Hyper-parameters Estimation in Bayesian Models

Guohui Song

Old Dominion University

Joint work with Feng Yu (UT El Paso) and Lixin Shen (Syracuse University)

Linear Inverse Problems

- Consider a linear inverse problem: $\boldsymbol{y} = \mathsf{F}\boldsymbol{x} + \boldsymbol{\epsilon}$ (observations: $\boldsymbol{y}$; unknown parameters: $\boldsymbol{x}$)
  - It is usually ill-posed.
  - We need some extra/prior information on $\boldsymbol{x}$.

Linear Inverse Problems

- [x] Consider a linear inverse problem: $\boldsymbol{y} = \mathsf{F}\boldsymbol{x} + \boldsymbol{\epsilon}$ (observations: $\boldsymbol{y}$; unknown parameters: $\boldsymbol{x}$)
  - It is usually ill-posed.
  - We need some extra/prior information on $\boldsymbol{x}$.

- [ ] Regularization approach: impose a regularization/penalty term $R(\boldsymbol{x})$

$$\arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \lambda R(\boldsymbol{x})$$

  - Ridge regression: $R(\boldsymbol{x}) = \|\boldsymbol{x}\|^2$
  - Lasso: $R(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$
  - Group Lasso: $R(\boldsymbol{x}) = \sum_{g=1}^{G} \|\boldsymbol{x}_g\|_2$

Bayesian Models

- Bayesian approach: introduce a prior distribution on $\boldsymbol{x}$
  - likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\mathsf{F}\boldsymbol{x}, \sigma^2\mathrm{I})$
  - prior: $p(\boldsymbol{x})$
  - compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$

Bayesian Models

☑ Bayesian approach: introduce a prior distribution on $\boldsymbol{x}$

- ⊘ likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\mathsf{F}\boldsymbol{x}, \sigma^2\mathrm{I})$
- ⊘ prior: $p(\boldsymbol{x})$
- ⊘ compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$

☐ MAP estimate

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 - \log p(\boldsymbol{x})$$

- ○ Equivalent to the regularization approaches for appropriate choices of the prior.

Bayesian Models

☑ Bayesian approach: introduce a prior distribution on $\boldsymbol{x}$

- ⊘ likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\mathsf{F}\boldsymbol{x}, \sigma^2\mathrm{I})$
- ⊘ prior: $p(\boldsymbol{x})$
- ⊘ compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$

☑ MAP estimate

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 - \log p(\boldsymbol{x})$$

- ⊘ Equivalent to the regularization approaches for appropriate choices of the prior.

☐ Uncertainty quantification

- ○ The posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ provides a full distribution of $\boldsymbol{x}$.

Bayesian Models

- Bayesian approach: introduce a prior distribution on $\boldsymbol{x}$
  - likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\mathsf{F}\boldsymbol{x}, \sigma^2\mathrm{I})$
  - prior: $p(\boldsymbol{x})$
  - compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$
- MAP estimate

$$\arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \frac{1}{2\sigma^2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 - \log p(\boldsymbol{x})$$

  - Equivalent to the regularization approaches for appropriate choices of the prior.
- Uncertainty quantification
  - The posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ provides a full distribution of $\boldsymbol{x}$.

Bayesian Models

- [x] Bayesian approach: introduce a prior distribution on $\boldsymbol{x}$
  - likelihood: $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}|\mathsf{F}\boldsymbol{x}, \sigma^2\mathrm{I})$
  - prior: $p(\boldsymbol{x})$
  - compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})}$
- [ ] Challenges
  - The computation of $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}$ is usually intractable.
  - The choice of the prior distribution is crucial.

Hierarchical Bayesian Learning

- Consider a hierarchical prior (scale mixtures of Gaussians)
  - $p(\boldsymbol{x}|\boldsymbol{\gamma}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathrm{diag}(\gamma_1, \ldots, \gamma_n))$ hyper-parameters
  - impose a hyper-prior on $\boldsymbol{\gamma}$

Hierarchical Bayesian Learning

☑ Consider a hierarchical prior (scale mixtures of Gaussians)

- ⊘ $p(\boldsymbol{x}|\boldsymbol{\gamma}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathrm{diag}(\gamma_1, \ldots, \gamma_n))$ hyper-parameters
- ⊘ impose a hyper-prior on $\boldsymbol{\gamma}$

☐ Benefits

- ○ Scale mixture of Gaussians contains all symmetric stable distributions, Laplace distributions, logistic distributions, and exponential power distributions.
- ○ The conditional posterior $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\gamma})$ is analytically tractable.
- ○ Individualized hyper-parameters encode sparsity.

Hierarchical Bayesian Learning

- ☑ Consider a hierarchical prior (scale mixtures of Gaussians)
  - ⊘ $p(\boldsymbol{x}|\boldsymbol{\gamma}) = \mathcal{N}(\boldsymbol{x}|\mathbf{0}, \mathrm{diag}(\gamma_1, \ldots, \gamma_n))$ hyper-parameters
  - ⊘ impose a hyper-prior on $\boldsymbol{\gamma}$
- ☑ Benefits
  - ⊘ Scale mixture of Gaussians contains all symmetric stable distributions, Laplace distributions, logistic distributions, and exponential power distributions.
  - ⊘ The conditional posterior $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\gamma})$ is analytically tractable.
  - ⊘ Individualized hyper-parameters encode sparsity.
- ☐ Challenge: how to estimate the hyper-parameters $\boldsymbol{\gamma}$?
  - ○ The dimension of hyper-parameters $\boldsymbol{\gamma}$ scales with $\boldsymbol{x}$.
  - ○ It is not easy to choose the hyper-prior.

Hyper-parameters Estimation

- Learning from data (Evidence approach)

$$\arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{y}|\boldsymbol{\gamma}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\gamma})d\boldsymbol{x}$$

Hyper-parameters Estimation

- [x] Learning from data (Evidence approach)

$$\arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{y}|\boldsymbol{\gamma}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\gamma})d\boldsymbol{x}$$

- [ ] Equivalent to

$$\arg\min_{\boldsymbol{\gamma}} \boldsymbol{y}^{\mathsf{T}}(\mathsf{S}(\boldsymbol{\gamma}))^{-1}\boldsymbol{y} + \log\det(\mathsf{S}(\boldsymbol{\gamma})), \quad \mathsf{S}(\boldsymbol{\gamma}) = \sigma^2\mathrm{I} + \mathsf{F}\,\mathrm{diag}(\boldsymbol{\gamma})\mathsf{F}^{\mathsf{T}}$$

  - It is a high-dimensional non-convex optimization problem
  - Existing methods: EM [Dempster et.al. 1977], MacKay [MacKay, 1992], Convex Bounding [Wipf and Nagarajan, 2009], etc.

Hyper-parameters Estimation

- [x] Learning from data (Evidence approach)

$$\arg\max_{\boldsymbol{\gamma}} p(\boldsymbol{y}|\boldsymbol{\gamma}) = \int p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{\gamma})d\boldsymbol{x}$$

- [ ] Equivalent to

$$\arg\min_{\boldsymbol{\gamma}} \boldsymbol{y}^{\mathsf{T}}(\mathsf{S}(\boldsymbol{\gamma}))^{-1}\boldsymbol{y} + \log\det(\mathsf{S}(\boldsymbol{\gamma})), \quad \mathsf{S}(\boldsymbol{\gamma}) = \sigma^2\mathrm{I} + \mathsf{F}\,\mathrm{diag}(\boldsymbol{\gamma})\mathsf{F}^{\mathsf{T}}$$

  - It is a high-dimensional non-convex optimization problem
  - Existing methods: EM [Dempster et.al. 1977], MacKay [MacKay, 1992], Convex Bounding [Wipf and Nagarajan, 2009], etc.

- [ ] Goal:develop more efficient algorithms

Alternating Minimization

- The objective function:

$$\min_{\boldsymbol{\gamma}} \boldsymbol{y}^{\mathsf{T}}(\mathsf{S}(\boldsymbol{\gamma}))^{-1}\boldsymbol{y} + \log\det(\mathsf{S}(\boldsymbol{\gamma})), \quad \mathsf{S}(\boldsymbol{\gamma}) = \sigma^2\mathrm{I} + \mathsf{F}\,\mathrm{diag}(\boldsymbol{\gamma})\mathsf{F}^{\mathsf{T}}$$

Alternating Minimization

☑ The objective function:

$$\min_{\boldsymbol{\gamma}} \boldsymbol{y}^{\mathsf{T}}(\mathsf{S}(\boldsymbol{\gamma}))^{-1}\boldsymbol{y} + \log\det(\mathsf{S}(\boldsymbol{\gamma})), \quad \mathsf{S}(\boldsymbol{\gamma}) = \sigma^2\mathrm{I} + \mathsf{F}\,\mathrm{diag}(\boldsymbol{\gamma})\mathsf{F}^{\mathsf{T}}$$

☐ Rewrite the first term: $\min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}}\,\mathrm{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x}$

Alternating Minimization

☑ The objective function:

$$\min_{\boldsymbol{\gamma}} \boldsymbol{y}^{\mathsf{T}}(\mathsf{S}(\boldsymbol{\gamma}))^{-1}\boldsymbol{y} + \log\det(\mathsf{S}(\boldsymbol{\gamma})), \quad \mathsf{S}(\boldsymbol{\gamma}) = \sigma^2\mathrm{I} + \mathsf{F}\,\mathrm{diag}(\boldsymbol{\gamma})\mathsf{F}^{\mathsf{T}}$$

☑ Rewrite the first term: $\min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}}\,\mathrm{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x}$

☐ Reformulate the objective function:

$$\min_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}}\,\mathrm{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$$

Alternating Minimization

☑ Reformulate the objective function:

$$\min_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \sigma^{-2} \|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$$

☐ Alternating Minimization:

Alternating Minimization

☑ Reformulate the objective function:

$$\min_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$$

☑ Alternating Minimization:

- ○ $\boldsymbol{x}$-update: $\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}((\boldsymbol{\gamma}^{(k)})^{-1})\boldsymbol{x}$

Alternating Minimization

☑ Reformulate the objective function:

$$\min_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$$

☑ Alternating Minimization:

- ⊘ $\boldsymbol{x}$-update: $\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}((\boldsymbol{\gamma}^{(k)})^{-1})\boldsymbol{x}$
- ○ $\boldsymbol{\gamma}$-update: $\boldsymbol{\gamma}^{(k+1)} = \arg\min_{\boldsymbol{\gamma}} (\boldsymbol{x}^{(k+1)})^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x}^{(k+1)} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$

Alternating Minimization

☑ Reformulate the objective function:

$$\min_{\boldsymbol{\gamma}} \min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$$

☑ Alternating Minimization:

- ⊘ $\boldsymbol{x}$-update: $\boldsymbol{x}^{(k+1)} = \arg\min_{\boldsymbol{x}} \sigma^{-2}\|\boldsymbol{y} - \mathsf{F}\boldsymbol{x}\|^2 + \boldsymbol{x}^{\mathsf{T}} \operatorname{diag}((\boldsymbol{\gamma}^{(k)})^{-1})\boldsymbol{x}$
- ⊘ $\boldsymbol{\gamma}$-update: $\boldsymbol{\gamma}^{(k+1)} = \arg\min_{\boldsymbol{\gamma}} (\boldsymbol{x}^{(k+1)})^{\mathsf{T}} \operatorname{diag}(\boldsymbol{\gamma}^{-1})\boldsymbol{x}^{(k+1)} + \log\det(\mathsf{S}(\boldsymbol{\gamma}))$

☐ The challenge comes from the log-determinant term.

- ○ Find a separable surrogate for it.
- ○ Use first-order Taylor expansion.

Alternating Minimization with Linearization

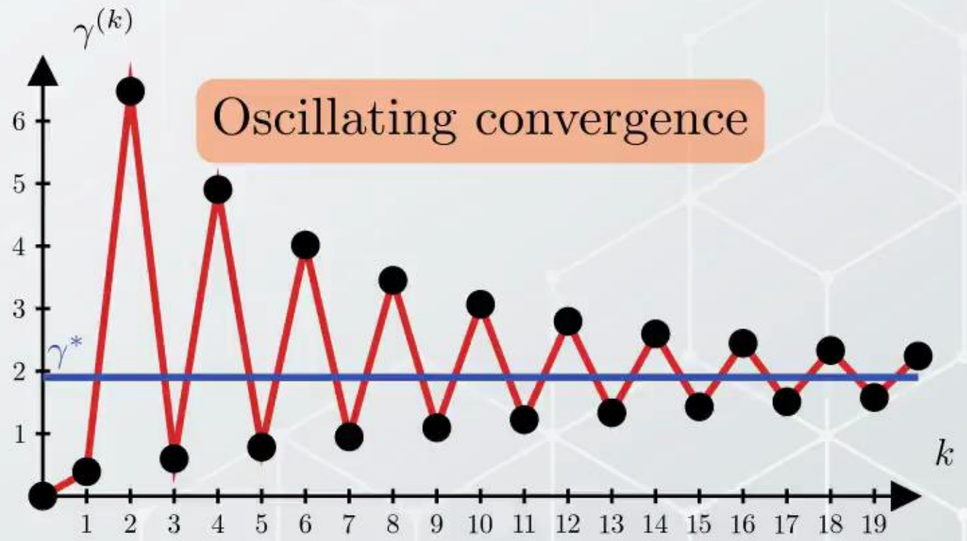
- Consider a linear approximation of $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$
  - We could update each γ_i separately.
  - We have a closed-form solution of each $\gamma_i^{(k+1)}$ in terms of $\gamma_i^{(k)}$

Alternating Minimization with Linearization

- ☑ Consider a linear approximation of $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$
  - ⊘ We could update each γ_i separately.
  - ⊘ We have a closed-form solution of each $\gamma_i^{(k+1)}$ in terms of $\gamma_i^{(k)}$
  - ○ We prove the convergence of the algorithm.
  - ○ The convergence might not be fast enough in some cases.

Alternating Minimization with Linearization

- Consider a linear approximation of $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$
  - We could update each γ_i separately.
  - We have a closed-form solution of each $\gamma_i^{(k+1)}$ in terms of $\gamma_i^{(k)}$
  - We prove the convergence of the algorithm.
  - The convergence might not be fast enough in some cases.



Alternating Minimization with Quadratic Surrogates

- Consider a quadratic surrogate for $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$
  - Add a penalty term $\tau\|\boldsymbol{\gamma} - \boldsymbol{\gamma}^{(k)}\|^2$ to the linear approximation.

Alternating Minimization with Quadratic Surrogates

- ☑ Consider a quadratic surrogate for $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$
  - ⊘ Add a penalty term $\tau\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(k)}\|^2$ to the linear approximation.
  - ○ Update each γ_i separately through minimizing a quadratic surrogate.
  - ○ Have a closed-form solution of each $\gamma_i^{(k+1)}$ in terms of $\gamma_i^{(k)}$

Alternating Minimization with Quadratic Surrogates

☑ Consider a quadratic surrogate for $\log\det(\mathsf{S}(\boldsymbol{\gamma}))$

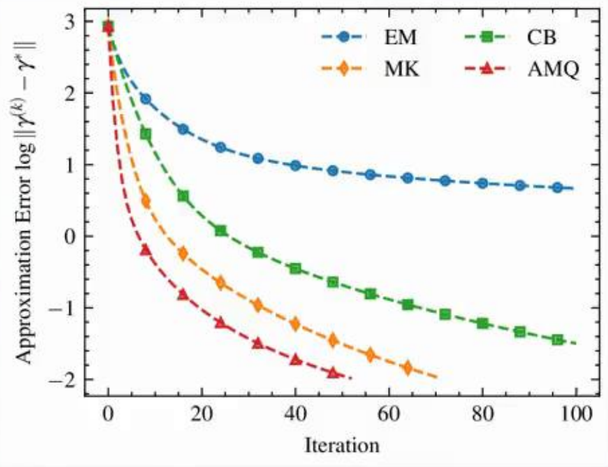
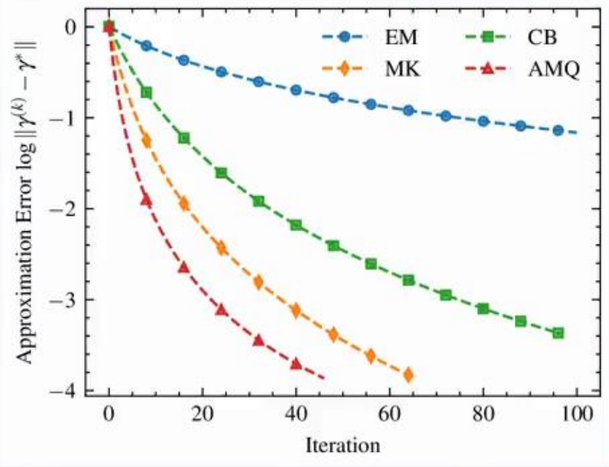
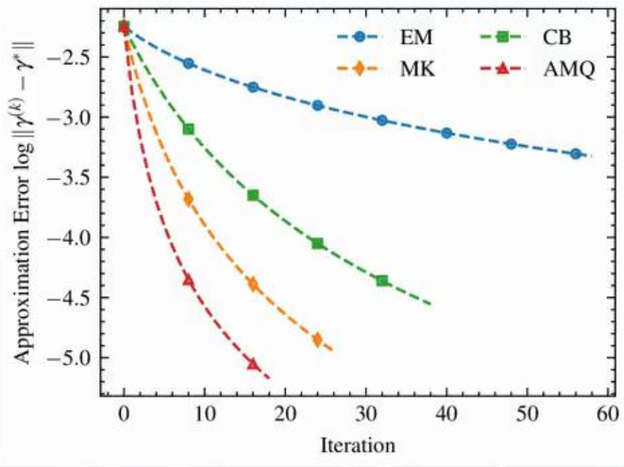
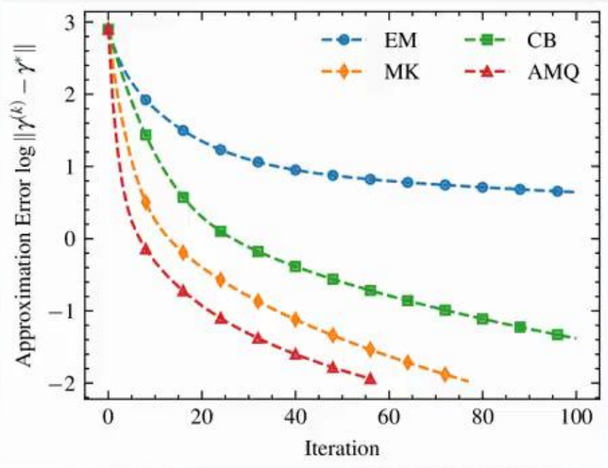
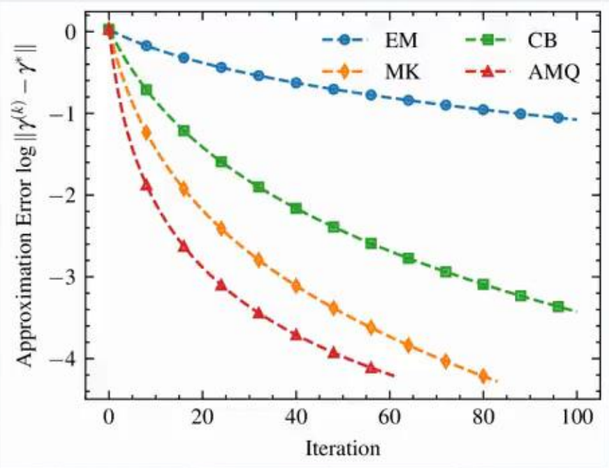
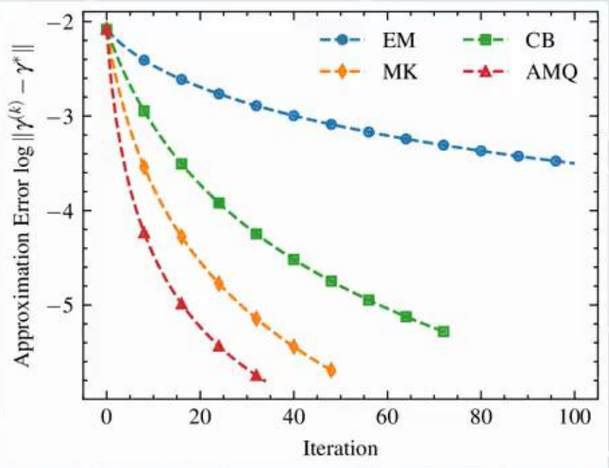
- ⊘ Add a penalty term $\tau\|\boldsymbol{\gamma}-\boldsymbol{\gamma}^{(k)}\|^2$ to the linear approximation.
- ⊘ Update each γ_i separately through minimizing a quadratic surrogate.
- ⊘ Have a closed-form solution of each $\gamma_i^{(k+1)}$ in terms of $\gamma_i^{(k)}$

Theorem [Feng, Shen, and S., 2024]

If τ_k is properly chosen, the proposed algorithm converges to a stationary point $\boldsymbol{\gamma}^*$ of the objective function $L(\boldsymbol{\gamma})$. That is, for each i, either $\gamma_i^*=0$ or $\frac{\partial L(\boldsymbol{\gamma}^*)}{\partial\gamma_i}=0$.

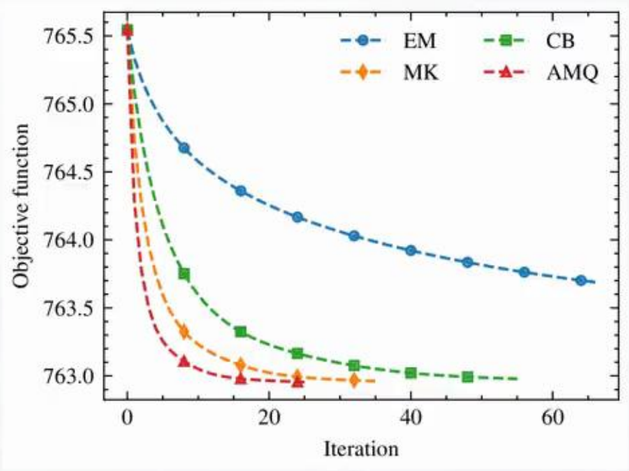
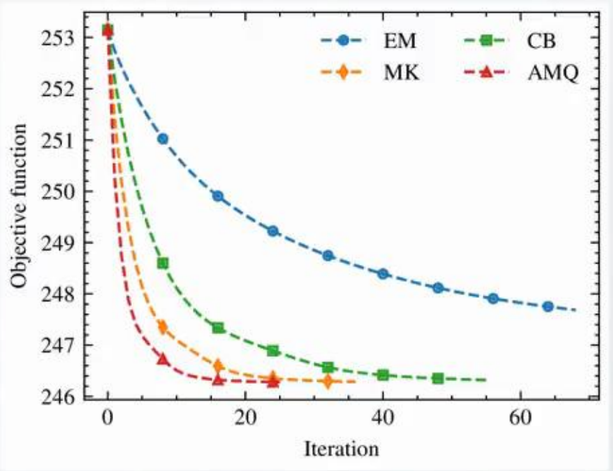
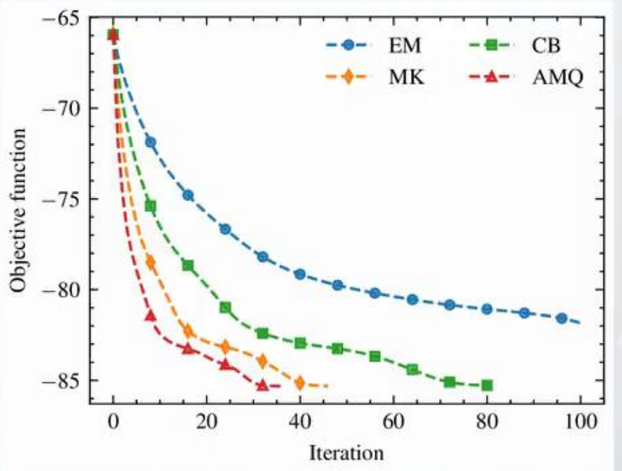
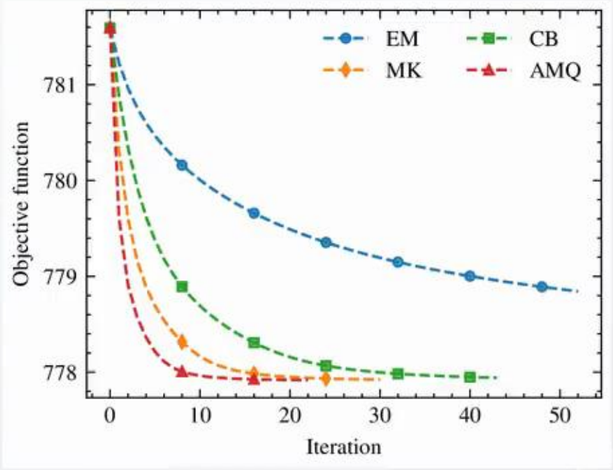
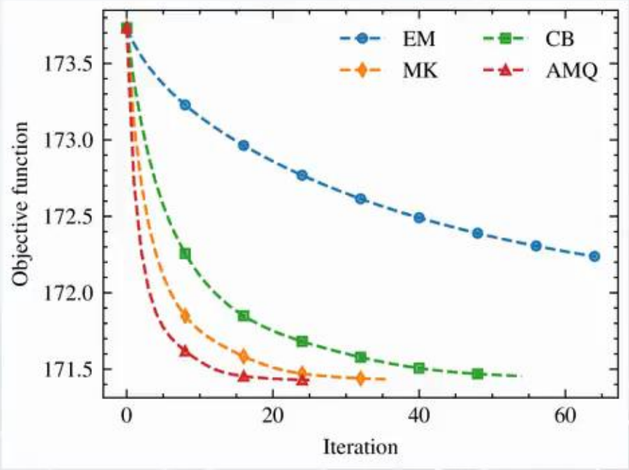
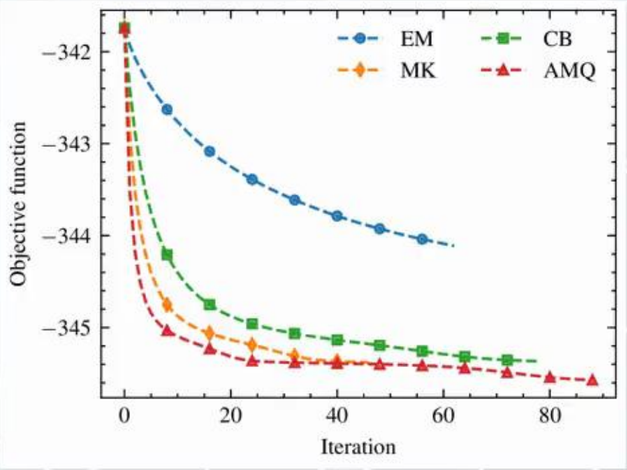
Numerical Results: Denoising

Numerical Results: Denoising

	$\sigma^2 = 10$	$\sigma^2 = 1$	$\sigma^2 = 0.1$
$s = 20\%$			
$s = 90\%$			

Numerical Results: Partial DCT

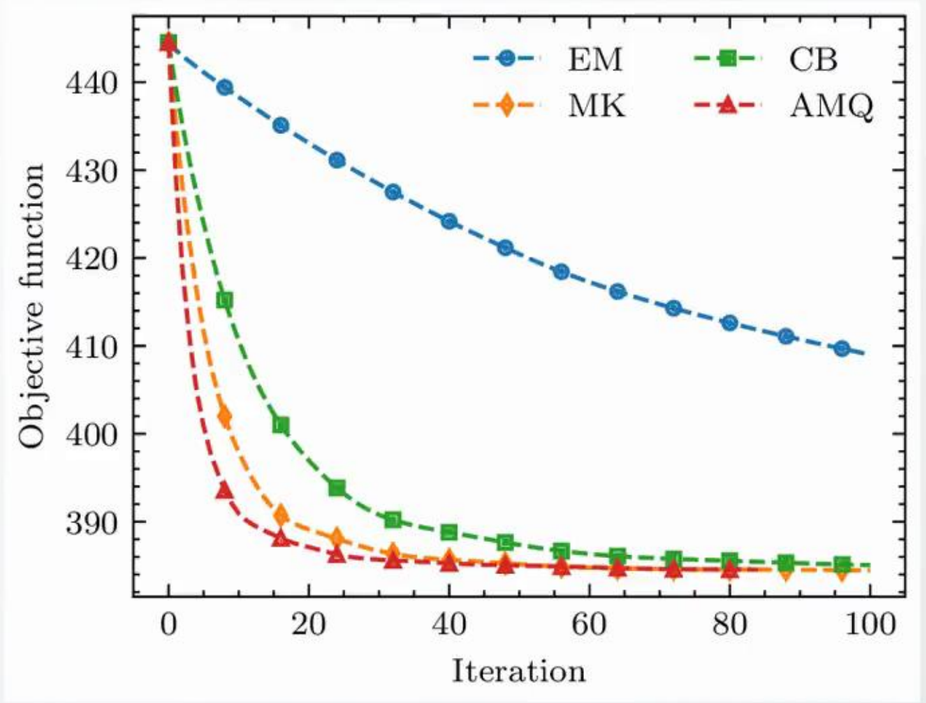
Numerical Results: Partial DCT

	$\sigma^2 = 10$	$\sigma^2 = 1$	$\sigma^2 = 0.1$
$s = 20\%$			
$s = 90\%$			

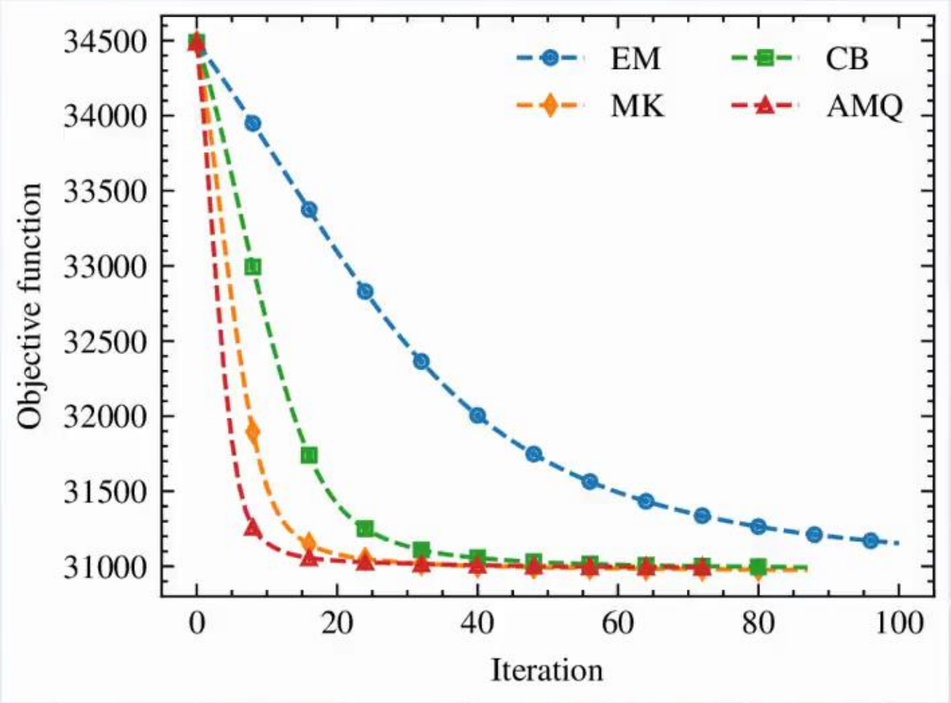
Numerical Results: SAR and EEG

Numerical Results: SAR and EEG

EEG



SAR



Summary

- Comments
  - Bayesian models offer uncertainty quantification.
  - Hierarchical Bayesian models provide flexibility and scalability in incorporating prior information.
  - We propose a new algorithm for hyper-parameters estimation in hierarchical (linear) Bayesian models.

Summary

☑ Comments

- Bayesian models offer uncertainty quantification.
- Hierarchical Bayesian models provide flexibility and scalability in incorporating prior information.
- We propose a new algorithm for hyper-parameters estimation in hierarchical (linear) Bayesian models.

☐ Issues

- Non-Gaussian distributions
- Non-linear models
- Challenge: more efficient sampling (Monte Carlo) algorithms

Thank you for your attention!

??

Guohui Song, gsong@odu.edu