The Abdus Salam International Centre for Theoretical Physics

2361-5

School on Large Scale Problems in Machine Learning and Workshop on Common Concepts in Machine Learning and Statistical Physics

20 - 31 August 2012

Large Scale Variational Bayesian Inference for Continuous Variable Models - Lecture Notes

Matthias SEEGER

Laboratory for Probabilistic Machine Learning, EPFL, CH-1015 Lausanne Switzerland

Strada Costiera, 11 - 34151 - Trieste - Italy • Tel. +39 0402240111 • Fax. +39 040224163 • sci_info@ictp.it • www.ictp.it

ICTP is governed by UNESCO, IAEA, and Italy, and it is a UNESCO Category 1 Institute

Large Scale Variational Bayesian Inference for Continuous Variable Models

Matthias Seeger

Laboratory for Probabilistic Machine Learning
Ecole Polytechnique Fédérale de Lausanne

http://lapmal.epfl.ch/

Outline

1. Motivation
2. Variational Inference Relaxations
   - Super-Gaussian Bounding
   - Expectation Propagation
   - Gaussian KL Minimization
   - Conjugate Gradients Algorithm
3. Scalable Variational Inference
   - Scaling up Super-Gaussian Bounding
   - Penalized Least Squares
   - Gaussian Variances
4. Application Example

Outline

Goals of Lecture

- Beyond point estimation:
  Bayesian inference for non-Gaussian continuous variable models
- Beyond message passing:
  Computational structure of variational inference relaxations
- The layer below:
  Scalability through reductions to convex optimization and numerical mathematics

Image Reconstruction

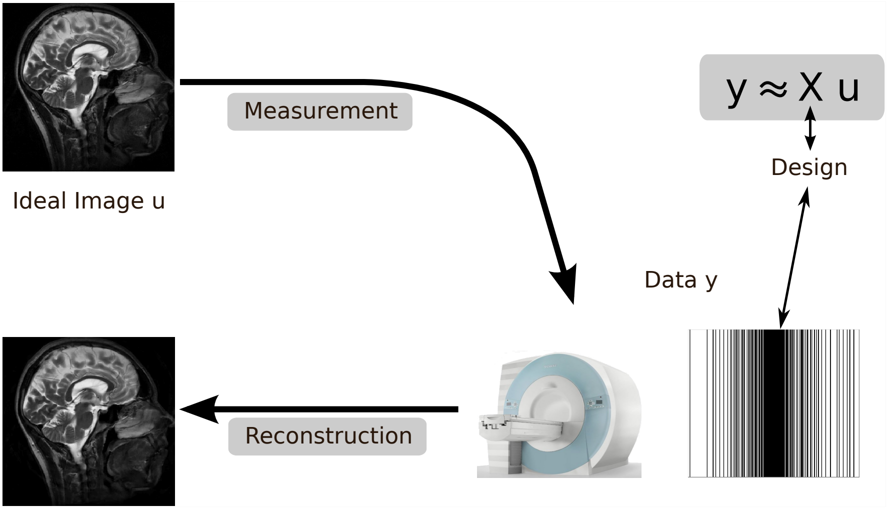
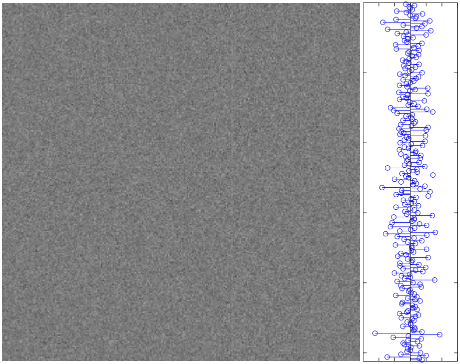
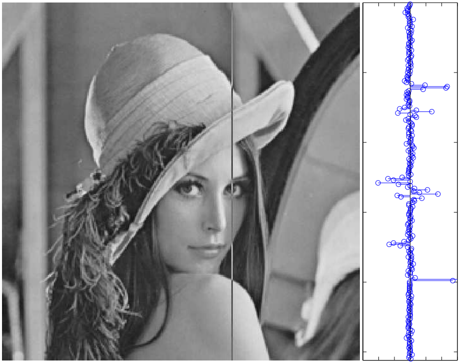


Image Statistics

Whatever images are . . .

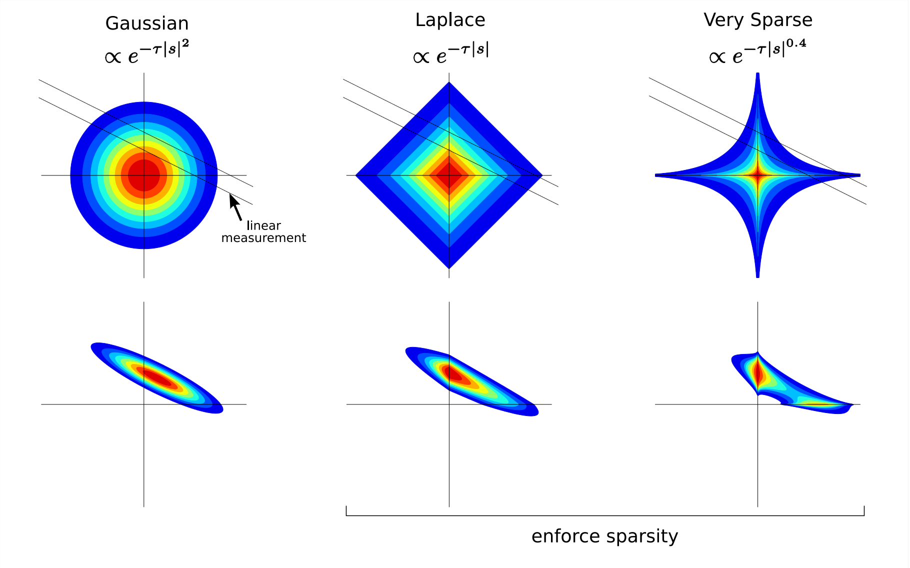
they are not Gaussian!





Sparsity Priors

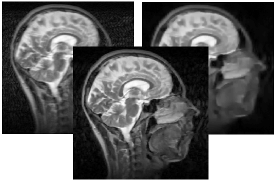
courtesy Florian Steinke

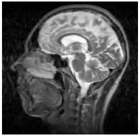


Posterior Distribution

- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit

$$P(\boldsymbol{y}|\boldsymbol{u})$$

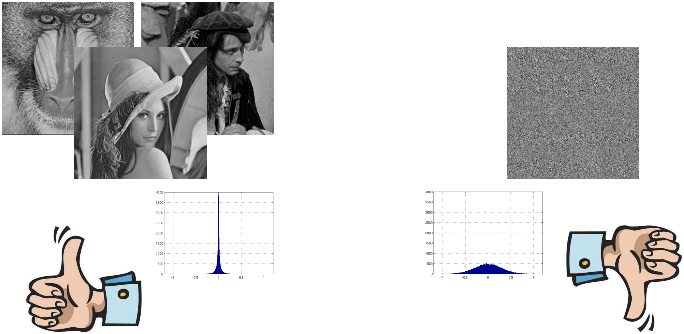




Posterior Distribution

- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit
- Prior $P(\boldsymbol{u})$: Signal properties

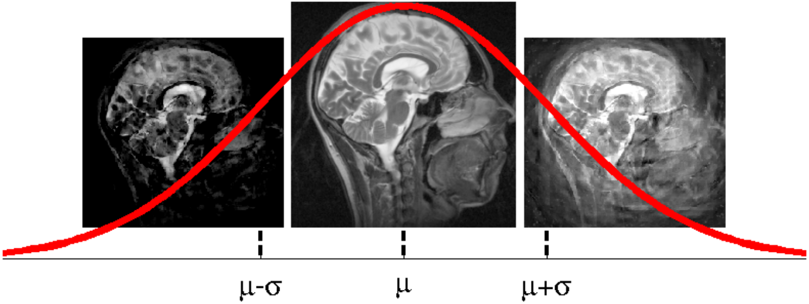
$$P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})$$



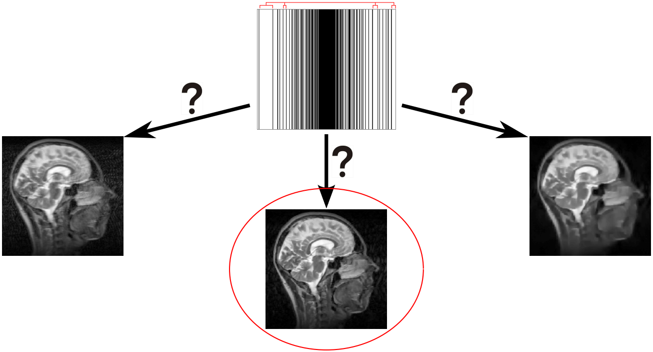
Posterior Distribution

- Likelihood $P(\boldsymbol{y}|\boldsymbol{u})$: Data fit
- Prior $P(\boldsymbol{u})$: Signal properties
- Posterior distribution $P(\boldsymbol{u}|\boldsymbol{y})$: Consistent information summary

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



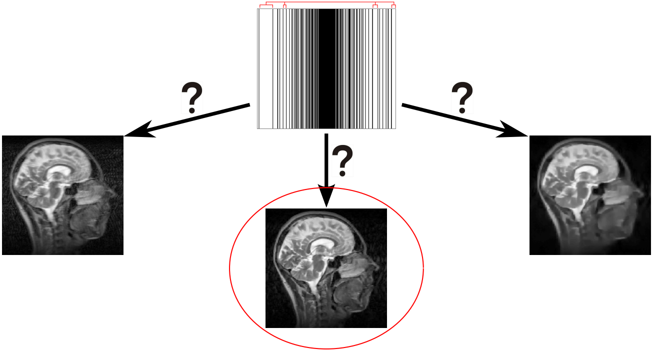
MAP Estimation



Maximum a Posteriori (MAP) Estimation

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

Why Move Beyond MAP?

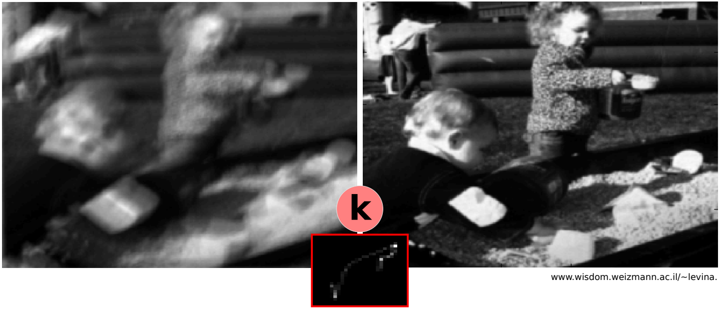


Maximum a Posteriori (MAP) Estimation

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

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 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 $\boldsymbol{u}$. Estimate model parameters θ.

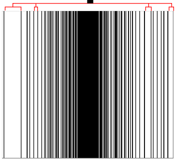
- Blind deconvolution (θ blur kernel)
- Multi-frame super-resolution (θ camera parameters, PSF)
- Image coding (θ codebook)
- Learning image priors ($P(\boldsymbol{u}) = P(\boldsymbol{u}|\theta)$)

Bayesian Experimental Design

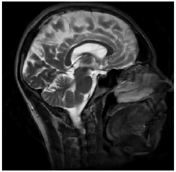
scan time $\propto$
phase encodes

$$y \approx X\, u$$

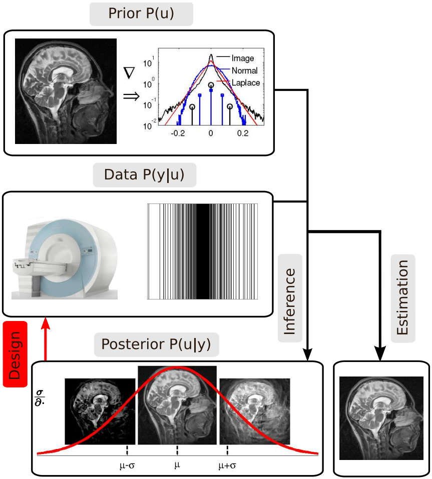
$$\boldsymbol{X} \leftarrow ?$$

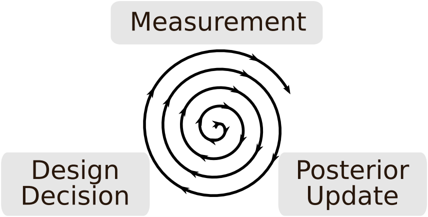


Reconstruction

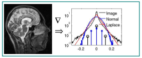


Bayesian Experimental Design

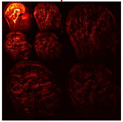




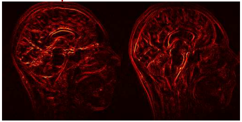
Sparse Linear Model



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

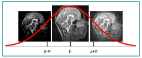


wavelet



gradient

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

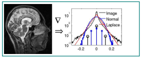


$$P(\mathbf{u}|\mathbf{y}) \propto P(\mathbf{u})P(\mathbf{y}|\mathbf{u})$$

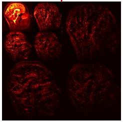
- $\boldsymbol{X}$, $\boldsymbol{B}$? Fast operators of your choice ($\boldsymbol{X}$ dictated by application)

  Denoising: $\boldsymbol{X}$ diagonal
  Deconvolution: $\boldsymbol{X}\boldsymbol{u} = \boldsymbol{k} \otimes \boldsymbol{u}$
  MRI reconstruction: $\boldsymbol{X} = \boldsymbol{I}_{J,\cdot}\boldsymbol{F}$, $\boldsymbol{F}$ DFT

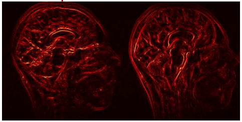
Sparse Linear Model



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

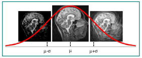


wavelet



gradient

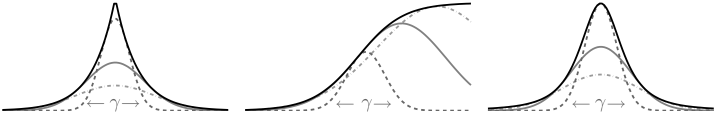
$P(\mathbf{y}|\mathbf{u}) = \mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{u}, \sigma^2 \mathbf{I})$



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

- $\boldsymbol{X}$, $\boldsymbol{B}$? Fast operators of your choice ($\boldsymbol{X}$ dictated by application)
- $t_i(s_i)$ Laplace here, but many other options

Nickisch, Seeger, ICML 2009



Outline

1. Motivation
2. Variational Inference Relaxations
   - Super-Gaussian Bounding
   - Expectation Propagation
   - Gaussian KL Minimization
   - Conjugate Gradients Algorithm
3. Scalable Variational Inference
   - Scaling up Super-Gaussian Bounding
   - Penalized Least Squares
   - Gaussian Variances
4. Application Example

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 $\log Z$

The Log Partition Function

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

Master function $\log Z$? Why this target?

- Physicist: Of course, it's the (negative) free energy!
- Probabilist: It generates posterior moments (cumulants)
- Variational definition of posterior distribution

$$\mathrm{E}_{Q(\boldsymbol{u}|\boldsymbol{y})}\left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_i t_i(s_i)}{Q(\boldsymbol{u}|\boldsymbol{y})}\right] \quad \begin{cases} \mathrm{argmax}_{Q(\boldsymbol{u}|\boldsymbol{y})}? & P(\boldsymbol{u}|\boldsymbol{y}) \\ \max_{Q(\boldsymbol{u}|\boldsymbol{y})}? & \log Z \end{cases}$$

- Bayesian inference: Optimization over distributions

Wainwright, Jordan, FTML 2008

Variational Approximations

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

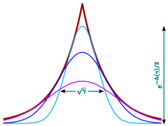
Variational approximation

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

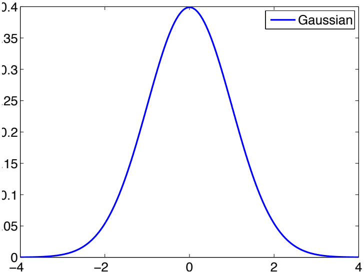
- Super-Gaussian bounding
- Expectation propagation
- Gaussian KL minimization

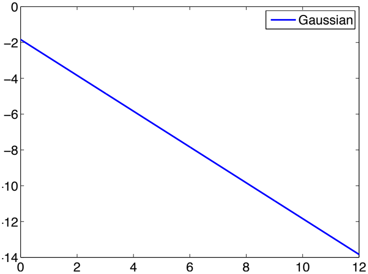
Super-Gaussian Potentials

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



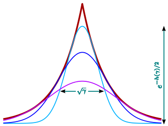
- $t(s)$ even and positive: Let's look at $s^2 \mapsto 2 \log t(s)$
- What's that for a Gaussian $t(s) = N(s|0, \sigma^2)$?
  A linear (affine) function





Super-Gaussian Potentials

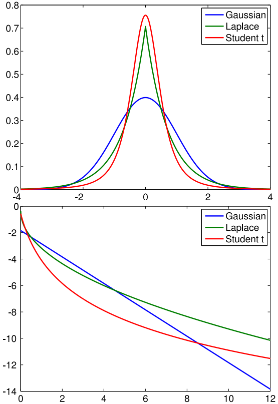
$$t(s) = \max_{\gamma \geq 0} e^{-s^2/(2\gamma)} e^{-h(\gamma)/2}$$



Sparsity potentials are super-Gaussian

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

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



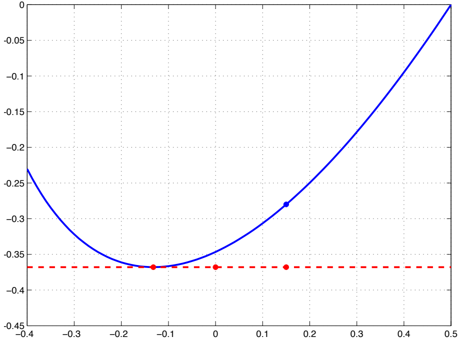
Convex (Fenchel) Duality

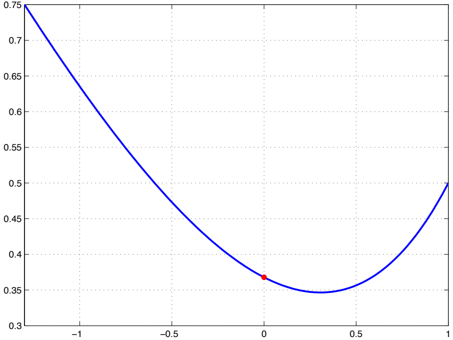
Super-Gaussian:
$t(s)$ even, $s^2 \mapsto 2\log t(s)$ convex.

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

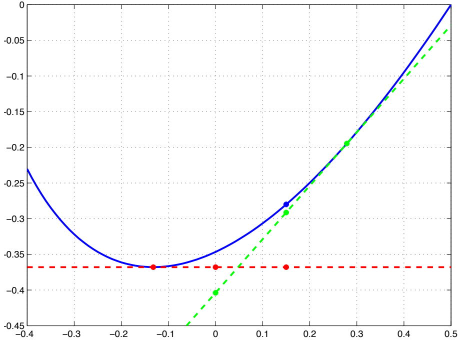


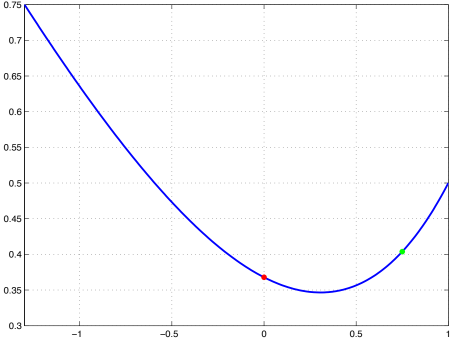


$$f(x) = \max_{\pi} x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

Convex (Fenchel) Duality

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

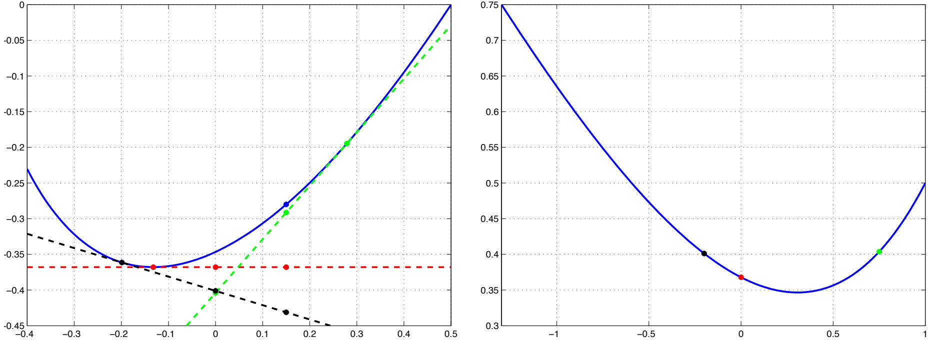




$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

Convex (Fenchel) Duality

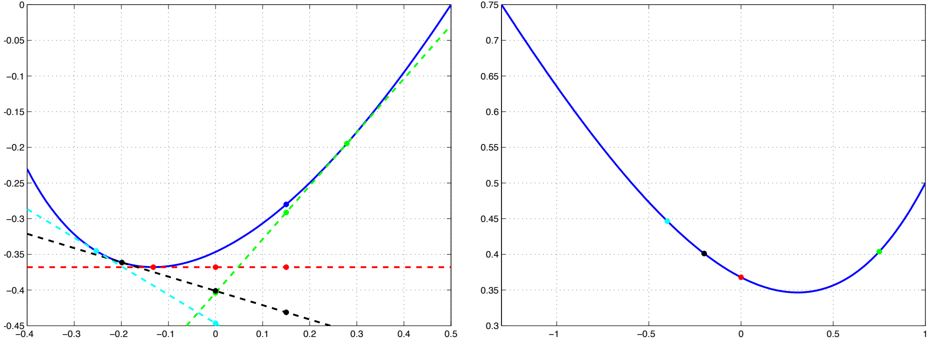
Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.



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

Convex (Fenchel) Duality

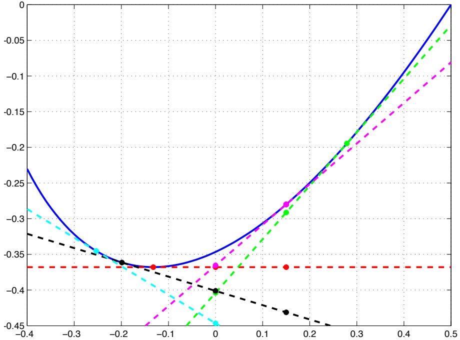
Super-Gaussian:
$t(s)$ even, $\{x = s^2\} \mapsto \{f(x) = 2\log t(s)\}$ convex.

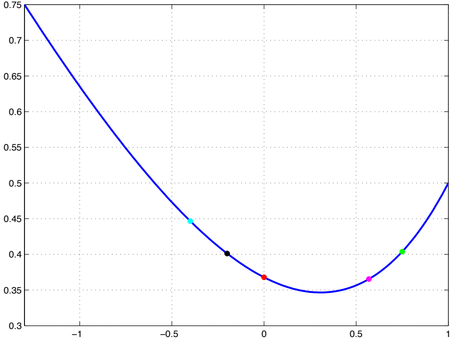


$f(x) = \max_{\pi} x\pi - f^*(\pi)$ $\quad$ $f^*(\pi) = \max_x \pi x - f(x)$

Convex (Fenchel) Duality

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





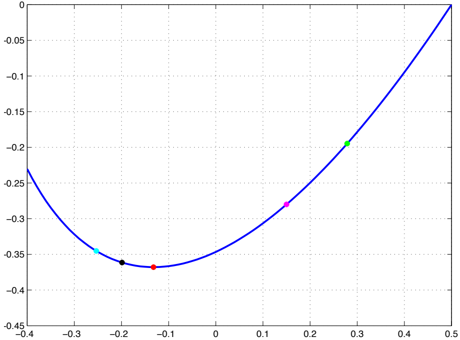
$$f(x) = \max_\pi x\pi - f^*(\pi)$$

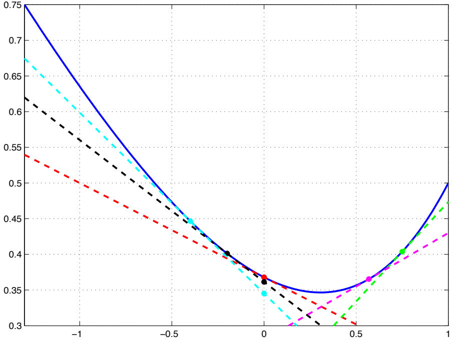
$$f^*(\pi) = \max_x \pi x - f(x)$$

Convex (Fenchel) Duality

Super-Gaussian: F1

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

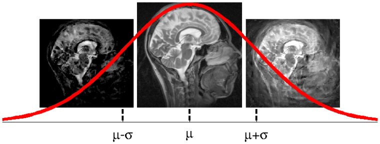




$$f(x) = \max_\pi x\pi - f^*(\pi) \qquad f^*(\pi) = \max_x \pi x - f(x)$$

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

Super-Gaussian Potentials



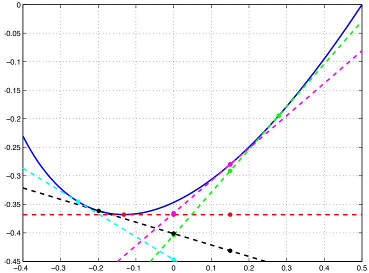
$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u}) \times P(\boldsymbol{u})}{P(\boldsymbol{y})}$$

Sparsity potentials are super-Gaussian

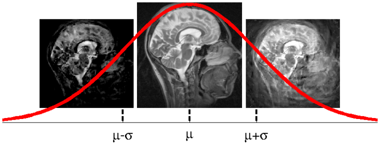
$$s_i^2 \mapsto 2\log t_i(s_i) \quad \text{is convex}$$

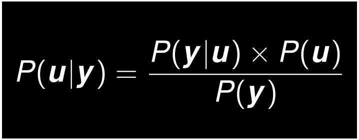
Convex (Fenchel) duality

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



Super-Gaussian Bounding

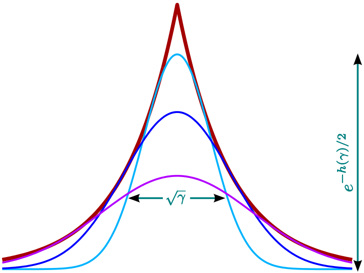




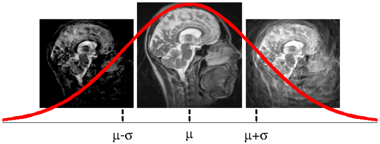
Sparsity potentials are super-Gaussian

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

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



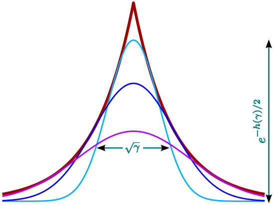
Super-Gaussian Bounding



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

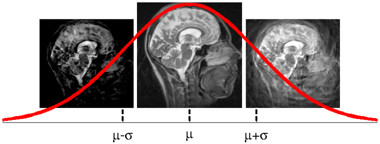
Exact representation

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



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

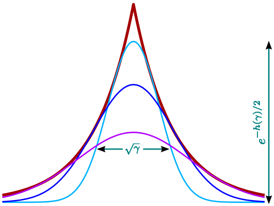
Super-Gaussian Bounding



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

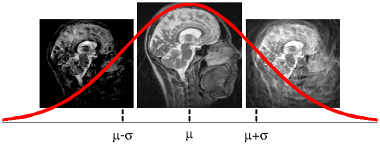
Lower bound

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



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

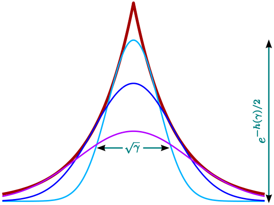
Super-Gaussian Bounding



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

Lower bound

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

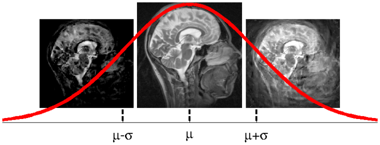


$$t_i(s_i) = \max_{\gamma_i \geq 0} e^{-s_i^2/(2\gamma_i) - h_i(\gamma_i)/2}$$

Gaussian approximation

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

Super-Gaussian Bounding



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

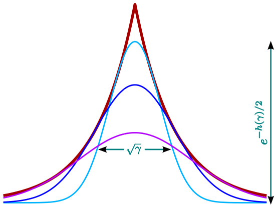
Variational problem: $Q(\boldsymbol{u}|\boldsymbol{y}) \approx P(\boldsymbol{u}|\boldsymbol{y})$

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

Gaussian approximation

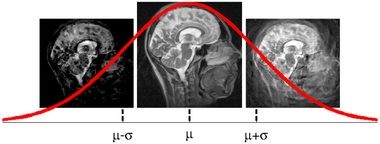
$$Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{-\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}/2}, \ \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u},$$

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



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

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: $t_i(s_i) = \max_{\gamma_i \geq 0} \dots$
  $\Rightarrow$ 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 Update 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

- Needs mean and variance of $Q(s_j|\boldsymbol{y})$ for each update
- Representation of $Q(\boldsymbol{u}|\boldsymbol{y})$: Backbone for Gaussian propagation. Moderate size problems: Cholesky representation

Seeger, JMLR 2008

Variational Approximations

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

Variational approximation

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

- Super-Gaussian bounding
- Expectation propagation
- Gaussian KL minimization

Expectation Propagation

Opper, Winther, Phys. Rev. E 2001
Minka, UAI 2001

$$P(\boldsymbol{u}|\boldsymbol{y}) \approx Q(\boldsymbol{u}|\boldsymbol{y};\gamma,\boldsymbol{b}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) \prod_i e^{b_i s_i - s_i^2/(2\gamma_i)}$$

- Best Gaussian approximation?

Expectation Propagation

Opper, Winther, Phys. Rev. E 2001
Minka, UAI 2001

$$Q(\boldsymbol{u}|\boldsymbol{y}) \xleftarrow{\text{MM}} P(\boldsymbol{u}|\boldsymbol{y}) \quad \Leftrightarrow \quad Q(\boldsymbol{u}|\boldsymbol{y}) = N(\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}], \mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}])$$

- Best Gaussian approximation? Moment matching of $P(\boldsymbol{u}|\boldsymbol{y})$

Intractable conditions for $Q(\boldsymbol{u}|\boldsymbol{y})$

$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_j e^{b_j s_j - s_j^2/(2\gamma_j)}$$
$$\uparrow \text{MM}$$
$$P(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)$$

Expectation Propagation

Opper, Winther, Phys. Rev. E 2001
Minka, UAI 2001

$$Q(\boldsymbol{u}|\boldsymbol{y}) \xleftarrow{\text{MM}} P(\boldsymbol{u}|\boldsymbol{y}) \quad \Leftrightarrow \quad Q(\boldsymbol{u}|\boldsymbol{y}) = N(\mathrm{E}[\boldsymbol{u}|\boldsymbol{y}], \mathrm{Cov}[\boldsymbol{u}|\boldsymbol{y}])$$

- Best Gaussian approximation? Moment matching of $P(\boldsymbol{u}|\boldsymbol{y})$

Intractable conditions for $Q(\boldsymbol{u}|\boldsymbol{y})$

$$\begin{aligned} Q(\boldsymbol{u}|\boldsymbol{y}) &\propto e^{b_i s_i - s_i^2/(2\gamma_i)} \times P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j\neq i} e^{b_j s_j - s_j^2/(2\gamma_j)} \\ &\uparrow \text{MM} \\ P(\boldsymbol{u}|\boldsymbol{y}) &\propto t_i(s_i) \qquad\qquad\quad \times P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j\neq i} t_j(s_j) \end{aligned}$$

Expectation Propagation

Opper, Winther, Phys. Rev. E 2001
Minka, UAI 2001

$$Q(\boldsymbol{u}|\boldsymbol{y}) \xleftarrow{\text{MM}} \hat{P}_i(\boldsymbol{u}) \quad \Leftrightarrow \quad Q(\boldsymbol{u}|\boldsymbol{y}) = N(\mathrm{E}_{\hat{P}_i}[\boldsymbol{u}], \mathrm{Cov}_{\hat{P}_i}[\boldsymbol{u}])$$

- Best Gaussian approximation? Moment matching of $P(\boldsymbol{u}|\boldsymbol{y})$
- Tractable surrogate: Moment matching for single potentials

Self-consistency conditions for $Q(\boldsymbol{u}|\boldsymbol{y})$

$$\begin{array}{ll} Q(\boldsymbol{u}|\boldsymbol{y}) \propto e^{b_i s_i - s_i^2/(2\gamma_i)} & \times P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j \neq i} e^{b_j s_j - s_j^2/(2\gamma_j)} \\ \qquad\quad \uparrow \text{MM} & \\ \hat{P}_i(\boldsymbol{u}) \propto t_i(s_i) & \times P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j \neq i} e^{b_j s_j - s_j^2/(2\gamma_j)} \end{array}$$

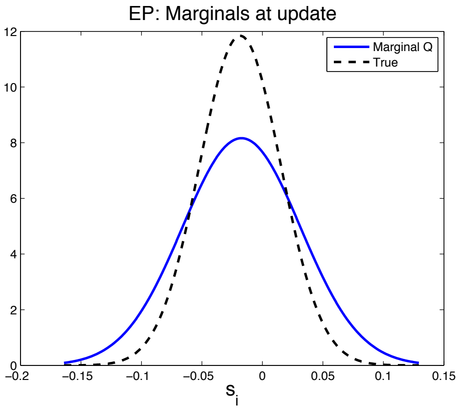
Expectation Propagation

$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) \prod_j e^{b_j s_j - s_j^2/(2\gamma_j)}$$

Expectation Propagation

$$Q(s_i|\boldsymbol{y}) \propto \int P(\boldsymbol{y}|\boldsymbol{u}) \prod_j e^{b_j s_j - s_j^2/(2\gamma_j)} \, d\{\boldsymbol{u} \setminus s_i\}$$

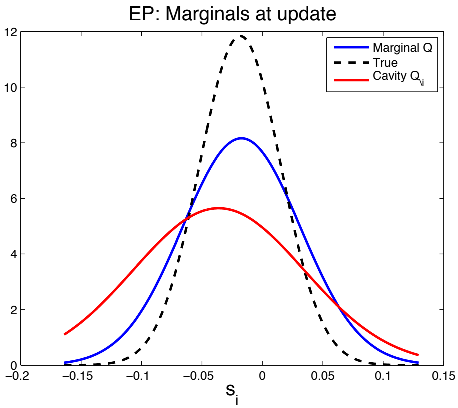
1. Marginal distribution: $Q(s_i|\boldsymbol{y})$



Expectation Propagation

$$Q_{-i}(s_i) \propto \int P(\boldsymbol{y}|\boldsymbol{u}) \prod\nolimits_{j\neq i} e^{b_j s_j - s_j^2/(2\gamma_j)} \, d\{\boldsymbol{u} \setminus s_i\}$$

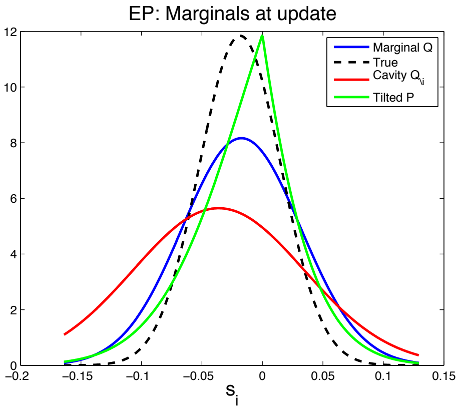
1. Marginal distribution: $Q(s_i|\boldsymbol{y})$
2. Cavity distribution: $Q_{-i}(s_i) \propto Q(s_i|\boldsymbol{y}) / e^{b_i s_i - s_i^2/(2\gamma_i)}$



Expectation Propagation

$$\hat{P}_i(s_i) \propto t_i(s_i) \int P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j \neq i} e^{b_j s_j - s_j^2/(2\gamma_j)} \, d\{\boldsymbol{u} \setminus s_i\}$$

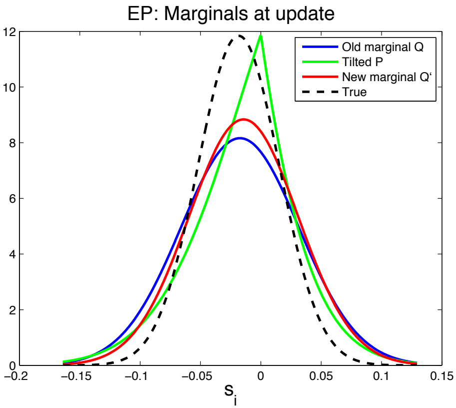
1. Marginal distribution:
   $Q(s_i|\boldsymbol{y})$
2. Cavity distribution:
   $Q_{-i}(s_i) \propto Q(s_i|\boldsymbol{y}) / e^{b_i s_i - s_i^2/(2\gamma_i)}$
3. Tilted distribution:
   $\hat{P}_i(s_i) \propto Q_{-i}(s_i) t_i(s_i)$



Expectation Propagation

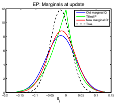
$$Q(s_i|\boldsymbol{y})' \propto e^{b_i' s_i - s_i^2/(2\gamma_i')} \int P(\boldsymbol{y}|\boldsymbol{u}) \prod_{j\neq i} e^{b_j s_j - s_j^2/(2\gamma_j)}\, d\{\boldsymbol{u} \setminus s_i\}$$

1. Marginal distribution:
   $Q(s_i|\boldsymbol{y})$
2. Cavity distribution:
   $Q_{-i}(s_i) \propto Q(s_i|\boldsymbol{y}) / e^{b_i s_i - s_i^2/(2\gamma_i)}$
3. Tilted distribution:
   $\hat{P}_i(s_i) \propto Q_{-i}(s_i) t_i(s_i)$
4. Moment matching update:
   $Q(s_i|\boldsymbol{y})' = N(\mathrm{E}_{\hat{P}_i}[s_i], \mathrm{Var}_{\hat{P}_i}[s_i])$



Expectation Propagation

$$Q(s_i|\boldsymbol{y})' \xleftarrow{\text{MM}} Z_i^{-1}\left(Q(s_i|\boldsymbol{y})/e^{b_i s_i - s_i^2/(2\gamma_i)}\right)t_i(s_i)$$



- Variational problem

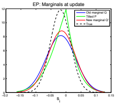
$$\phi_{\text{EP}}(\gamma) = -2\log Z_Q + \sum\nolimits_i h_i^{\text{EP}}(\gamma_i, Q(s_i|\boldsymbol{y}))$$

$$h_i^{\text{EP}}(\gamma_i, Q(s_i|\boldsymbol{y})) = -2\left(\log \mathrm{E}_{Q_{-i}}[t_i(s_i)] - \log \mathrm{E}_{Q_{-i}}[e^{b_i s_i - s_i^2/(2\gamma_i)}]\right)$$

- Arbitrary potentials
- Empirically very accurate — Nickisch *et.al.*, JMLR 2008
- Algorithmically difficult
  - Saddlepoint, not optimum of $\phi_{\text{EP}}(\gamma)$
  - EP coordinate update algorithm lacks convergence proof
  - Difficult to scale up — Seeger, Nickisch, AISTATS 2011
- Robust for log-concave potentials — Seeger, JMLR 2008

Expectation Propagation

$$Q(s_i|\boldsymbol{y})' \xleftarrow{\text{MM}} Z_i^{-1}\left(Q(s_i|\boldsymbol{y})/e^{b_i s_i - s_i^2/(2\gamma_i)}\right)t_i(s_i)$$



Expectation propagation much more general

- Related to cavity methods (adaptive TAP) — Opper, Winther, JMLR 2005
- Discrete graphical models (generalizes loopy belief propagation). Tree expectation propagation — Minka, Qi, NIPS 2004
- Dynamical systems: Natural generalization of moment matching (assumed density) filtering — Zoeter, Heskes, UAI 2002
- Inference in hybrid models (discrete and continuous)

research.microsoft.com/en-us/um/people/minka/papers/ep/roadmap.html

Variational Approximations

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

Variational approximation

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

- Super-Gaussian bounding
- Expectation propagation
- Gaussian KL minimization

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Variational inference

$$\log Z = \max_{Q(\boldsymbol{u}|\boldsymbol{y})} \mathrm{E}_Q \left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})} \right]$$

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Why not use Gaussians directly?

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y})\in\mathcal{Q}_{\text{tract}}} \mathrm{E}_Q\left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u})\prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})}\right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{ Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T\boldsymbol{s}-\frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}} \right\}$$

- Equivalent to

$$\min_{Q(\boldsymbol{u}|\boldsymbol{y})\in\mathcal{Q}_{\text{tract}}} \mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$$

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Why not use Gaussians directly?

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y}) \in \mathcal{Q}_{\text{tract}}} \mathrm{E}_Q \left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})} \right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{ Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T \boldsymbol{s} - \frac{1}{2} \boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}} \right\}$$

- Working out the variational problem:

$$-2 \log Z \leq \min_{\gamma, \boldsymbol{b}} 2\mathrm{E}_Q \left[\log \frac{Q(\boldsymbol{u}|\boldsymbol{y})}{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)} \right]$$

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Why not use Gaussians directly?

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y}) \in \mathcal{Q}_{\text{tract}}} \mathrm{E}_Q \left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})} \right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{ Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T \boldsymbol{s} - \frac{1}{2} \boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}} \right\}$$

- Working out the variational problem:

$$-2 \log Z \leq \min_{\gamma, \boldsymbol{b}} 2\mathrm{E}_Q \left[\log Z_Q^{-1} \prod_j \frac{e^{b_j s_j - s_j^2/(2\gamma_j)}}{t_j(s_j)} \right]$$

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Why not use Gaussians directly?

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y})\in\mathcal{Q}_{\text{tract}}} \mathrm{E}_Q\left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u})\prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})}\right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T\boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\right\}$$

- Working out the variational problem:

$$-2\log Z \leq \min_{\gamma,\boldsymbol{b}} -2\log Z_Q + \sum\nolimits_j 2\mathrm{E}_Q[-\log t_j(s_j) - s_j^2/(2\gamma_j) + b_j s_j]$$

Gaussian KL Minimization

Seeger, Dipl. 1999
Opper, Archambeau, N. Comp. 2009

Why not use Gaussians directly?

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y}) \in \mathcal{Q}_{\text{tract}}} \mathrm{E}_Q\left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})}\right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{Q(\boldsymbol{u}|\boldsymbol{y}) = Z_Q^{-1} P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T \boldsymbol{s} - \frac{1}{2}\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}}\right\}$$

- Working out the variational problem:

$$-2\log Z \leq \min_{\gamma, \boldsymbol{b}} -2\log Z_Q + \sum_j h_j^{\text{KL}}(\gamma_j, b_j; Q(s_j|\boldsymbol{y}))$$

Gaussian KL Minimization

$$\min_{\gamma,\boldsymbol{b}} -2\log Z_Q + \sum_j h_j^{\mathrm{KL}}(\gamma_j, b_j; Q(s_j|\boldsymbol{y}))$$

Comparison to super-Gaussian bounding:

- More general (t_j need not be super-Gaussian; $\boldsymbol{b}$ parameters)
- More difficult to solve
  - h_j^{KL} depends on $Q(s_j|\boldsymbol{y})$, so on all of γ, $\boldsymbol{b}$
  - Non-convex in general
  - No large scale algorithm so far
- Tighter bound on log partition function $\log Z$ Exercise

Gaussian KL Minimization

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y}) \in \mathcal{Q}_{\text{tract}}} \mathrm{E}_Q \left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})} \right],$$

$$\mathcal{Q}_{\text{tract}} = \left\{ Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) e^{\boldsymbol{b}^T \boldsymbol{s} - \frac{1}{2} \boldsymbol{s}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{s}} \right\}$$

- Why this form? Why not any Gaussian $Q(\boldsymbol{u}|\boldsymbol{y}) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$?
- Any Gaussian maximizer lies in $\mathcal{Q}_{\text{tract}}$ Seeger, Dipl. 1999; Exercise

$$Q^*(\boldsymbol{u}|\boldsymbol{y}) \in \underset{\text{Gaussian}}{\operatorname{argmin}} \mathrm{D}[Q(\boldsymbol{u}|\boldsymbol{y}) \,\|\, P(\boldsymbol{u}|\boldsymbol{y})]$$

$$\Rightarrow \mathrm{Cov}_{Q^*}[\boldsymbol{u}|\boldsymbol{y}]^{-1} = \sigma^{-2} \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}^T (\operatorname{diag} \gamma_*)^{-1} \boldsymbol{B}$$

Gaussian KL Minimization

$$\log Z \geq \max_{Q(\boldsymbol{u}|\boldsymbol{y})=N(\boldsymbol{\mu},\boldsymbol{\Sigma})} \mathrm{E}_Q \left[\log \frac{P(\boldsymbol{y}|\boldsymbol{u}) \prod_j t_j(s_j)}{Q(\boldsymbol{u}|\boldsymbol{y})} \right]$$

- Log-concave potentials $t_j(s_j)$:
  Problem jointly convex in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}^{\frac{1}{2}}$ — Challis, Barber, AISTATS 2011
  - Reduced Cholesky parameterizations
  - Factorization assumptions
- Coordinate update algorithm — Seeger, Dipl. 1999
- Open problem:
  Scalable algorithm for $\mathcal{Q}_{\text{tract}}$ parameterization (γ, $\boldsymbol{b}$)

What About Large Models?

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

- Needs mean and variance of $Q(s_j|\boldsymbol{y})$ for each update
- Moderate size problems: Cholesky representation
- Moderate-sized high-resolution image: $n = 65536$ pixels.

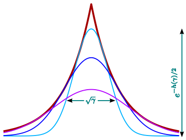
  Storage: 32G (single matrix)
  Time for Cholesky decomposition: $\approx$ 3h (if enough memory)

Out of the question

Gaussian Computations

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

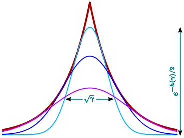
$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = ?$$



$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{s}}$$

F2

Gaussian Computations



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

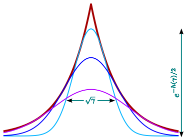
$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = ?$$

$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{u}^T\boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}\boldsymbol{u}}$$

Gaussian Computations

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

$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = ?$$

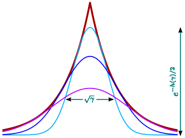


$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto e^{-\frac{1}{2}(\sigma^{-2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{u}^T\boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B}\boldsymbol{u})}$$

Gaussian Computations

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

$$\mathrm{Cov}_Q[\boldsymbol{u}|\boldsymbol{y}] = ?$$



$$Q(\boldsymbol{u}|\boldsymbol{y}) \propto e^{-\frac{1}{2}\boldsymbol{u}^T(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}+\boldsymbol{B}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{B})\boldsymbol{u}+\dots}$$

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

If $\boldsymbol{v} = \boldsymbol{A}^{-1}(\boldsymbol{B}^T\delta_j)$:

$$\mathrm{E}_Q[s_j|\boldsymbol{y}] = \boldsymbol{v}^T(\sigma^{-1}\boldsymbol{X}^T\boldsymbol{y}), \quad \mathrm{Var}_Q[s_j|\boldsymbol{y}] = \boldsymbol{v}^T(\boldsymbol{B}^T\delta_j)$$

Iterative Solvers

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

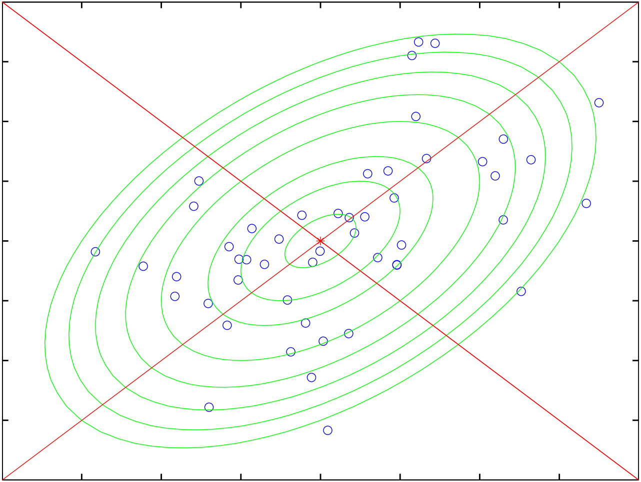
- Can multiply with $\boldsymbol{A}$ rapidly:
  $\boldsymbol{X}$, $\boldsymbol{X}^T$: FFT. $\boldsymbol{B}$, $\boldsymbol{B}^T$: Simple filters
- Solve systems by iterating over matrix-vector multiplications
- Equivalent to linear least squares estimation

$$\mathrm{E}_Q[\boldsymbol{u}|\boldsymbol{y}] = \operatorname*{argmin}_{\boldsymbol{u}} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}$$

Minimizing Quadratic Functions

Positive definite $\boldsymbol{A}$:

$$\boldsymbol{x}_* = \operatorname{argmin}\{q(\boldsymbol{x}) = \tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}\} \quad \Leftrightarrow \quad \boldsymbol{A}\boldsymbol{x}_* = \boldsymbol{b}$$



Minimizing Quadratic Functions

$$q(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}, \quad \boldsymbol{g}(\boldsymbol{x}) = \nabla q(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}$$

Require: Operator $\boldsymbol{A}$. Initial $\boldsymbol{x}_0$

for $k = 1, 2, \dots$ **do**

Pick search direction $\boldsymbol{d}_k$, based on $\boldsymbol{g}_{k-1} = \boldsymbol{g}(\boldsymbol{x}_{k-1})$, $\{\boldsymbol{d}_l : l < k\}$

Line minimization:

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \alpha_k \boldsymbol{d}_k, \quad \alpha_k = \mathrm{argmin}_\alpha\, q(\boldsymbol{x}_{k-1} + \alpha \boldsymbol{d}_k)$$

end for

Conjugate Directions

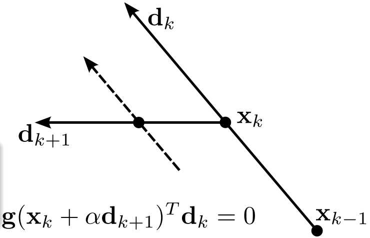
- Why, of course down as steep as possible

$$q(\boldsymbol{x}_{k-1} + d\boldsymbol{x}) = q(\boldsymbol{x}_{k-1}) + \underbrace{\boldsymbol{g}_{k-1}^T (d\boldsymbol{x})}_{\text{Smallest:}\, d\boldsymbol{x} \propto -\boldsymbol{g}_{k-1}} + O(\|d\boldsymbol{x}\|^2)$$

  Steepest descent: $\boldsymbol{d}_k = -\boldsymbol{g}_{k-1}$

- Wrong: For steepest descent: $\boldsymbol{d}_{k+1}^T \boldsymbol{d}_k = 0$
  $\Rightarrow$ Improvements from previous iterations rapidly tempered with

New gradients $\perp$ old directions?
$\Rightarrow$ Retains previous efforts:
$\boldsymbol{g}_k^T \boldsymbol{d}_k = 0 \rightarrow \boldsymbol{g}_{k+1}^T \boldsymbol{d}_k = 0 \ldots$



Conjugate Directions

$$\boldsymbol{d}_k^T \boldsymbol{A} \boldsymbol{d}_j = 0 \text{ for all } j < k$$

Towards Conjugate Gradients

Details: Handout

1. Directions conjugate: Gradient $\boldsymbol{g}_k \perp$ all previous directions: $\boldsymbol{g}_k^T \boldsymbol{d}_j = 0$ for all $j \leq k$
2. After n steps we are done: $\boldsymbol{g}_n = \boldsymbol{0}$
3. Construct conjugate directions by recurrence: $\boldsymbol{d}_k = -\boldsymbol{g}_{k-1} + \beta_{k-1} \boldsymbol{d}_{k-1}$
4. All gradients are orthogonal: $\boldsymbol{g}_k^T \boldsymbol{g}_j = 0$, $j < k$
   [Bit of misnomer: Directions are conjugate]
5. What is α_k? From line minimization:
$$\alpha_k = \frac{\|\boldsymbol{g}_{k-1}\|^2}{\boldsymbol{d}_k^T \boldsymbol{A} \boldsymbol{d}_k}$$
6. What is β_k? The great synthesis!
$$\beta_k = \frac{\|\boldsymbol{g}_k\|^2}{\|\boldsymbol{g}_{k-1}\|^2}$$

Conjugate Gradients Algorithm

Require: Operator $\boldsymbol{A}$. Initial $\boldsymbol{x}_0$. $\boldsymbol{g}_0 = \boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b}$

for $k = 1, 2, \dots$ (no more than n) **do**

  $\rho_{k-1} = \|\boldsymbol{g}_{k-1}\|^2$

  if $k = 1$ **then**

    $\boldsymbol{d}_1 = -\boldsymbol{g}_0$

  else

    $\beta_{k-1} = \rho_{k-1}/\rho_{k-2}$; $\boldsymbol{d}_k = -\boldsymbol{g}_{k-1} + \beta_{k-1}\boldsymbol{d}_{k-1}$

  end if

  $\boldsymbol{q}_k = \boldsymbol{A}\boldsymbol{d}_k$; $\alpha_k = \rho_{k-1}/(\boldsymbol{d}_k^T\boldsymbol{q}_k)$

  $\boldsymbol{x}_k = \boldsymbol{x}_{k-1} + \alpha_k\boldsymbol{d}_k$; $\boldsymbol{g}_k = \boldsymbol{g}_{k-1} + \alpha_k\boldsymbol{q}_k$

  Check for convergence (say $\|\boldsymbol{g}_k\| < \varepsilon\|\boldsymbol{b}\|$)

end for

Conjugate Gradients Algorithm

Let $\mathcal{K}_k = \boldsymbol{x}_0 + \mathrm{span}\{\boldsymbol{d}_1, \dots, \boldsymbol{d}_k\}$. Then:

$$\boldsymbol{g}_k^T \boldsymbol{d}_j = 0,\ j \le k \quad \Rightarrow \quad \boldsymbol{x}_k = \mathrm{argmin}_{\boldsymbol{x} \in \mathcal{K}_k}\, q(\boldsymbol{x})$$

But $\mathcal{K}_k = \boldsymbol{x}_0 + \mathrm{span}\{\boldsymbol{A}^j \boldsymbol{g}_0 \mid j < k\}$
$\Rightarrow$ Optimal with k ($\boldsymbol{A}\cdot$) multiplications!

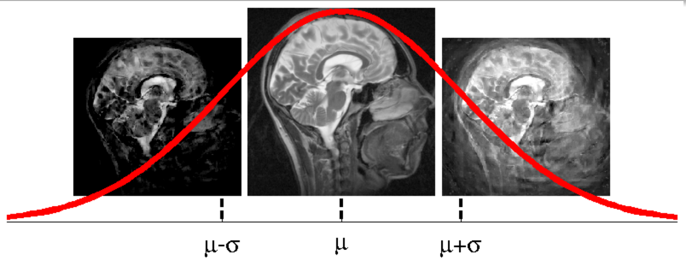
- $\mathcal{K}_k \subset \mathcal{K}_{k+1} \subset \dots$, $\boldsymbol{x}_* \in \mathcal{K}_n$ (Cayley/Hamilton)
- What about $k \ll n$ for huge n? Depends on eigenspectrum of $\boldsymbol{A}$.
  $\boldsymbol{x}_k \approx \boldsymbol{x}_*$ in surprisingly many cases in practice
  $\Rightarrow$ Krylov subspace view key to convergence analysis — Exercise
- Preconditioning: $\boldsymbol{M} = \boldsymbol{C}\boldsymbol{C}^T \approx \boldsymbol{A}$, but easy to solve systems with
  - Work on $(\boldsymbol{C}^{-T}\boldsymbol{A}\boldsymbol{C}^{-1})\boldsymbol{C}\boldsymbol{x} = \boldsymbol{C}^{-T}\boldsymbol{b}$
    $\Rightarrow$ Better spectral properties $\rightarrow$ Faster convergence
  - CG as before, with one ($\boldsymbol{M}^{-1}\cdot$) per iteration
  - Art of iterative linear solvers

Outline

Recap

- Sparse linear models:
  Inverse problems, Bayesian calibration and sampling optimization
- Super-Gaussian bounding:
  From local max-of-Gaussian representations to global bound
- Expectation propagation:
  Tractable self-consistency by local moment matching
- Gaussian KL minimization:
  Tighter, but more difficult than super-Gaussian bounding
- Conjugate gradients:
  Large scale linear solvers by iterated matrix-vector multiplications

Need for Scalability



Bayesian inference over full images (256 $\times$ 256)?

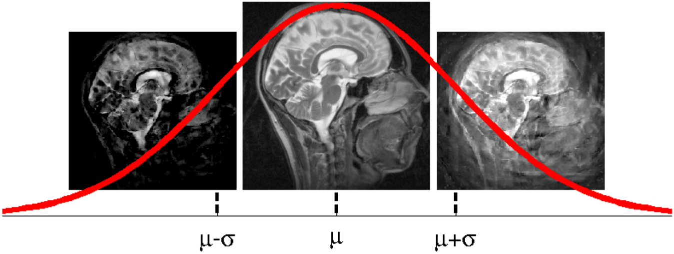
$\Rightarrow \boldsymbol{u} \in \mathbb{C}^{65536}$, $\gamma \in \mathbb{R}^{196096}$

Need for Scalability

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

- Needs mean and variance of $Q(s_j|\boldsymbol{y})$ for each update

Need for Scalability

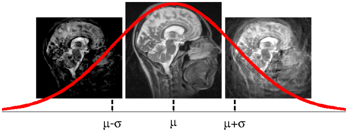


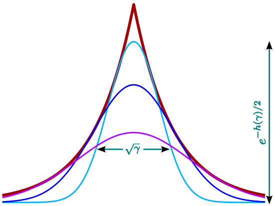
Bayesian inference over full images (256×256)?
$\Rightarrow \boldsymbol{u} \in \mathbb{C}^{65536}$, $\gamma \in \mathbb{R}^{196096}$

- Coordinate update algorithm (most previous methods)
  - Linear system $\boldsymbol{A}^{-1}\boldsymbol{r}$ for each update (needs conjugate gradients)
  - At least 196096 systems (visit each γ_i once)

Out of the question

Properties of Super-Gaussian Bounding





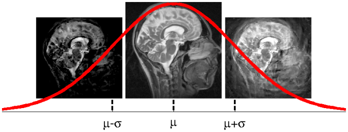
$$\min_{\gamma} -2\log \int P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{s}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u} + h(\gamma)$$

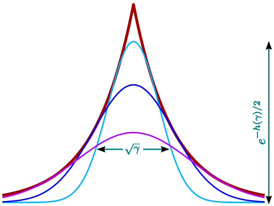
Super-Gaussian bounding stands out

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

Why is that?

MAP estimation will help solving it!

Towards Scalable 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

$$\log Z$$

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

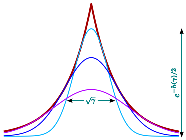
$$\geq$$

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

Towards Scalable Variational Inference

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

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



- Harder than MAP estimation. But why?
- Convert integration to optimization

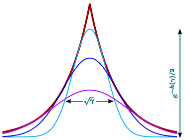
F4

$$\int P(\boldsymbol{y}|\boldsymbol{u}) e^{-\frac{1}{2}\boldsymbol{s}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{s}}\, d\boldsymbol{u} \overset{!}{=} |2\pi \boldsymbol{A}^{-1}|^{1/2} \max_{\boldsymbol{u}_*} P(\boldsymbol{y}|\boldsymbol{u}_*) e^{-\frac{1}{2}\boldsymbol{s}_*^T \boldsymbol{\Gamma}^{-1}\boldsymbol{s}_*}$$

Towards Scalable Variational Inference

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

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



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

Super-Gaussian bounding

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

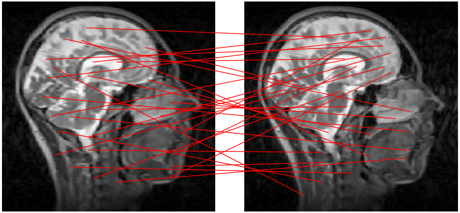
Decoupling by Convex Duality

$$-2 \log Z \le \min_{\gamma, \boldsymbol{u}_*} \log |\boldsymbol{A}(\gamma)| + \phi_{\cup}(\boldsymbol{u}_*, \gamma)$$

- Dependencies in posterior $P(\boldsymbol{u}|\boldsymbol{y})$
  $\Rightarrow$ Difficult coupling term $\log |\boldsymbol{A}|$ in criterion ϕ

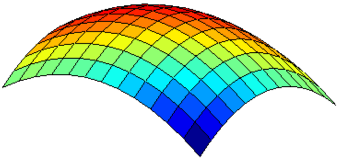
$$\boldsymbol{A} = \sigma^{-2} \boldsymbol{X}^T \boldsymbol{X} + \boldsymbol{B}^T \boldsymbol{\Gamma}^{-1} \boldsymbol{B}$$

- $\boldsymbol{A} \mapsto \log |\boldsymbol{A}|$ concave
- $\gamma^{-1} \mapsto \log |\boldsymbol{A}|$ concave



Decoupling by Convex Duality

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

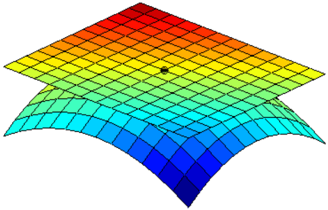


Decoupling by Convex Duality

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

Convex (Fenchel) duality

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

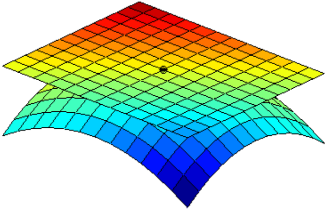


Decoupling by Convex Duality

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

Convex (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 $\boldsymbol{z}$]

F5

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

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 $(\boldsymbol{u}_*, \gamma)$]

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

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 en masse (thanks to MAP "gold rush")

2. Gaussian variances

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

   - More difficult
   - Methods from numerical maths, spatial statistics, solid state physics

Summary

$$\text{Variational problem:} \quad \min_{\gamma} -2\log Z_Q + h(\gamma)$$

1. Variational representation of Gaussian log partition function

$$-2\log Z_Q \doteq \min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + (\boldsymbol{s}_*{}^2)^T(\gamma^{-1}) + \log|\boldsymbol{A}|$$

Summary

$$\text{Variational problem:} \quad \min_{\gamma} -2\log Z_Q + h(\gamma)$$

1. Variational representation of Gaussian log partition function

$$-2\log Z_Q \doteq \min_{\boldsymbol{z},\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + (\boldsymbol{z} + \boldsymbol{s}_*{}^2)^T(\gamma^{-1}) - g^*(\boldsymbol{z})$$

2. Choose computationally favourable ordering of updates

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

   Variances $\boldsymbol{z}$ expensive: Fix them as long as sensible

3. Convergence guarantee: Tangential bound to $\log|\boldsymbol{A}|$ Wipf *et.al.*, NIPS 2008

Factorization Assumptions

$$\min_{\boldsymbol{z}} \left(\min_{\boldsymbol{u}_*} \min_{\gamma} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) \right), \quad \boldsymbol{z} \leftarrow \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$$

- Real time? As fast as MAP estimation?
  Gaussian variances can be a real problem

Factorization Assumptions

$$\min_{\boldsymbol{z}} \left(\min_{\boldsymbol{u}_*} \min_{\gamma} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) \right)$$

- Real time? As fast as MAP estimation?
  Gaussian variances can be a real problem
- Factorization assumptions:

$$Q(\boldsymbol{u}|\boldsymbol{y}) = \prod_{i=1}^{n} Q(u_i|\boldsymbol{y}) \quad \Rightarrow \quad Z_Q = \prod_{i=1}^{n} Z_{Q(u_i|\boldsymbol{y})}$$

$\Rightarrow$ Update $\boldsymbol{z}$ in $O(q) = O(n)$

Factorization Assumptions

$$\min_{\boldsymbol{u}_*} \left(\min_{\boldsymbol{z}} \min_{\gamma} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) \right)$$

- Real time? As fast as MAP estimation?
  Gaussian variances can be a real problem
- Factorization assumptions:

$$Q(\boldsymbol{u}|\boldsymbol{y}) = \prod_{i=1}^{n} Q(u_i|\boldsymbol{y}) \quad \Rightarrow \quad Z_Q = \prod_{i=1}^{n} Z_{Q(u_i|\boldsymbol{y})}$$

  $\Rightarrow$ Update $\boldsymbol{z}$ in $O(q) = O(n)$
- One penalized least squares problem (coupled regularizer)

Factorization Assumptions

$$\min_{\boldsymbol{u}_*} \sigma^{-2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + \left(\min_{\boldsymbol{z}} \min_{\gamma} \mathcal{R}_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma) \right)$$

- Real time? As fast as MAP estimation?
  Gaussian variances can be a real problem
- Factorization assumptions:

$$Q(\boldsymbol{u}|\boldsymbol{y}) = \prod_{i=1}^{n} Q(u_i|\boldsymbol{y}) \quad \Rightarrow \quad Z_Q = \prod_{i=1}^{n} Z_{Q(u_i|\boldsymbol{y})}$$

$\Rightarrow$ Update $\boldsymbol{z}$ in $O(q) = O(n)$

- One penalized least squares problem (coupled regularizer)
- Convexity properties are retained

Factorization Assumptions

$$Q(\boldsymbol{u}|\boldsymbol{y}) = \prod_{i=1}^{n} Q(u_i|\boldsymbol{y})$$

- Advantages
  - Essentially as fast as MAP estimation
  - Gaussian KL minimization convex — Challis, Barber, AISTATS 2011
  - It might just work . . .
- Drawbacks
  - True posterior tightly and strongly coupled:
    Expect better results without factorizations
  - Bayesian experimental design (active sampling) relies on covariances

Penalized Least Squares

Computational primitives driving large scale inference

1. Penalized least squares (≈ MAP estimation)

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

Penalized Least Squares

Computational primitives driving large scale inference

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

$$\min_{\boldsymbol{u}_*} \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + \sum_{i=1}^{q} \psi_i(s_{*i}), \quad \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$$

Iteratively Reweighted Least Squares

$$\min_{\boldsymbol{u}_*} \left\{ \phi_{\boldsymbol{z}}(\boldsymbol{u}_*) = \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + \sum\nolimits_{i=1}^{q} \psi_i(s_{*i}) \right\}, \quad \boldsymbol{s}_* = \boldsymbol{B}\boldsymbol{u}_*$$

- ψ_i twice differentiable: Newton-Raphson optimization
- Taylor approximation at $\boldsymbol{u}_* = \boldsymbol{u}_k$:

$$\phi_{\boldsymbol{z}}(\boldsymbol{u}_*) \approx \sigma^{-2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}_*\|^2 + \boldsymbol{s}_*^T(\operatorname{diag}\boldsymbol{h}_k)\boldsymbol{s}_* - 2\boldsymbol{g}_k^T\boldsymbol{s}_* + C_k$$

- Newton search direction by conjugate gradients:

$$\boldsymbol{d}_k = \boldsymbol{A}_k^{-1}\left(\sigma^{-2}\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{B}^T\boldsymbol{g}_k\right), \quad \boldsymbol{A}_k = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{B}^T(\operatorname{diag}\boldsymbol{h}_k)\boldsymbol{B}$$

- Line search in $O(q)$:

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \alpha_*\boldsymbol{d}_k, \quad \alpha_* = \operatorname*{argmin}_{\alpha>0} \phi_{\boldsymbol{z}}(\boldsymbol{u}_k + \alpha\boldsymbol{d}_k)$$

Iteratively Reweighted Least Squares

- Advantages
  - Rapid (quadratic) convergence
  - Reuse code for conjugate gradients algorithm
- Drawbacks
  - Two nested loops: Difficult to fine-tune
  - MAP estimation: ψ_i may not be twice differentiable
  - For our setup: Systems

$$\boldsymbol{u} = \left(\boldsymbol{X}^T\boldsymbol{X} + \rho\boldsymbol{B}^T\boldsymbol{B}\right)^{-1}\boldsymbol{r}, \quad \rho > 0$$

can be solved analytically.

Augmented Lagrangian Solvers

$$\min_{\boldsymbol{u}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 - \sum\nolimits_j \log e^{-\tau|s_j|}, \quad \boldsymbol{s} = \boldsymbol{B}\boldsymbol{u}$$

- Consider MAP estimation problem: Not differentiable

Augmented Lagrangian Solvers

$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1, \quad \kappa = \tau\sigma^2$$

- Consider MAP estimation problem: Not differentiable

Augmented Lagrangian Solvers

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

- Consider MAP estimation problem: Not differentiable
- Rewrite: Operator splitting.
  $\Rightarrow$ Would be simple without constraint

Augmented Lagrangian Solvers

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

- Consider MAP estimation problem: Not differentiable
- Rewrite: Operator splitting.
  $\Rightarrow$ Would be simple without constraint
- Dualize constraint (Lagrange multipliers $\boldsymbol{b}$)

Augmented Lagrangian Solvers

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

- Consider MAP estimation problem: Not differentiable
- Rewrite: Operator splitting.
  $\Rightarrow$ Would be simple without constraint
- Dualize constraint (Lagrange multipliers $\boldsymbol{b}$)
- Augmented Lagrangian technique (additional smoothing) F6

Alternating Direction Method of Multipliers

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

Alternating Direction Method of Multipliers

Iterate:

- Linear least squares (fixed $\boldsymbol{s}$, $\boldsymbol{b}$)

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

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

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

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

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

The Proximal Map

$$\mathrm{prox}_f(\boldsymbol{r}) := \operatorname*{argmin}_{\boldsymbol{s}} f(\boldsymbol{s}) + \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{r}\|^2$$

- Recall: $\boldsymbol{s} \leftarrow \mathrm{prox}_{(\kappa/\lambda)\|\cdot\|_1}(\boldsymbol{B}\boldsymbol{u} + \boldsymbol{b})$ F7

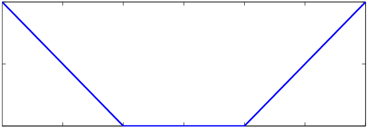
The Proximal Map

$$\mathrm{prox}_f(\boldsymbol{r}) := \operatorname*{argmin}_{\boldsymbol{s}} f(\boldsymbol{s}) + \frac{1}{2}\|\boldsymbol{s} - \boldsymbol{r}\|^2$$

- Recall: $\boldsymbol{s} \leftarrow \mathrm{prox}_{(\kappa/\lambda)\|\cdot\|_1}(\boldsymbol{Bu} + \boldsymbol{b})$
- Simple for decoupling $f(\boldsymbol{s}) = \sum_i f_i(s_i)$: $\mathrm{prox}_f(\boldsymbol{r}) = [\mathrm{prox}_{f_i}(r_i)]$

- For Laplace (ℓ_1): Soft thresholding:

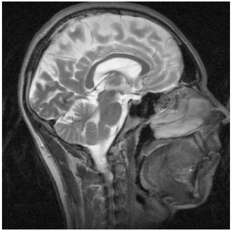
$$\mathrm{prox}_{\alpha|\cdot|}(r) = \frac{\max\{|r| - \alpha, 0\}}{|r|} r$$



$\Rightarrow$ Sparsity in $\boldsymbol{s}$

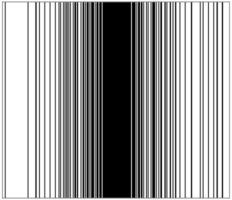
MRI Reconstruction

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



MRI Reconstruction

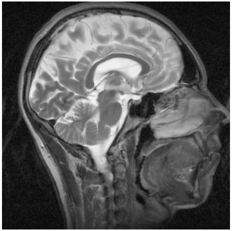
$$\min_{\boldsymbol{u}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \kappa\|\boldsymbol{B}\boldsymbol{u}\|_1$$



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

MRI Reconstruction

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



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

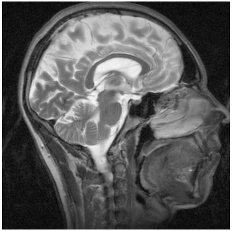
- Linear least squares: F8

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

MRI Reconstruction

$$\min_{\boldsymbol{u}} \tfrac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{u}\|^2 + \tfrac{\lambda}{2}\|\boldsymbol{B}\boldsymbol{u} - (\boldsymbol{s} - \boldsymbol{b})\|^2$$

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

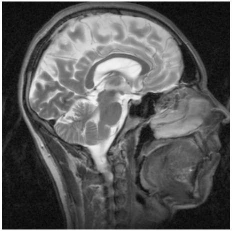


- Linear least squares:

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

MRI Reconstruction

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



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

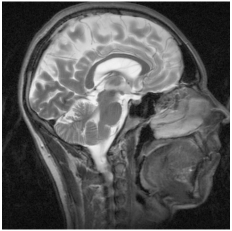
- Linear least squares:

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

MRI Reconstruction

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

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



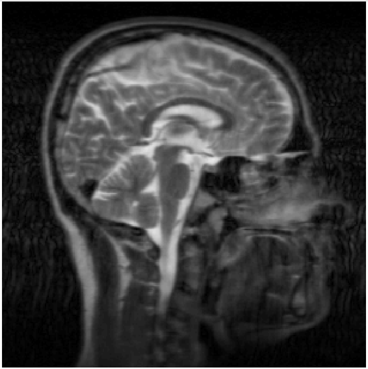
- Linear least squares:

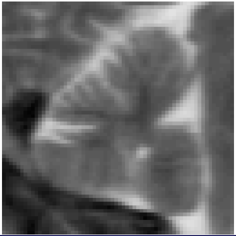
$$\underbrace{\left(\boldsymbol{I}_{\cdot,J}\boldsymbol{I}_{J,\cdot} + \boldsymbol{D}\right)}_{\text{diagonal}} \boldsymbol{F}\boldsymbol{u} = \boldsymbol{F}\boldsymbol{r}$$

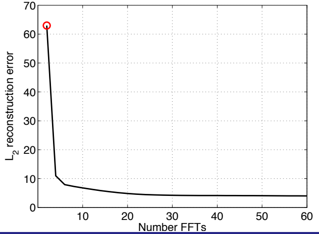
$\Rightarrow$ Two fast Fourier transforms only!

MRI Reconstruction

courtesy Mateusz Malinowski

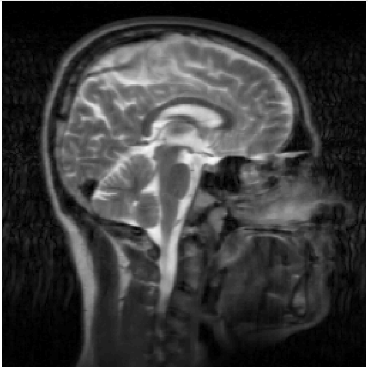


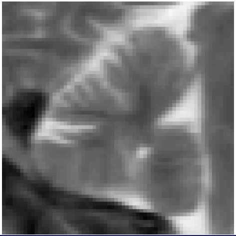


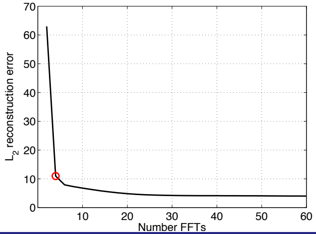


MRI Reconstruction

courtesy Mateusz Malinowski

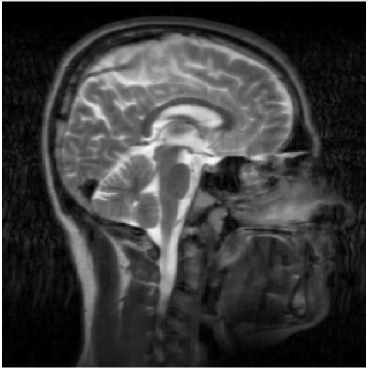


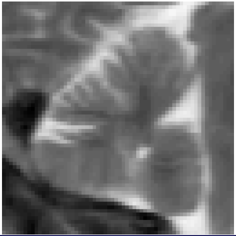


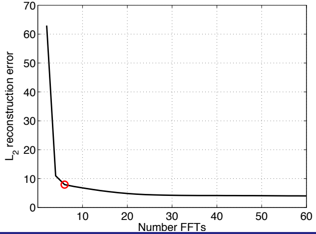


MRI Reconstruction

courtesy Mateusz Malinowski

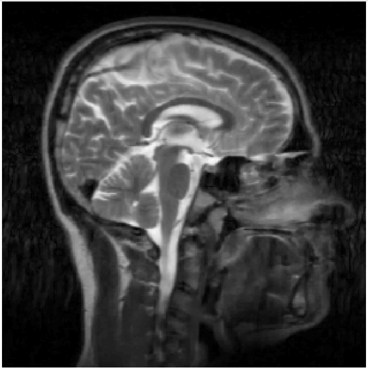


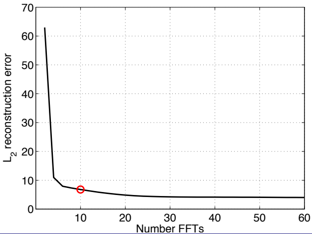




MRI Reconstruction

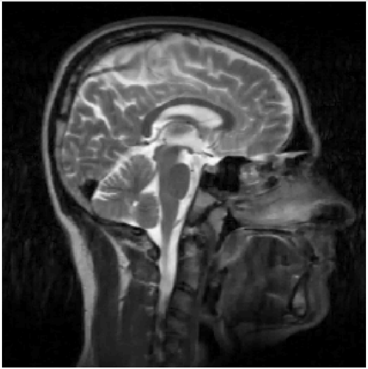
courtesy Mateusz Malinowski

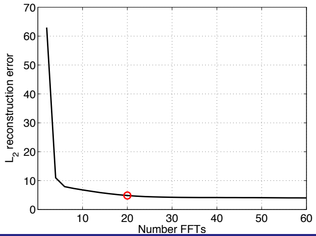




MRI Reconstruction

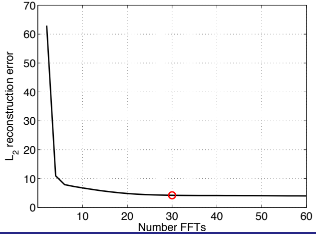
courtesy Mateusz Malinowski





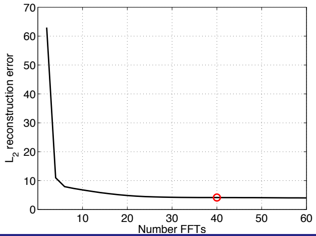
MRI Reconstruction

courtesy Mateusz Malinowski



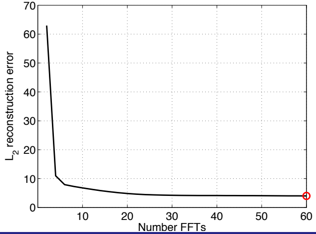
MRI Reconstruction

courtesy Mateusz Malinowski



MRI Reconstruction

courtesy Mateusz Malinowski



ADMM versus IRLS

Inner loop problems are smooth (twice differentiable)

- Use ADMM only if linear least squares step has direct solution (e.g., two FFTs)
- ADMM simpler to code and run (no CG inside)
- With complex data terms (many parts): ADMM easier to parallelize
- IRLS has much better convergence rate

Gaussian Variances

Computational primitives driving large scale inference

2. Gaussian variances

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

A Difficult Problem

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

- Gaussian variances much more difficult than Gaussian means
  - All means? One linear system
  - All variances? n linear systems!
- Situation for Gaussian loopy belief propagation (LBP)
  - If LBP converges: Means are exact — Weiss *et.al.*, JCOMP 2001
  - Variances are wrong in general: — Malioutov *et.al.*, JMLR 2006
    Major part of computation (typically) not done by LBP
- Some tractable cases
  - Tree-structured graphical model: Both means and variances in $O(n)$
  - $\boldsymbol{A}$ admits sparse Cholesky factorization: — van Gerven *et.al.*, Neuroimage 2010
    Variances by Takahashi equation

Our $\boldsymbol{A}$ is none of these (densely coupled)

A Difficult Problem

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

- Other fields need them as well
  - Electronic structure calculations
  - Uncertainty quantifications for PDEs
  - Gaussian MRFs for remote sensing

Low Rank Approximations

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

- Pick $\boldsymbol{V} \in \mathbb{R}^{n \times L}$, $L \ll n$:

$$\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) \approx \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{B}^T) = \sum_{l=1}^{L}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{v}_l) \circ (\boldsymbol{B}\boldsymbol{v}_l)$$

- Solve L linear systems instead of n
- How to choose $\boldsymbol{V}$?

Hadamard Vectors

Bekas *et.al.*, App. Num. Math. 2007

$$\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) \approx \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{B}^T)$$

- Hadamard matrix:

$$\boldsymbol{H}_n \in \{-1, +1\}^{n\times n}, \quad \boldsymbol{H}_n^T\boldsymbol{H}_n = n\boldsymbol{I}$$

- Pick L columns:

$$\boldsymbol{V} = \frac{1}{\sqrt{L}}(\boldsymbol{H}_n)_{\cdot,L}$$

- Deterministic estimator.
  Intuition: Maximize smallest angle between any pair of rows of $\boldsymbol{V}$

Perturb&MAP

Papandreou, Yuille, NIPS 2010

$$\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) \approx \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{V}\boldsymbol{V}^T\boldsymbol{B}^T)$$

- Draw L independent samples $\boldsymbol{q}_l \sim N(\boldsymbol{0}, \boldsymbol{A}^{-1})$:

$$\mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T) = \mathrm{E}[(\boldsymbol{B}\boldsymbol{q}_1)^2] \approx \tfrac{1}{L}\sum_{l=1}^{L}(\boldsymbol{B}\boldsymbol{q}_l)^2$$

  Optimal Monte Carlo estimator (no knowledge about $\boldsymbol{A}$)

- One linear system per sample: F9

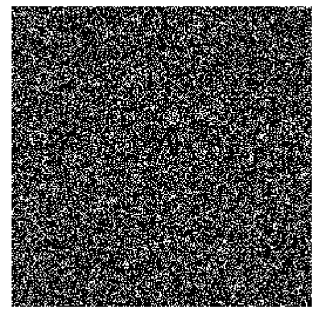
$$\boldsymbol{w}_l \sim N(\boldsymbol{0}, \boldsymbol{A}), \quad \boldsymbol{q}_l = \boldsymbol{A}^{-1}\boldsymbol{w}_l$$

Outline

Ill-Posed Inverse Problems

- Undersampled image reconstruction

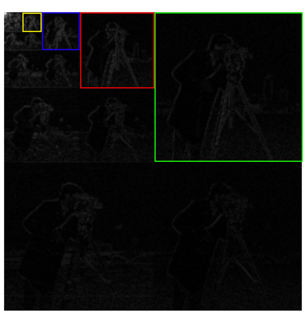
Linear measurements



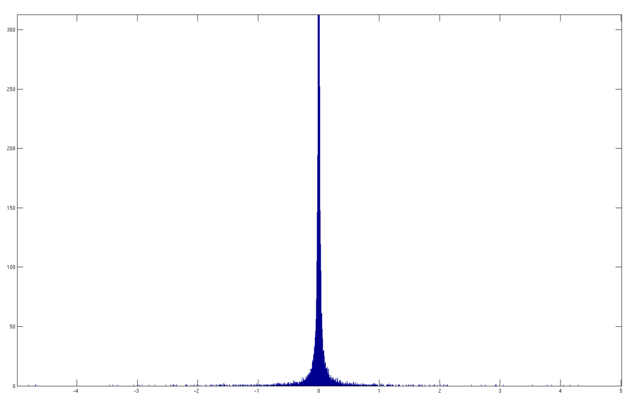
Reconstruction

$$\boldsymbol{u} \qquad\qquad \boldsymbol{y} = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\varepsilon} \qquad\qquad \hat{\boldsymbol{u}}$$

- Resolve ambiguities from prior knowledge (transform sparsity, ...)



hist



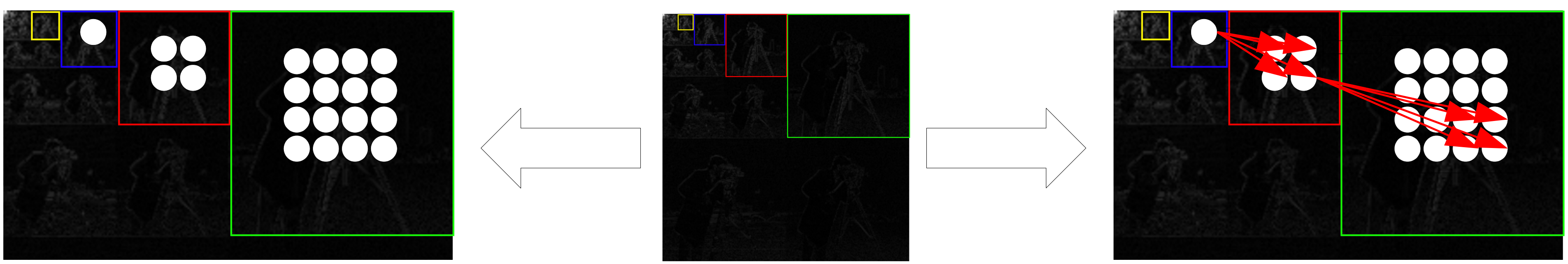
$$P(\boldsymbol{u})$$

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{u}$$

Model: Factorial vs. Structured

- Independent sparsity potentials
- No structure beyond component-wise sparsity

- Discrete tree-structured backbone
- **Mixtures** of sparsity potentials



$$P(\boldsymbol{u}) = \prod_j t_j(s_j)$$

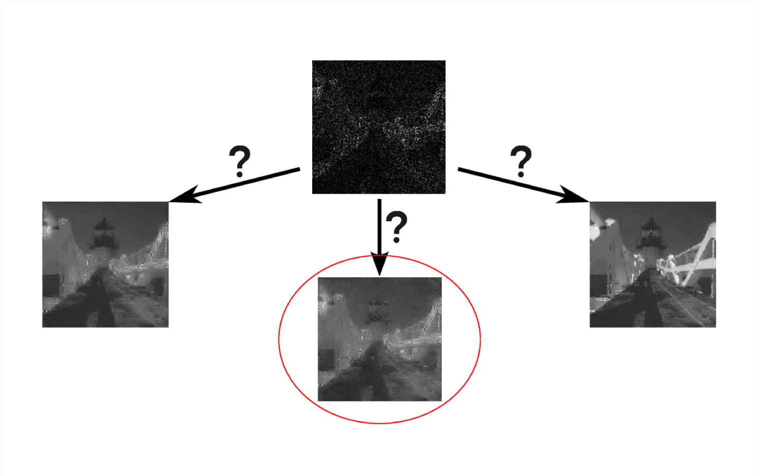
$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{u}$$

$$P(\boldsymbol{u}) = \sum_{\boldsymbol{\delta}} \prod_j t_j(s_j; \delta_j) P(\boldsymbol{\delta})$$

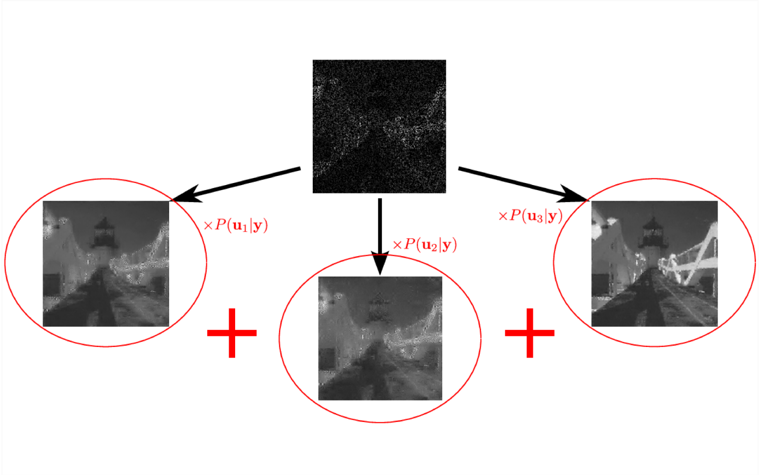
Method: MAP vs. Inference

- Estimate single maximum point
- Many fast algorithms

- Bayesian inference over posterior distribution
- Integrate, don't just maximize



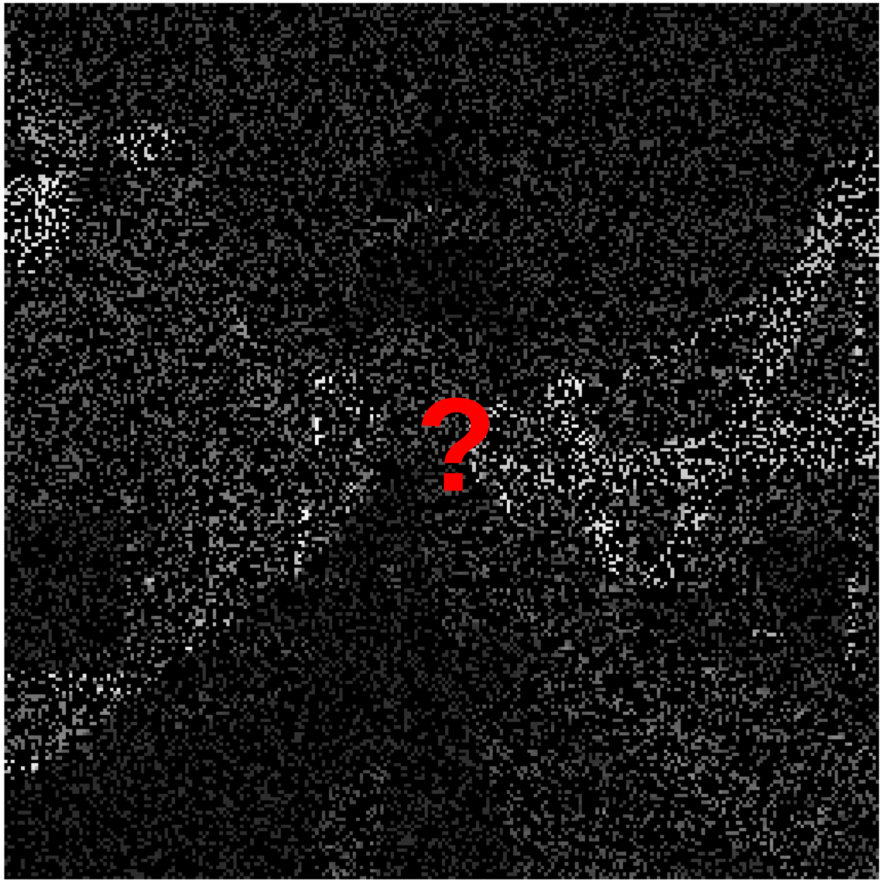
$$\hat{\boldsymbol{u}} = \underset{\boldsymbol{u}}{\operatorname{argmax}}\, P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})$$

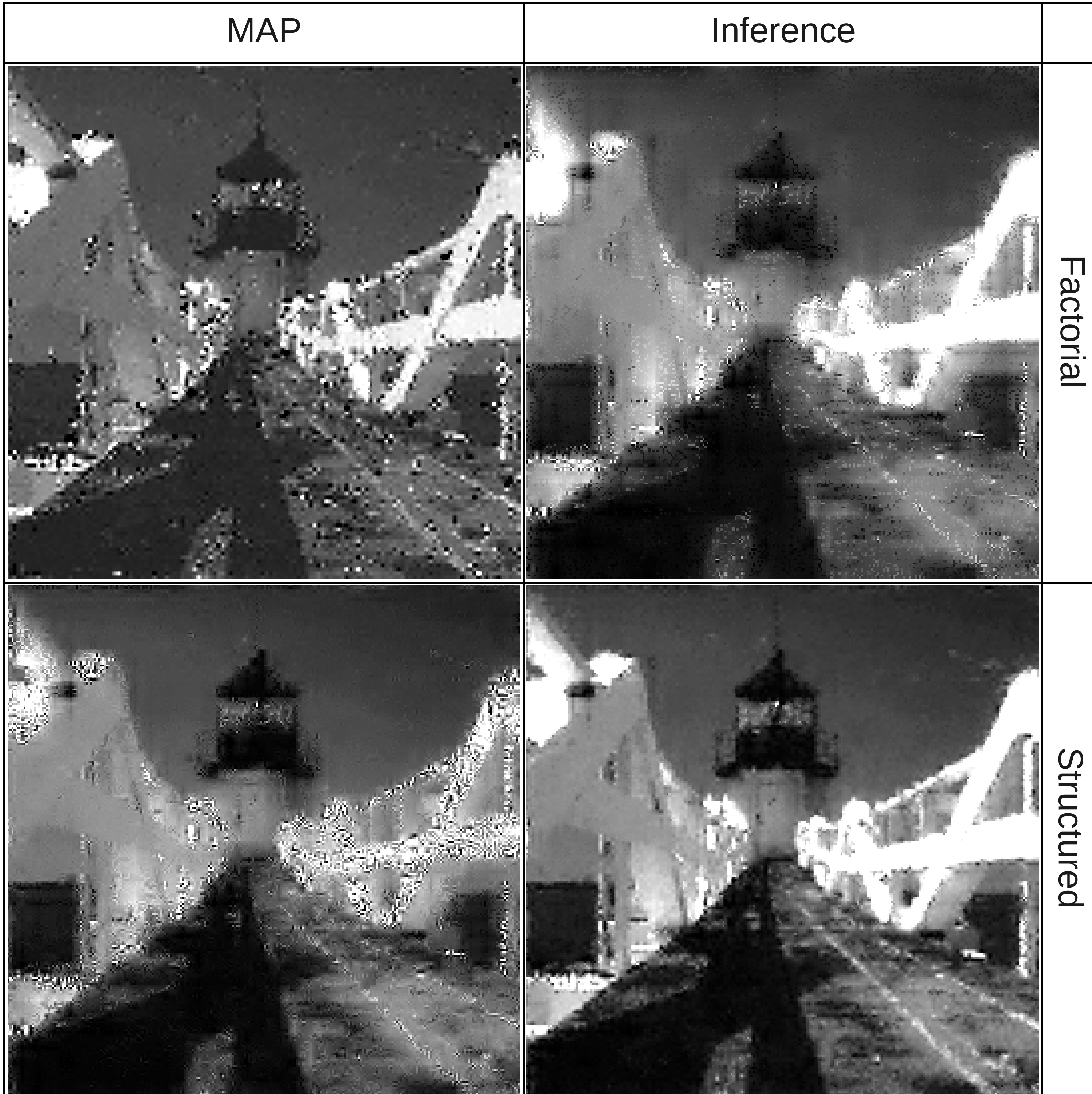


$$P(\boldsymbol{u}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})}{\int P(\boldsymbol{y}|\boldsymbol{u})P(\boldsymbol{u})d\boldsymbol{u}}$$

Example: Image Inpainting

- Problem: 75% of pixels randomly removed



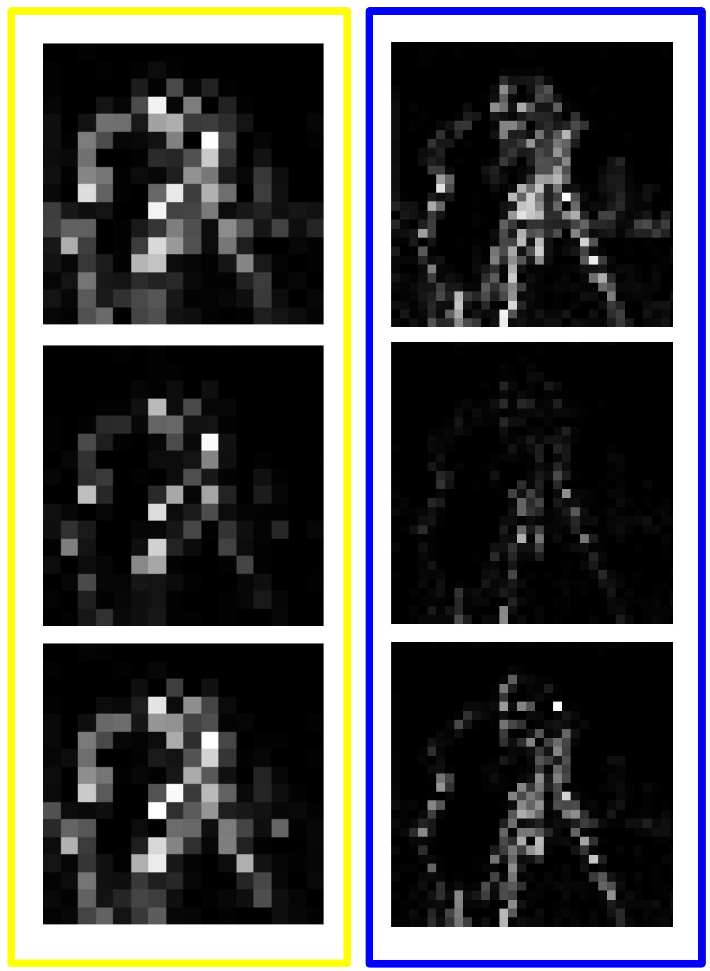

MAP
Inference
Factorial
Structured

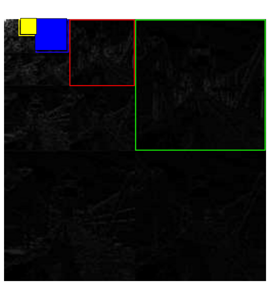
Multi-Scale Wavelet Analysis

$$\boldsymbol{s} = \boldsymbol{W}\boldsymbol{u}$$

$$\hat{\boldsymbol{s}}_{MAP} = \boldsymbol{W}\hat{\boldsymbol{u}}_{MAP}$$

$$\hat{\boldsymbol{s}}_{INF} = \boldsymbol{W}\hat{\boldsymbol{u}}_{INF}$$



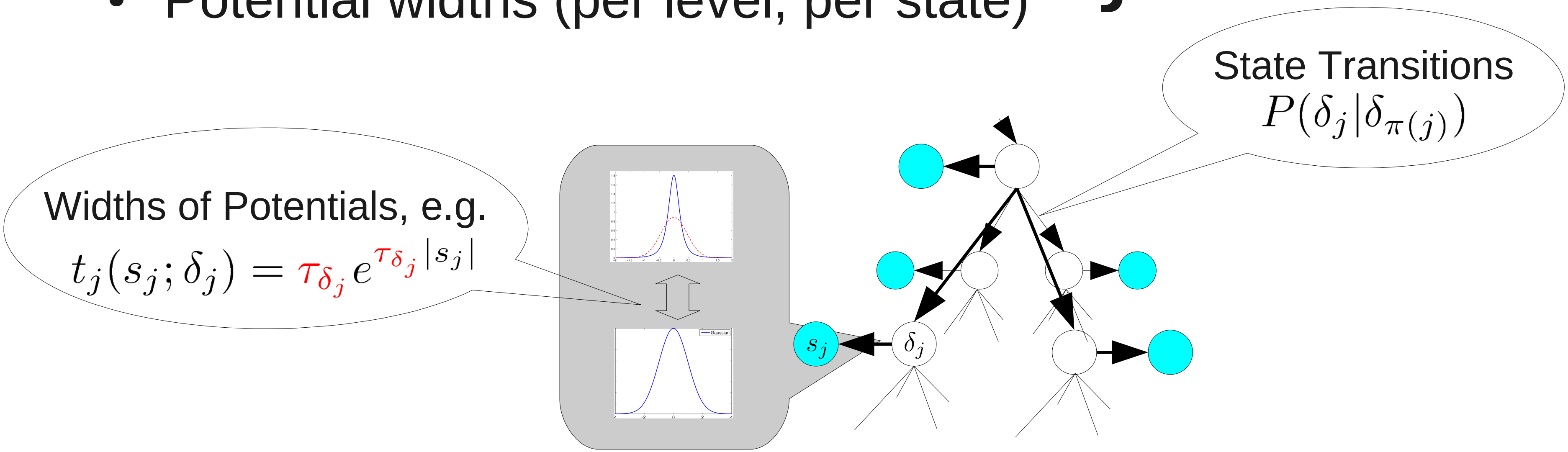


- Quad-tree: Correspondence between scales
- Energy percolates down the tree
- Better recovery where it really matters

Learning Structured Models

- Rich parametrization
  - Transition probabilities (per level)
  - Potential widths (per level, per state)

$\Big\} \ \boldsymbol{\theta}$



- Parameters learned automatically from raw data
  - Simple closed form updates
  - No expensive cross validation

Variational Bayesian Inference

Ko, Seeger, ICML 2012

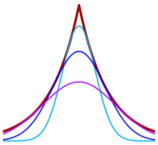
$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding

F10

$$\log P(\boldsymbol{y}) \geq$$

$$\max_{\gamma} \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{Q(\boldsymbol{u}|\delta;\gamma)}_{\text{Gaussian}} e^{-\frac{1}{2}h(\gamma;\delta)} P(\delta) \, d\boldsymbol{u}$$



Variational Bayesian Inference

Ko, Seeger, ICML 2012

$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding
2. Factorize: $Q(\boldsymbol{u}, \delta|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\delta|\boldsymbol{y})$

F11

$$\log P(\boldsymbol{y}) \geq \max_{\gamma, Q(\delta|\boldsymbol{y})} \log \int \underbrace{P(\boldsymbol{y}|\boldsymbol{u})Q(\boldsymbol{u}|\langle\delta\rangle_Q; \gamma)}_{\text{Gaussian}} \, d\boldsymbol{u}$$

$$- \mathrm{D}[Q(\delta|\boldsymbol{y}) \,\|\, P(\delta)] - \tfrac{1}{2} h(\gamma; \langle\delta\rangle_Q)$$

Variational Bayesian Inference

Ko, Seeger, ICML 2012

$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding
2. Factorize: $Q(\boldsymbol{u}, \delta|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\delta|\boldsymbol{y})$

$$\log P(\boldsymbol{y}) \geq \log Z_Q(\langle\delta\rangle_Q, \gamma) \\ - \mathrm{D}[Q(\delta|\boldsymbol{y}) \,\|\, P(\delta)] - \tfrac{1}{2} h(\gamma; \langle\delta\rangle_Q)$$

Variational Bayesian Inference

Ko, Seeger, ICML 2012

$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding
2. Factorize: $Q(\boldsymbol{u}, \delta|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\delta|\boldsymbol{y})$
3. Variational representation of $\log Z_Q$

$$-2\log P(\boldsymbol{y}) \le \min_{\gamma, Q(\delta|\boldsymbol{y})} \min_{\boldsymbol{z}, \boldsymbol{u}_*} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma, \langle\delta\rangle_Q) + 2\mathrm{D}[Q(\delta|\boldsymbol{y}) \,\|\, P(\delta)]$$

Variational Bayesian Inference

$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding
2. Factorize: $Q(\boldsymbol{u}, \delta|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\delta|\boldsymbol{y})$
3. Variational representation of $\log Z_Q$

$$-2\log P(\boldsymbol{y}) \le \min_{\boldsymbol{z}} \left(\min_{Q(\delta|\boldsymbol{y}), \boldsymbol{u}_*, \gamma} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma, \langle\delta\rangle_Q) + 2\mathrm{D}[Q(\delta|\boldsymbol{y}) \,\|\, P(\delta)] \right)$$

- Inner loop problem: Alternating minimization
  - Penalized least squares over $\boldsymbol{u}_*$ (eliminate γ)

F12a

Variational Bayesian Inference

$$\log P(\boldsymbol{y}) = \log \sum_{\delta} \int P(\boldsymbol{y}|\boldsymbol{u}) \underbrace{P(\boldsymbol{u}|\delta)}_{\text{mixture}} \underbrace{P(\delta)}_{\text{tree}} \, d\boldsymbol{u}$$

1. Super-Gaussian bounding
2. Factorize: $Q(\boldsymbol{u}, \delta|\boldsymbol{y}) = Q(\boldsymbol{u}|\boldsymbol{y})Q(\delta|\boldsymbol{y})$
3. Variational representation of $\log Z_Q$

$$-2\log P(\boldsymbol{y}) \leq \min_{\boldsymbol{z}} \left(\min_{Q(\delta|\boldsymbol{y}), \boldsymbol{u}_*, \gamma} \phi_{\boldsymbol{z}}(\boldsymbol{u}_*, \gamma, \langle\delta\rangle_Q) + 2\mathrm{D}[Q(\delta|\boldsymbol{y}) \,\|\, P(\delta)] \right)$$

- Inner loop problem: Alternating minimization
  - Penalized least squares over $\boldsymbol{u}_*$ (eliminate γ)
  - Belief propagation on tree for $Q(\delta|\boldsymbol{y})$ (eliminate γ)

F12b

No (Variational) Inference Without ...

Gaussian Variances

$$\boldsymbol{z} \leftarrow \mathrm{diag}^{-1}(\boldsymbol{B}\boldsymbol{A}^{-1}\boldsymbol{B}^T)$$

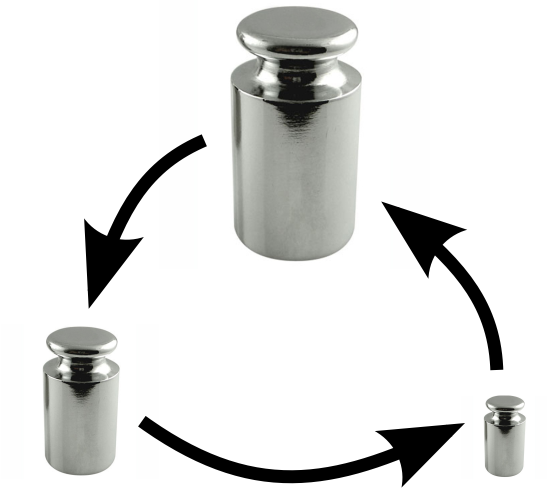
Linear Systems

$$\boldsymbol{u}_* \leftarrow \mathrm{argmin}\ \boldsymbol{u}_*^T\boldsymbol{A}\boldsymbol{u}_* - 2\boldsymbol{r}^T\boldsymbol{u}_*$$

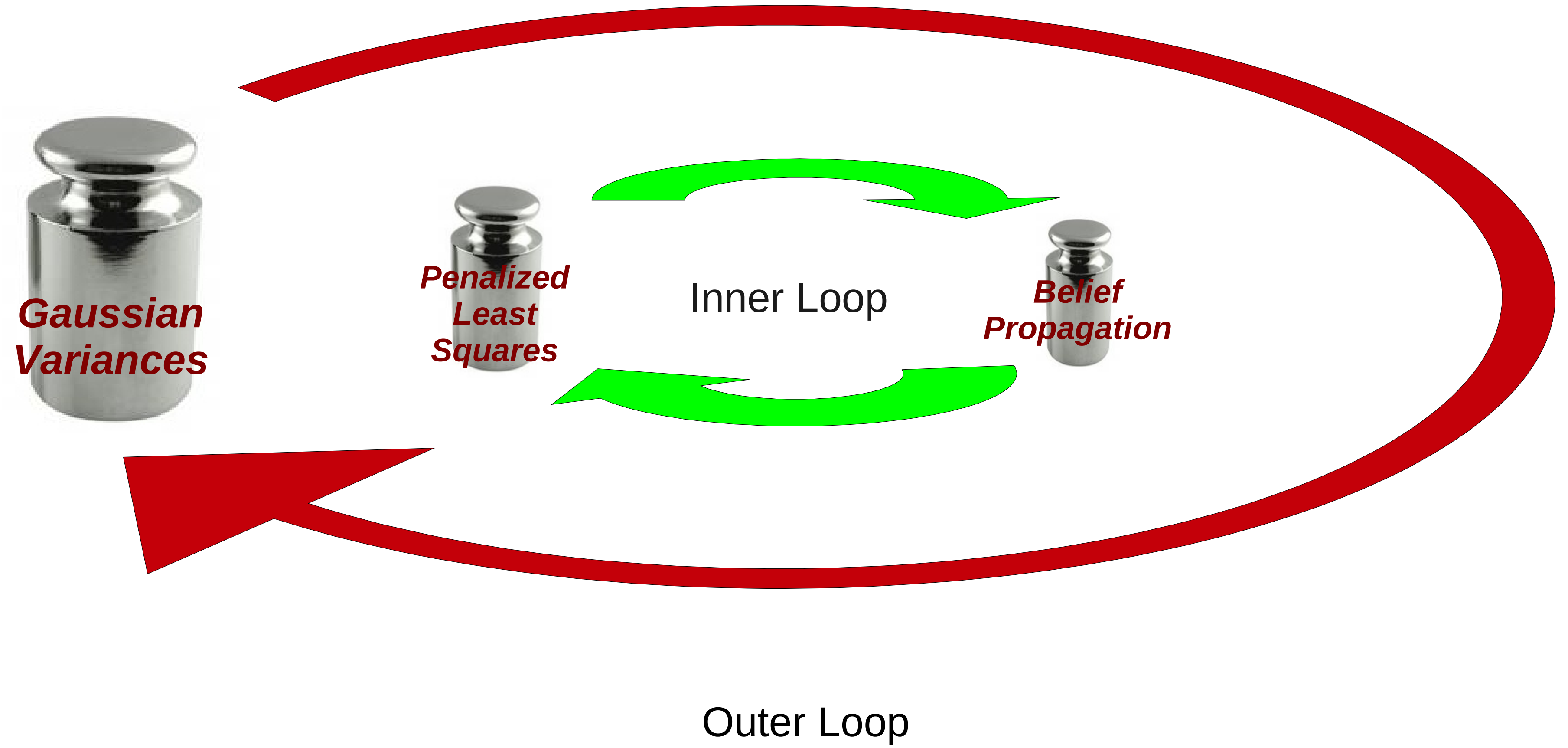
Peanuts

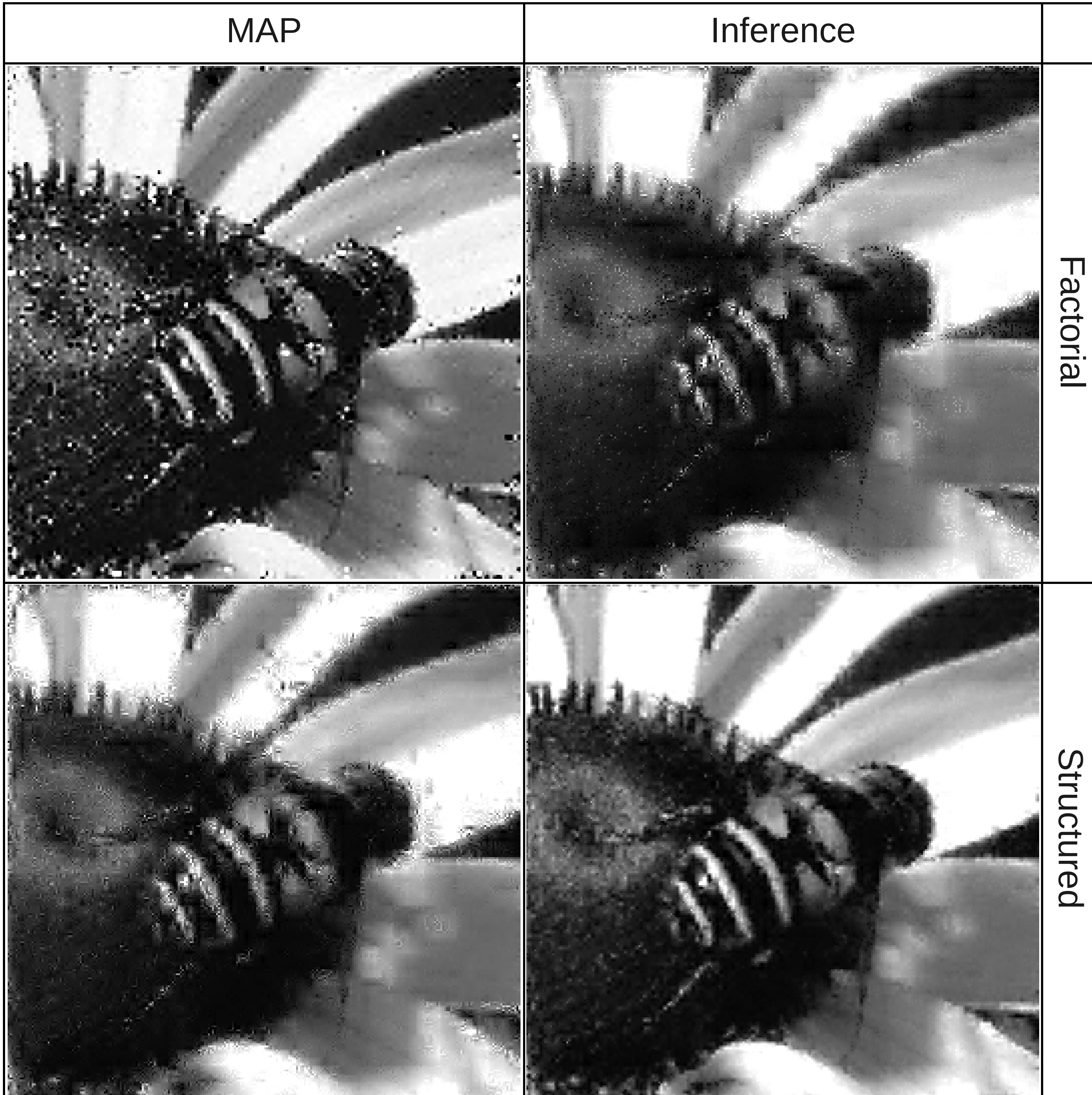
$O(q)$

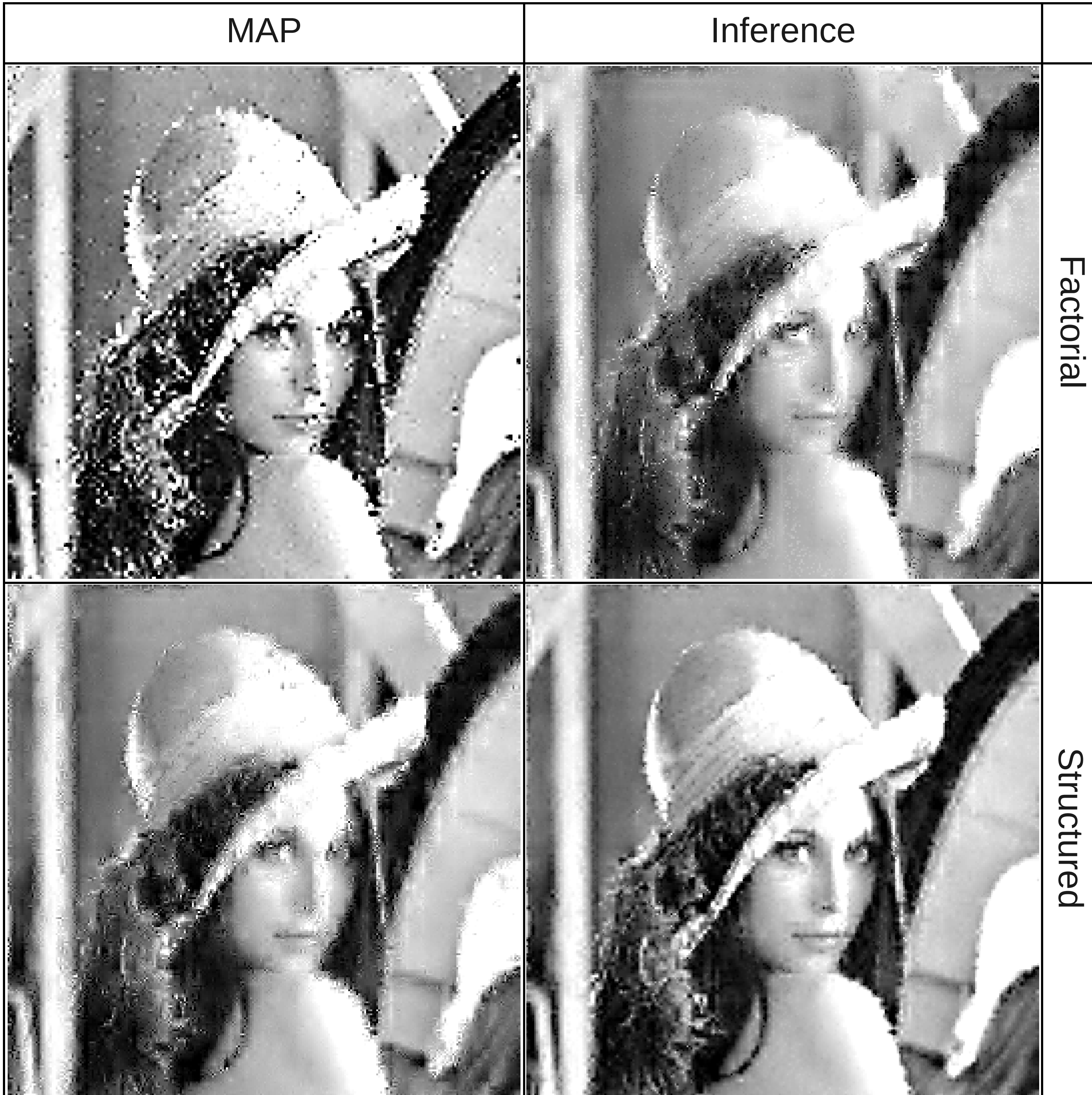
Why Inference Algorithms Can Be Slow



Attaining Scalability




MAP
Inference
Factorial
Structured


MAP
Inference
Factorial
Structured

Large Scale Variational Inference

- Inference beyond MAP estimation
  - Robust solutions for ill-posed problems
  - Model calibration from raw data
  - Bilinear models (blind deconvolution, dictionary learning)
  - Decision making (experimental design)
- Algorithms beyond belief propagation
  - Exploit workhorses from computational mathematics
  - Reductions to convex optimization
  - Randomized techniques may be part of the solution
- What about expectation propagation?
  - More difficult to scale up
  - Talk at workshop

Seeger, Nickisch, AISTATS 2011

Try This At Home

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

Physics and Bayesian Machine Learning

- You love physics?
  - Predictive models for real-world phenomena
  - Intuitive analysis of large complex systems
  - Statistical evaluation by clever experiments

  You'll love Bayesian machine learning!
- Most Bayesian concepts come from (statistical) physics
- Challenges of our time
  - Huge, extremely complex datasets
  - Connectivity at all scales
  - Human-level recognition and decision-making

  Needs physics thinking more than ever

Bayesian Machine Learning @ EPFL

Care for new vistas? Postdoc, PhD, Internship . . .