Gaussian bounding stands out

Seeger, Nickisch, SIAM IS 2011

- Convex problem iff MAP estimation is convex
- Can be solved at much larger scales than others

Properties of Super-Gaussian Bounding





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

Super-Gaussian bounding stands out

Seeger, Nickisch, SIAM IS 2011

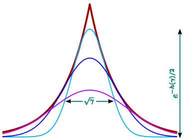
- Convex problem iff MAP estimation is convex
- Can be solved at much larger scales than others

MAP estimation will help solving it!

Towards Scalable Variational Inference

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

$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} = \sigma^{-2} \boldsymbol{X}^H \boldsymbol{X} + \boldsymbol{B}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{B}$$

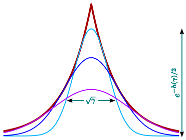


- Harder than MAP estimation. But why?

Towards Scalable Variational Inference

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

$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} = \sigma^{-2} \boldsymbol{X}^H \boldsymbol{X} + \boldsymbol{B}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{B}$$



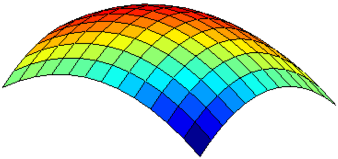
- Harder than MAP estimation. Because of $\log |\boldsymbol{A}|$.

Super-Gaussian bounding

$$\min_{\gamma,\boldsymbol{u}} \left\{ \phi(\boldsymbol{u}, \gamma) = \underbrace{\sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s} + h(\gamma)}_{\text{MAP criterion}} + \log |\boldsymbol{A}| \right\}$$

Decoupling by Fenchel Duality

$$\min_{\gamma,\boldsymbol{u}_*} \phi(\boldsymbol{u}_*,\gamma) = \min_{\gamma,\boldsymbol{u}_*} \underbrace{\log|\boldsymbol{A}(\gamma^{-1})|}_{\text{concave}} + \underbrace{\phi_\cup(\boldsymbol{u}_*,\gamma)}_{\text{convex}}$$

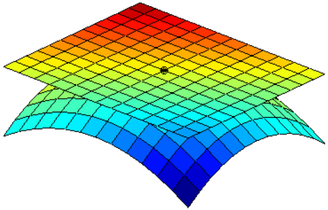


Decoupling by Fenchel Duality

$$\min_{\gamma, \boldsymbol{u}_*} \phi(\boldsymbol{u}_*, \gamma) = \min_{\gamma, \boldsymbol{u}_*} \underbrace{\log |\boldsymbol{A}(\gamma^{-1})|}_{\text{concave}} + \underbrace{\phi_\cup(\boldsymbol{u}_*, \gamma)}_{\text{convex}}$$

Fenchel duality

$$\log |\boldsymbol{A}(\gamma^{-1})| = \min_{\boldsymbol{z}} \boldsymbol{z}^T(\gamma^{-1}) - g^*(\boldsymbol{z})$$

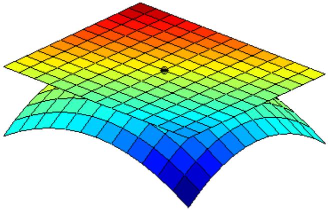


Decoupling by Fenchel Duality

$$\log|\boldsymbol{A}(\gamma^{-1})| + \phi_\cup(\boldsymbol{u}_*,\gamma) = \min_{\boldsymbol{z}} \underbrace{\boldsymbol{z}^T(\gamma^{-1}) + \phi_\cup(\boldsymbol{u}_*,\gamma) - g^*(\boldsymbol{z})}_{\phi_{\boldsymbol{z}}(\boldsymbol{u}_*,\gamma)\ \text{(convex, decoupled)}}$$

Fenchel duality

$$\log|\boldsymbol{A}(\gamma^{-1})| = \min_{\boldsymbol{z}} \boldsymbol{z}^T(\gamma^{-1}) - g^*(\boldsymbol{z})$$



Scalable Double Loop Algorithm

Double loop algorithm

Seeger *et.al.*, NIPS 2009; insp. by Wipf *et.al.*, NIPS 2008

- Inner loop optimization: $\min_{\gamma} \min_{\boldsymbol{u}_*} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) + g^*(\boldsymbol{z})$ [fixed $\boldsymbol{z}$]

$$\min_{\boldsymbol{u}_*} \min_{\gamma} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X} \boldsymbol{u}_*\|^2 + \boldsymbol{z}^T(\gamma^{-1}) + \boldsymbol{s}_*^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}_* + h(\gamma)$$

Scalable Double Loop Algorithm

Double loop algorithm

Seeger *et.al.*, NIPS 2009; insp. by Wipf *et.al.*, NIPS 2008

- Inner loop optimization: $\min_\gamma \min_{\boldsymbol{u}_*} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) + g^*(\boldsymbol{z})$ [fixed $\boldsymbol{z}$]
  Smoothed MAP Reconstruction

$$\min_{\boldsymbol{u}_*} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - 2\sum_{i=1}^{q} \log t_i\left(\sqrt{z_i + s_{*i}^2}\right), \quad z_i > 0$$

Scalable Double Loop Algorithm

Double loop algorithm

Seeger *et.al.*, NIPS 2009; insp. by Wipf *et.al.*, NIPS 2008

- Inner loop optimization: $\min_{\gamma} \min_{\boldsymbol{u}_*} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) + g^*(\boldsymbol{z})$ [fixed $\boldsymbol{z}$]
  Smoothed MAP Reconstruction
- Outer loop update: $\min_{\boldsymbol{z}} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma)$ [fixed $(\boldsymbol{u}_*, \gamma)$]

$$\text{Tangent}: \ \boldsymbol{z} \leftarrow \nabla_{\gamma^{-1}} \log |\boldsymbol{A}|, \quad \boldsymbol{A} = \sigma^{-2} \boldsymbol{X}^H \boldsymbol{X} + \boldsymbol{B}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{B}$$

Scalable Double Loop Algorithm

Double loop algorithm

Seeger *et.al.*, NIPS 2009; insp. by Wipf *et.al.*, NIPS 2008

- Inner loop optimization: $\min_\gamma \min_{\boldsymbol{u}_*} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) + g^*(\boldsymbol{z})$ [fixed $\boldsymbol{z}$]
  Smoothed MAP Reconstruction
- Outer loop update: $\min_{\boldsymbol{z}} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma)$ [fixed $(\boldsymbol{u}_*, \gamma)$]
  Gaussian (Co)Variances

$$\boldsymbol{z} \leftarrow \nabla_{\gamma^{-1}} \log|\boldsymbol{A}| = \mathrm{diag}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = (\mathrm{Var}_Q[s_i|\boldsymbol{y}])$$

Reductions

Computational primitives driving large scale inference

1. Penalized least squares ($\approx$ MAP estimation)

$$\min_{\boldsymbol{u}_*} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - 2\sum_{i=1}^{q} \log t_i\left(\sqrt{z_i + s_{*i}^2}\right)$$

- MAP special case: $z_i = 0$
- Scalable algorithms (thanks to MAP “gold rush”)

Reductions

Computational primitives driving large scale inference

1. Penalized least squares ($\approx$ MAP estimation)

$$\min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 - 2\sum_{i=1}^{q} \log t_i\left(\sqrt{z_i + s_{*i}^2}\right)$$

   - MAP special case: $z_i = 0$
   - Scalable algorithms (thanks to MAP "gold rush")

2. Gaussian variances

$$\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T), \quad \boldsymbol{A} = \sigma^{-2}\boldsymbol{X}^H\boldsymbol{X} + \boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}$$

   - More difficult
   - Methods from numerical mathematics, spatial statistics

Where Are We?

- Bayesian inference: Optimization over distributions.
  Variational approximations: Relaxations thereof
- Super-Gaussian bounding:
  Exploit latent Gaussian representations of t_i
  - Convex iff MAP estimation is convex
  - Scalable by reductions (double loop algorithm)

Where Are We?

- Bayesian inference: Optimization over distributions.
  Variational approximations: Relaxations thereof
- Super-Gaussian bounding:
  Exploit latent Gaussian representations of t_i
  - Convex iff MAP estimation is convex
  - Scalable by reductions (double loop algorithm)
- Other relaxations available
- General variational inference
- Randomized approximations
  (MCMC, brief Gibbs sampling)

Seeger, J. Phys. Conf. 2009
Nickisch *et.al.*, JMLR 2008

Wainwright, Jordan, FTML 2008

Teh *et.al.*, JMLR 2003
Roth, Black, IJCV 2009

Outline

1. Sparse Modelling
2. Sparse Estimation
3. Sparse Bayesian Inference
4. Sparse Estimation vs. Sparse Inference

Questions

- When do I get exact zeros?
- Why is sparse inference more expensive than sparse estimation?
- Can I drive Bayesian experimental design with sparse estimation (RVM/ARD)?

Automatic Relevance Determination

Tipping 2001,
Wipf *et.al.* 2004

$$\min_{\gamma} \left\{ \phi_{\text{ARD}}(\gamma) = -2 \log \int P(\boldsymbol{y}|\boldsymbol{u}) N(\boldsymbol{s}|\boldsymbol{0}, \boldsymbol{\Gamma})\, d\boldsymbol{u} \right\}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$$

- ARD $\leftrightarrow$ Relevance Vector Machine
- Sparsity by $\gamma_i \to 0$
- Sparse estimation, not sparse inference: Seeger, Wipf, IEEE SPM 2010
  Zero-temperature limit of variational inference with Student t prior
- Algorithms:
  - Sequential greedy — Tipping, Faul, AISTATS 2003
  - Double loop (reweighted ℓ_1) — Wipf *et.al.*, NIPS 2008

Exact Sparsity Kills Posterior Uncertainty

$$\gamma_i = 0 \quad \Rightarrow \quad \mathrm{E}_Q[s_i^2 | \boldsymbol{y}] = 0$$

- Exact sparsity controlled by γ_i

Exact Sparsity Kills Posterior Uncertainty

$$\gamma_i = 0 \quad \Rightarrow \quad \mathrm{E}_Q[s_i^2 | \boldsymbol{y}] = 0$$

- Exact sparsity controlled by γ_i
- Exact sparsity kills posterior uncertainty:

$$\boldsymbol{\gamma}_J = \mathbf{0} \quad \Rightarrow \quad \mathrm{E}_Q[\boldsymbol{s}_J | \boldsymbol{y}] = \mathbf{0}, \;\; \mathrm{Cov}_Q[\boldsymbol{s}_J | \boldsymbol{y}] = \mathbf{0}$$

Exact Sparsity Kills Posterior Uncertainty

$$\gamma_i = 0 \quad \Rightarrow \quad \mathrm{E}_Q[s_i^2 | \boldsymbol{y}] = 0$$

- Exact sparsity controlled by γ_i
- Exact sparsity kills posterior uncertainty:

$$\gamma_J = \mathbf{0} \quad \Rightarrow \quad \mathrm{E}_Q[\boldsymbol{s}_J | \boldsymbol{y}] = \mathbf{0}, \;\; \mathrm{Cov}_Q[\boldsymbol{s}_J | \boldsymbol{y}] = \mathbf{0}$$

- Good for computation: Heavy scaling only in nonzeros $\|\gamma\|_0$

Exact Sparsity Kills Posterior Uncertainty

$$\gamma_i = 0 \quad \Rightarrow \quad \mathrm{E}_Q[s_i^2|\boldsymbol{y}] = 0$$

- Exact sparsity controlled by γ_i
- Exact sparsity kills posterior uncertainty:

$$\gamma_J = \mathbf{0} \quad \Rightarrow \quad \mathrm{E}_Q[\boldsymbol{s}_J|\boldsymbol{y}] = \mathbf{0}, \;\; \mathrm{Cov}_Q[\boldsymbol{s}_J|\boldsymbol{y}] = \mathbf{0}$$

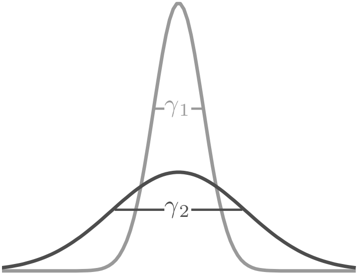
- Good for computation: Heavy scaling only in nonzeros $\|\gamma\|_0$
- Bad for Bayesian inference:
  Uncertainty is eliminated (the more so, the less data!)
  - Bayesian experimental design:
    Cannot be based on sparse estimation

Ji, Carin, ICML 2007

Sparse Estimation vs. Sparse Inference

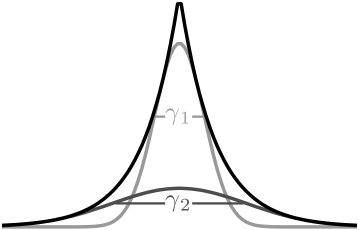
Sparse Estimation

$$\phi_{\text{ARD}} = -2\times$$
$$\log \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{N(\boldsymbol{s}|\boldsymbol{0}, \boldsymbol{\Gamma})}_{\text{normalized}} \, d\boldsymbol{u}$$



Sparse Inference

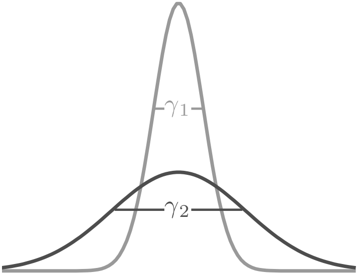
$$\phi_{\text{SGB}} = -2\times$$
$$\log \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{e^{-\frac{1}{2}((\boldsymbol{s}^2)^T\gamma^{-1}+h(\gamma))}}_{\text{lower bound}} \, d\boldsymbol{u}$$



Sparse Estimation vs. Sparse Inference

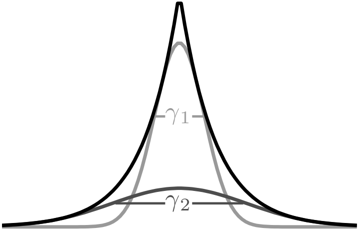
Sparse Estimation

- Encourages $\gamma_i \to 0$



Sparse Inference

- Forbids $\gamma_i \to 0$ $(\phi_{\text{SGB}} \to \infty)$



Answers

Seeger, Nickisch, SIAM IS 2001

- When do I get exact zeros?
  Controlled by $\gamma_i \to 0$.
  Happens for sparse estimation, not for sparse inference

Answers

Seeger, Nickisch, SIAM IS 2001

- When do I get exact zeros?
  Controlled by $\gamma_i \to 0$.
  Happens for sparse estimation, not for sparse inference
- Why is sparse inference more expensive than sparse estimation?
  Because sparse inference maintains posterior uncertainty
  (full covariance)

Answers

- When do I get exact zeros?
  Controlled by $\gamma_i \to 0$.
  Happens for sparse estimation, not for sparse inference
- Why is sparse inference more expensive than sparse estimation?
  Because sparse inference maintains posterior uncertainty
  (full covariance)
- Can I drive Bayesian experimental design with sparse estimation?
  No free lunch! Sparse estimation kills posterior uncertainty
  (highly degenerate covariance)

Conclusions

- Variational Bayesian inference very active field
  - Loopy belief propagation and generalizations
  - Convex relaxations. LP relaxations
  - Gaussian/discrete Markov random fields

Wainwright, Jordan
FTML 2008

Conclusions

- Variational Bayesian inference very active field — Wainwright, Jordan FTML 2008
    - Loopy belief propagation and generalizations
    - Convex relaxations. LP relaxations
    - Gaussian/discrete Markov random fields
- Broad application impact
    - Coding, information transmission
    - Expert systems
    - Low level computer vision, adaptive robotics and control
    - Discrete optimization — Mézard, Montanari, 2009
    - Geostatistics, spatial modelling

Conclusions

- Sparse inference beyond MAP estimation
  - Robust reconstruction
  - Active, adaptive data acquisition
  - Learning for inverse problems
  - Sequential decision-making

Conclusions

- Sparse inference beyond MAP estimation
  - Robust reconstruction
  - Active, adaptive data acquisition
  - Learning for inverse problems
  - Sequential decision-making
- Bayesian experimental design, Bayesian optimization
  - Medical imaging sampling optimization
  - Computational photography
  - Intelligent user interfaces
  - Active calibration of cameras

Conclusions

- Sparse inference beyond MAP estimation
  - Robust reconstruction
  - Active, adaptive data acquisition
  - Learning for inverse problems
  - Sequential decision-making
- Bayesian experimental design, Bayesian optimization
  - Medical imaging sampling optimization
  - Computational photography
  - Intelligent user interfaces
  - Active calibration of cameras
- Modern variational inference algorithms:
  Layers on top of what you already know
  - Penalized least squares (MAP) reconstruction
  - Gaussian covariance approximation (PCA)

Software and Acknowledgments

glm-ie: Toolbox by Hannes Nickisch

`mloss.org/software/view/269/`

- Generalized sparse linear models
- MAP reconstruction and variational Bayesian inference (double loop algorithm for super-Gaussian bounding)
- Matlab 7.x, GNU Octave 3.2.x

- Hannes Nickisch (Philips Hamburg, ex-MPI Tübingen)
- Rolf Pohmann, Bernhard Schölkopf (MPI Tübingen)
- David Wipf (MSR China)