Expectation-Propagation methods for scalable inference in imaging inverse problems

Dr Yoann Altmann

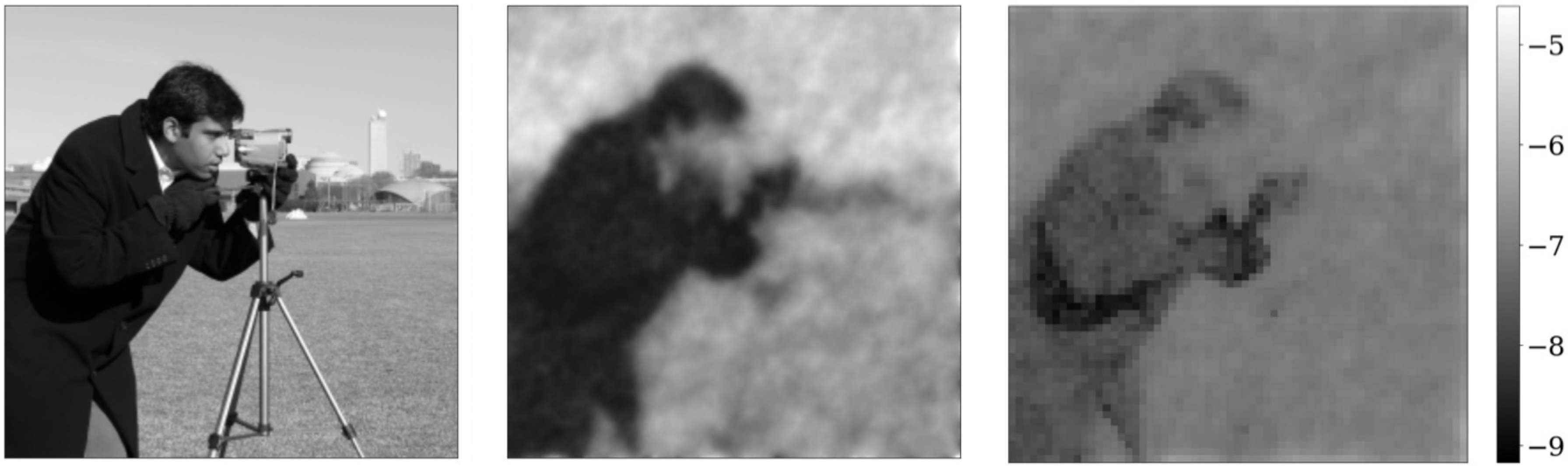
RAEng Research Fellow
School of Engineering and Physical Sciences
Heriot-Watt University, Edinburgh, U.K.

Joint work with: Dan Yao, Zeng Li, Abdullah Abdulaziz, Stephen McLaughlin

Nice, June, 25th 2022

LEADERS IN IDEAS AND SOLUTIONS

Why?

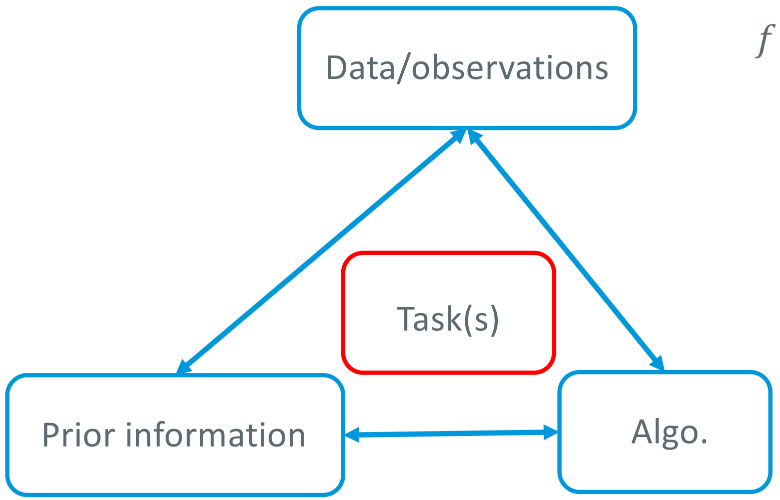


Outline

- Approximate methods and EP
- Linear spectral unmixing
- Image restoration
- Conclusion

Bayesian inference: key ingredients

$f(\boldsymbol{y}|\boldsymbol{x})$ - likelihood



$f(\boldsymbol{x}|\boldsymbol{\theta})$ - prior

$$f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})} \propto f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\theta})$$ - posterior

HERIOT WATT UNIVERSITY

Estimation strategies

- Point estimation
  - MAP: convex/non-convex optimization
  - Limited uncertainty quantification

- “Exact” methods: Importance sampling - SMC - MCMC

- Approximate methods
  - Proximal MCMC
  - Variational Bayes (VB) methods
  - Approximate message passing (AMP/EP)

HERIOT WATT UNIVERSITY

Bayesian modeling

- Exact model

$$f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\theta})}{f(\boldsymbol{y}|\boldsymbol{\theta})}$$

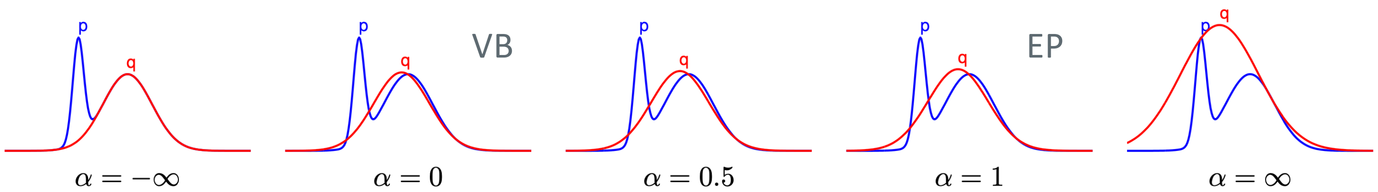
- Approximating distribution
  – VB/EP: Divergence-based

On the choice of divergence(s)

- Different families of similarity measures
- Classical choice: Kullback-Leibler

$$KL(q(x)\,||\,p(x)) = \int q(x) \ln \frac{q(x)}{p(x)}\, dx$$

- More general families
  - $\boldsymbol{\alpha}$-divergences



Extracted from Minka (2005)

Mean-Field Variational Bayes (MFVB)

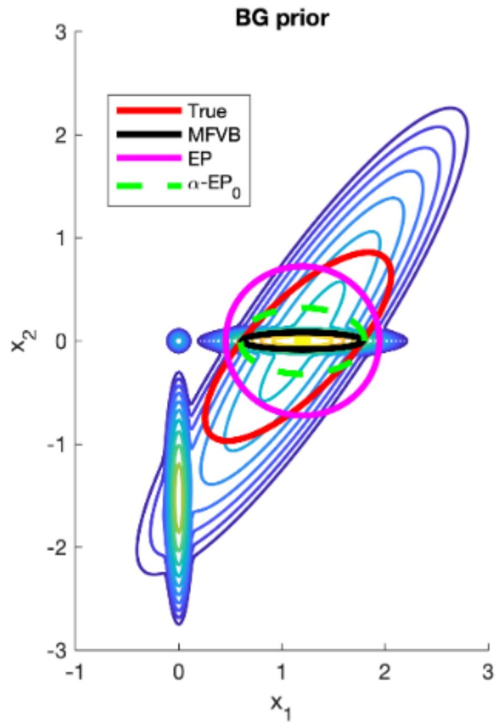
$f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \approx q(\boldsymbol{x})$ (often using mean field approx.)

$$\min_{q(\boldsymbol{x})} KL\big(q(\boldsymbol{x}) \,||\, f(\boldsymbol{x}|\boldsymbol{y}, \theta)\big)$$

$$KL(q(x) \,||\, p(x)) = \int q(x) \ln \frac{q(x)}{p(x)} \, dx$$

Expectation-Propagation

$$f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) = f(\boldsymbol{y}|\boldsymbol{x})f(\boldsymbol{x}|\boldsymbol{\theta})$$
$q(\boldsymbol{x}) \propto q_1(\boldsymbol{x})q_0(\boldsymbol{x})$: user-defined



- Based on the reverse KL-divergence

$$\min_{Z,\, q(\boldsymbol{x})\in\mathcal{F}} KL\left(f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\theta}) \,||\, Z_{\boldsymbol{y},\boldsymbol{\theta}}\, q(\boldsymbol{x})\right)$$

$$Z_{\boldsymbol{y},\boldsymbol{\theta}} \approx f(\boldsymbol{y}|\boldsymbol{\theta})$$

EP: preserves better the marginals than MFVB

EP: Iterative minimization

$$\min_{Z,\, q(\boldsymbol{x})} KL\left(f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\theta}) \;||\; Z_{y,\theta}\, q(\boldsymbol{x})\right)$$

Actual model:

$f(\boldsymbol{y}, \boldsymbol{x}|\theta) = f(\boldsymbol{y}|\boldsymbol{x}) f(\boldsymbol{x}|\theta)$

Approximate model:

$q(\boldsymbol{x}) \propto q_1(\boldsymbol{x}) q_0(\boldsymbol{x})$

Repeat:

- $\min_{q_1(\boldsymbol{x})} KL\left(f(\boldsymbol{y}|\boldsymbol{x}) q_0(\boldsymbol{x}) || q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\right)$
- $\min_{q_0(\boldsymbol{x})} KL\left(q_1(\boldsymbol{x}) f(\boldsymbol{x}|\theta) || q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\right)$

… until convergence

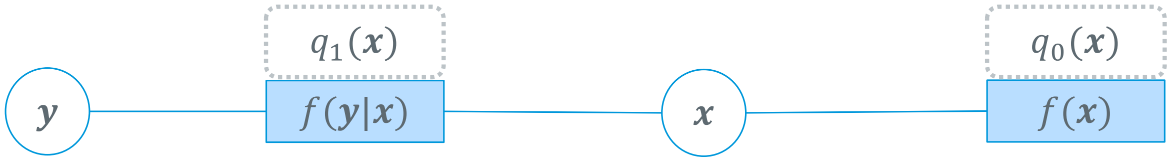
Can be applied with more than 2 factors

Divergence minimization

- With Gaussian approximations: moment matching
  Most challenging step: moments of the *tilted distributions*
  - $f(\boldsymbol{y}|\boldsymbol{x})q_0(\boldsymbol{x})$ and $q_1(\boldsymbol{x})f(\boldsymbol{x}|\theta)$

- Can be simplified using the structure of $q(\boldsymbol{x})$
  - Additional constraints
  - Tradeoff accuracy/complexity

Message passing interpretation



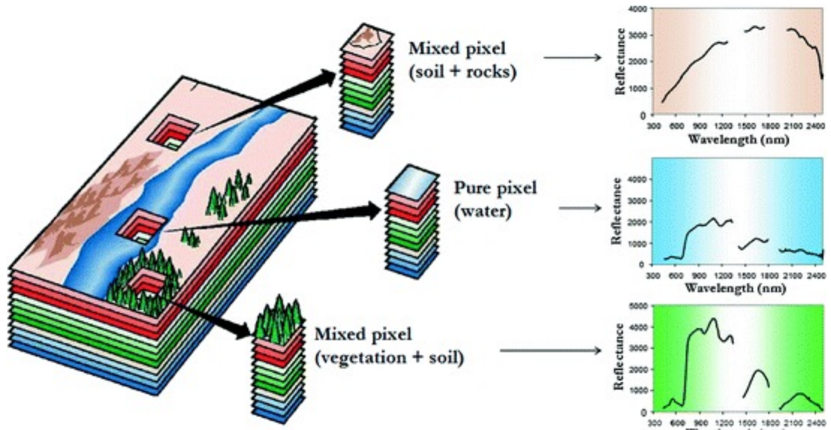
$$\min_{q_0(\boldsymbol{x})} KL(q_1(\boldsymbol{x})f(\boldsymbol{x}) \,||\, q_1(\boldsymbol{x})q_0(\boldsymbol{x}))$$

$$\min_{q_1(\boldsymbol{x})} KL(f(\boldsymbol{y}|\boldsymbol{x})q_0(\boldsymbol{x}) \,||\, q_1(\boldsymbol{x})q_0(\boldsymbol{x}))$$

Outline

- Approximate methods and EP
- Linear spectral unmixing
- Image restoration using TV prior
- Conclusion

EP for source separation



$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e} \Rightarrow f(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) = \prod_{m=1}^{M} f(y_m|\boldsymbol{a}_m^T x)$$

$$f(\boldsymbol{x}) = \prod_{n=1}^{N} f_n(x_n)$$

- $\boldsymbol{A}$: sources, spectral signatures…
- $\boldsymbol{x} = [x_1, \dots, x_N]^T$: fractions, amount of each source
- Arbitrary observation noise (Gaussian, Poisson,…)
- Separable prior model (not necessarily log-concave)
- $N \ll M$ (and N small)

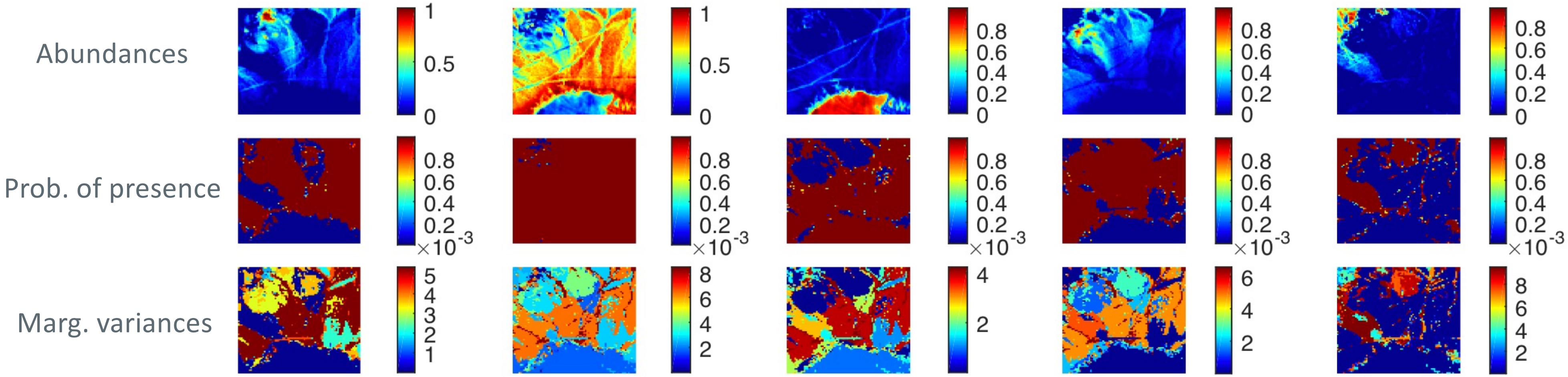
EP with compact graph

- Simple model: 2 updates

$$\min_{q_0(\boldsymbol{x})} KL(q_1(\boldsymbol{x})f(\boldsymbol{x}) \,||q_1(\boldsymbol{x})q_0(\boldsymbol{x}) \,)$$
$$\min_{q_1(\boldsymbol{x})} KL(f(\boldsymbol{y}|\boldsymbol{Ax})q_0(\boldsymbol{x})||q_1(\boldsymbol{x})q_0(\boldsymbol{x}) \,)$$

- More challenging with non-Gaussian (e.g. Poisson) noise…

Hyperspectral unmixing

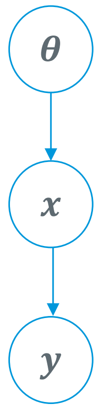


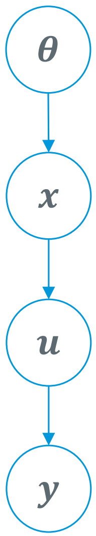
Sparse unmixing using spike-and-slab prior

Z. Li et al. "Sparse Linear Spectral Unmixing of Hyperspectral images using Expectation-Propagation", IEEE TGRS, 2021.

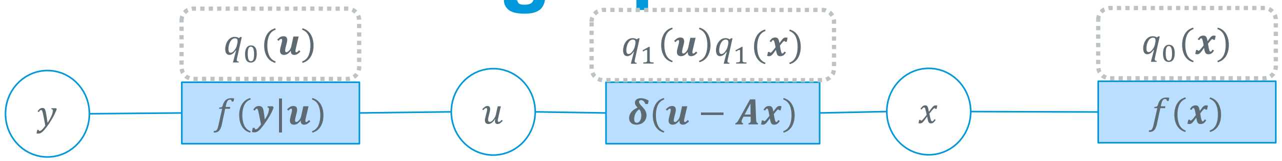
EP with extended graphs

- Extended models: split into smaller problems
  - $f(\boldsymbol{x}|\boldsymbol{\theta})$
  - $f(\boldsymbol{u}|\boldsymbol{x}) = \boldsymbol{\delta}(\boldsymbol{u} - \boldsymbol{A}\boldsymbol{x})$
  - $f(\boldsymbol{y}|\boldsymbol{u})$
- $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = \int f(\boldsymbol{u}, \boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) d\boldsymbol{u}$
  - $q(\boldsymbol{x}) = \int q(\boldsymbol{u}, \boldsymbol{x}) d\boldsymbol{u}$





EP with extended graphs



$$\min_{q_0(\boldsymbol{u})} \quad KL(f(\boldsymbol{y}|\boldsymbol{u})q_1(\boldsymbol{u})||q_0(\boldsymbol{u})q_1(\boldsymbol{u}))$$

$$\min_{q_1(\boldsymbol{u}),q_1(\boldsymbol{x})} \quad KL\left(\boldsymbol{\delta}(\boldsymbol{u}-\boldsymbol{Ax})\, q_0(\boldsymbol{u})q_0(\boldsymbol{x})||q_1(\boldsymbol{u})q_1(\boldsymbol{x})q_0(\boldsymbol{u})q_0(\boldsymbol{x})\right)$$

$$\min_{q_0(\boldsymbol{x})} KL(q_1(\boldsymbol{x})f(\boldsymbol{x})\, ||q_1(\boldsymbol{x})q_0(\boldsymbol{x}))$$

- Choosing $q_0(\boldsymbol{u}), q_0(\boldsymbol{x})$ and $q_1(\boldsymbol{u},\boldsymbol{x})$
  - Tradeoff accuracy/complexity
  - Problem-dependent

Links to traditional splitting methods

- Similar schemes as ADMM for MAP estimation
  - Local MAP estimation $\rightarrow$ local MMSE estimation
  - Local uncertainty estimation
  - Allows automatic adjustment of the splitting parameters

Outline

- Approximate methods and EP
- Linear unmixing
- Image restoration
- Conclusion

EP for imaging inverse problems

$$y = Hx + e \Rightarrow f(y|Hx) = \prod_{m=1}^{M} f(y_m|h_m^T x)$$

- $f(x)$: image prior
- H: convolution, subsampling, sensing matrix
- x: image to be recovered
- Arbitrary observation noise (Gaussian, Poisson,…)

Challenges:

- High-dimensional
- Non-convex priors

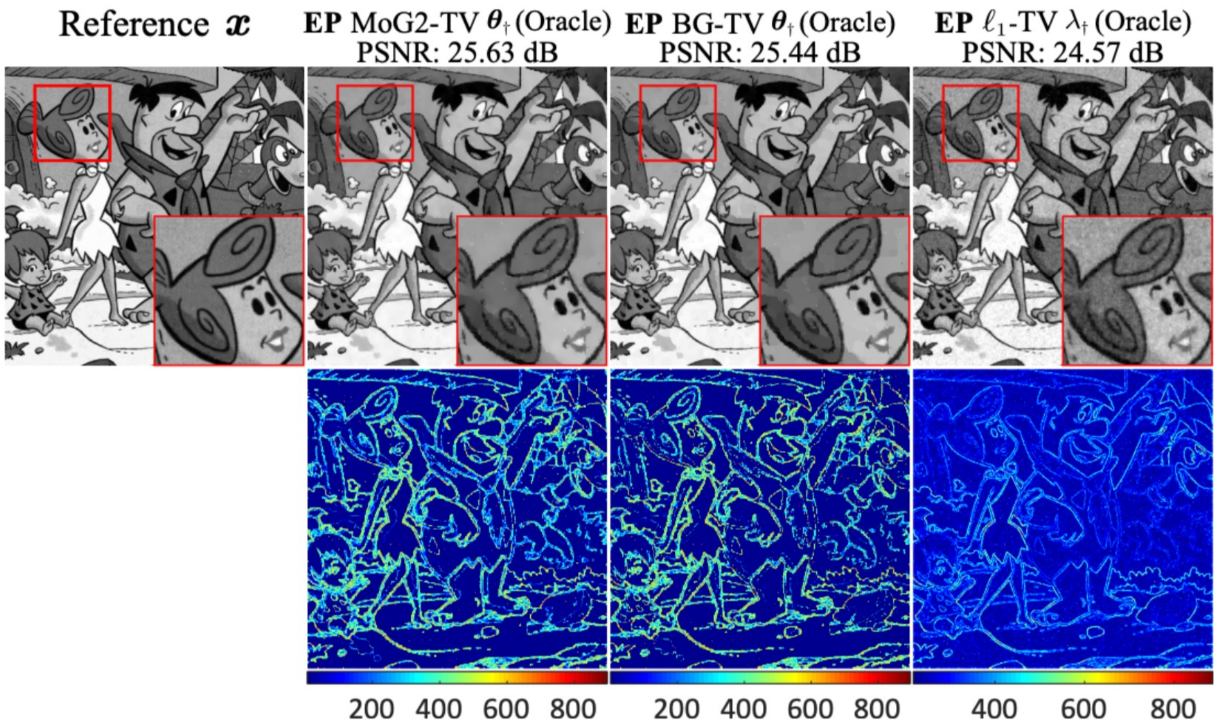
- Avoiding handling large covariance matrices

EP with Total variation priors

$$f_x(\boldsymbol{x}|\boldsymbol{\theta}) \propto \prod_{(i,j)\in V} \phi\left(x_i - x_j; \boldsymbol{\theta}\right)$$

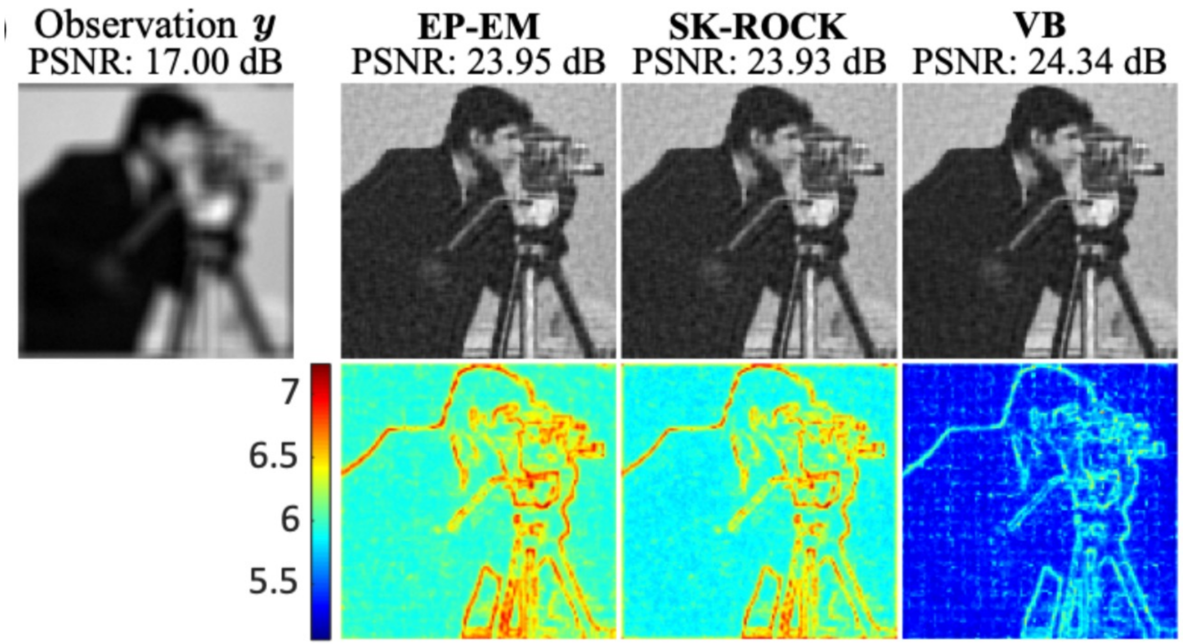
- Classical anisotropic TV: $\phi\left(x_i - x_j; \boldsymbol{\theta}\right) = \exp\left(-\lambda\ \left|x_i - x_j\right|\right), \boldsymbol{\theta} = \lambda$
- MoG prior: $\phi\left(x_i - x_j; \boldsymbol{\theta}\right)$ is a Gaussian mixture
- Special case: Bernoulli-Gaussian (BG) mixture (spike-and-slab)

EP with Total variation priors (denoising)



HERIOT WATT UNIVERSITY

EP with Total variation priors (deconvolution)



Hyperparameter estimation

- What if $\boldsymbol{\theta}$ is unknown in $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$?
- Option 1:

$$f(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) f(\boldsymbol{x}|\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

Joint estimation of $(\boldsymbol{x}, \boldsymbol{\theta})$: MCMC or EP (extended model)

Hyperparameter estimation

- What if $\boldsymbol{\theta}$ is unknown in $f(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta})$?
- Option 2:

$$f\left(\boldsymbol{x}\middle|\boldsymbol{y}, \widehat{\boldsymbol{\theta}}\right) = \frac{f(\boldsymbol{y}|\boldsymbol{x})f\left(\boldsymbol{x}\middle|\widehat{\boldsymbol{\theta}}\right)}{f(\boldsymbol{y})}$$

where $\widehat{\boldsymbol{\theta}} = \operatorname{argmax} f(\boldsymbol{\theta}|\boldsymbol{y})$ (or $f(\boldsymbol{y}|\boldsymbol{\theta})$).

Hyperparameter estimation

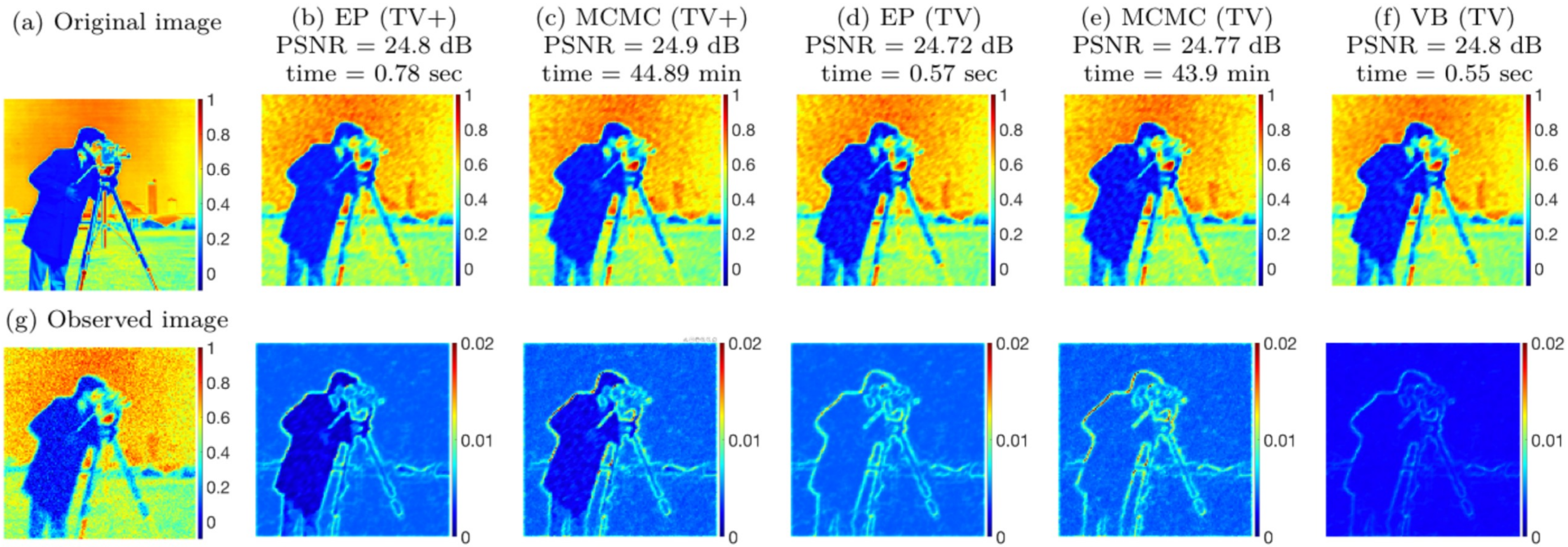
- Finding $\widehat{\boldsymbol{\theta}} = \text{argmax}\, f(\boldsymbol{\theta}|\boldsymbol{y})$ can often be done (in principle) using EM-like methods
  - Requires expectations w.r.t. $f(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ (during the E-step)
- Idea: replace by expectations w.r.t $q(\boldsymbol{x})$
  - Like variational-EM: EP-EM

Application to blind deconvolution

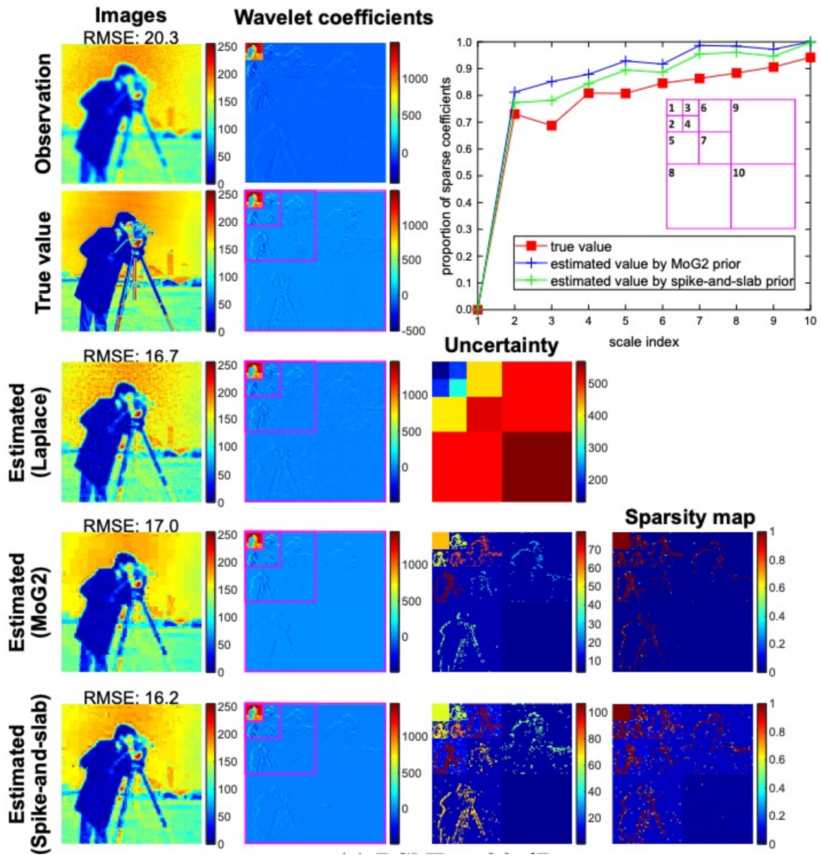
- Deconvolution with Gaussian noise and TV prior
  - Comparison with existing alternative approaches
    - VB and proximal MCMC

- Here, the noise variance and TV hyperparameter are known, the blur is unknown.

Deconvolution results (TV prior)



Using structured sparsity in transformed domains



Beyond KL divergences: Power EP

EP

$$\min_{q_0(\boldsymbol{x})} KL(q_1(\boldsymbol{x}) f(\boldsymbol{x}) \,||q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\,)$$

$$\min_{q_1(\boldsymbol{x})} KL(f(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) q_0(\boldsymbol{x}) || q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\,)$$

Power EP

$$\min_{q_0(\boldsymbol{x})} D_{\alpha_0}(q_1(\boldsymbol{x}) f(\boldsymbol{x}) \,||q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\,)$$

$$\min_{q_1(\boldsymbol{x})} D_{\alpha_1}(f(\boldsymbol{y}|\boldsymbol{A}\boldsymbol{x}) q_0(\boldsymbol{x}) || q_1(\boldsymbol{x}) q_0(\boldsymbol{x})\,)$$

Conclusion

- Variational inference possible beyond VB
  - EP: Scalable/distributed
  - Fast convergence (no guarantees)
- Good estimation performance
- But can be difficult to implement (need a bit of practice)
  - How to split?
  - Level of approximation
  - More expensive than MFVB
- How to detect/manipulate correlations in high dimensions?

Thanks for your attention!

Contact: Y.Altmann@hw.ac.uk