Automatic Penalization of Model Complexity

Bayesian Inference adds Robustness Against Noise and Fitting too Many Components

Jesper Løve Hinrich, KU Food, Chemometrics and Analytical Technology.

Talk at dsk.2020, Nov. 5, 2020

KØBENHAVNS UNIVERSITET

Research Interests and Background

- Background
  - PhD (17'-20'), Section for Cognitive Systems, Technical University of Denmark (DTU)
  - Visiting Scholar (6 months), Department of Statistical Science, Duke University, NC, USA
  - Research Assistant (16'-17'), Section for Cognitive Systems, DTU
  - M.Sc. Eng. (14'-16'), Mathematical Modelling and Computation, DTU
- Research
  - Statistical Science and Machine Learning
    - Unsupervised and multi-modal learning, uncertainty quantification, and Bayesian statistics
  - Application areas
    - Chemistry and Chemometrics, Neuroscience, Computer Science
- Aims
  - Improving capabilities (automating, speedups).
  - Expanding capabilities (new insights).
  - Bridging statistics with domain experts.

Contact at jlh@food.ku.dk

or

jlhinrich

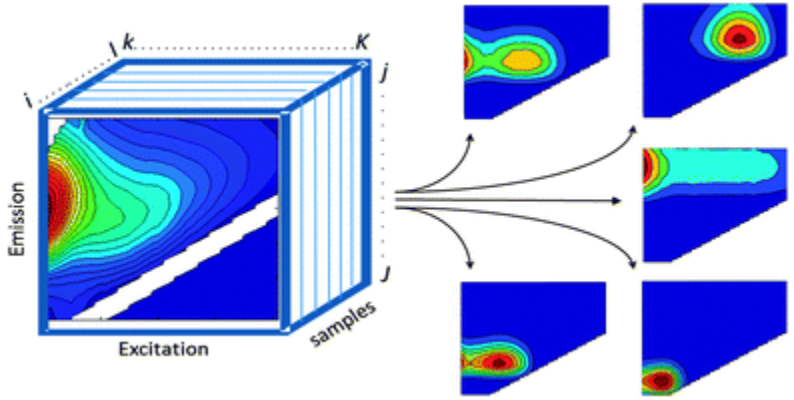
JesperLH

0000-0003-0258-7151

A Few Applications

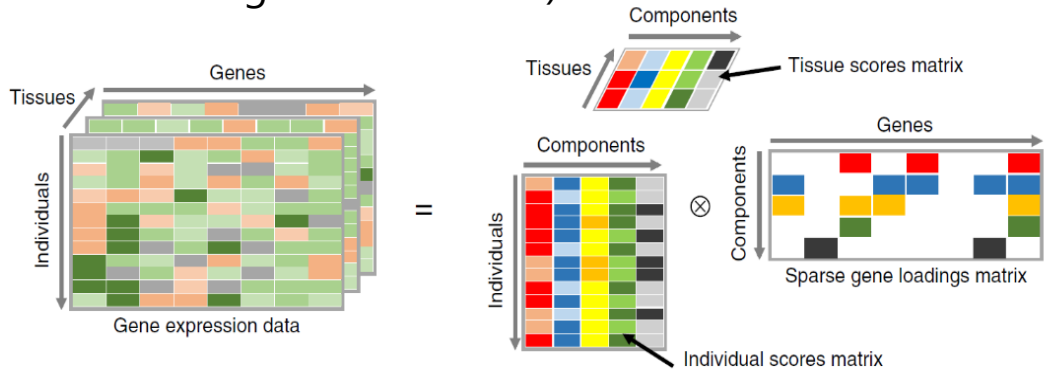
Fluorescence Spectroscopy

(emission spectra x excitation spectra x samples)



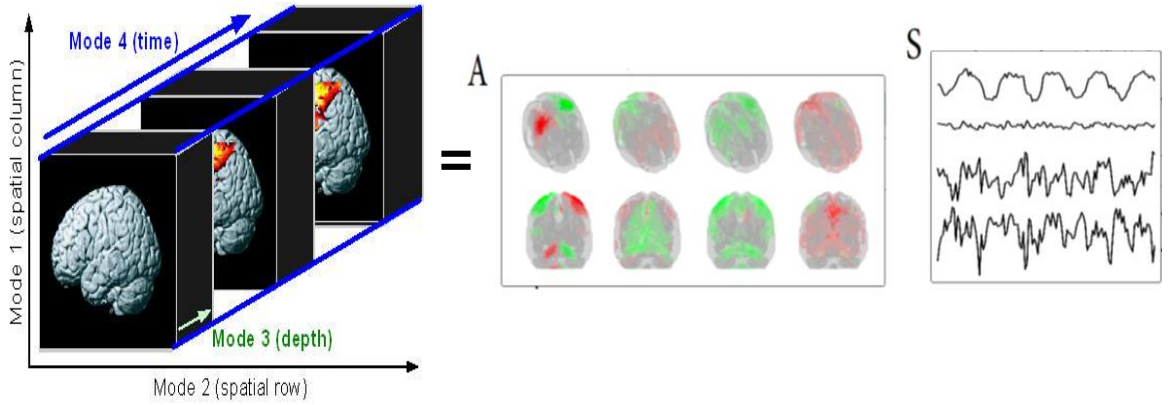
Gene expressions

(Individuals x genes x tissues)



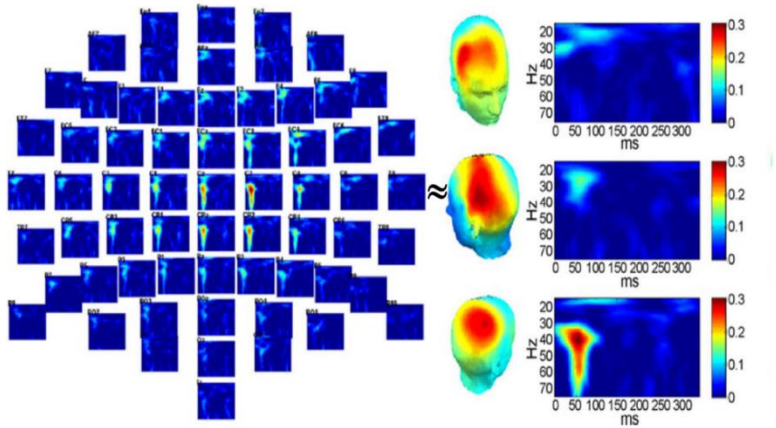
Functional MRI

(x-axis x y-axis x z-axis x time)



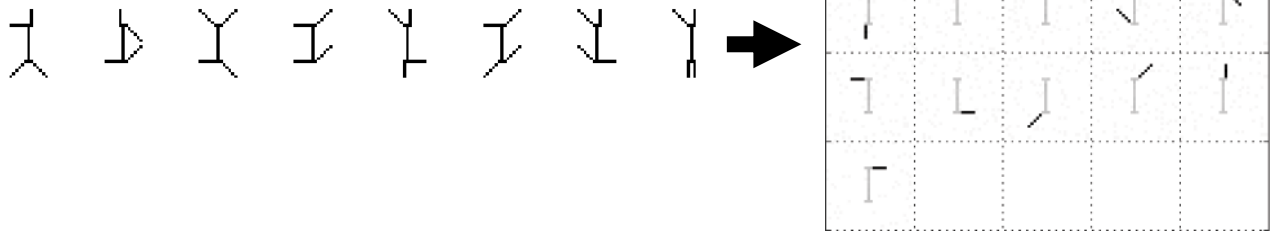
Electroencephalogram (EEG)

(channels x time x frequency)



A Few Applications

Computer Science
Parts based representation



Chemistry and Chemometrics

- Gas Chromatography
- Liquid Chromatography
- Fluorescence spectroscopy
- Near infrared spectroscopy
- 1D or 2D separation

Planetary and Space Science
ESA/ExoMars Trace Gas Orbiter
(~10m spectras, ~10k channels)

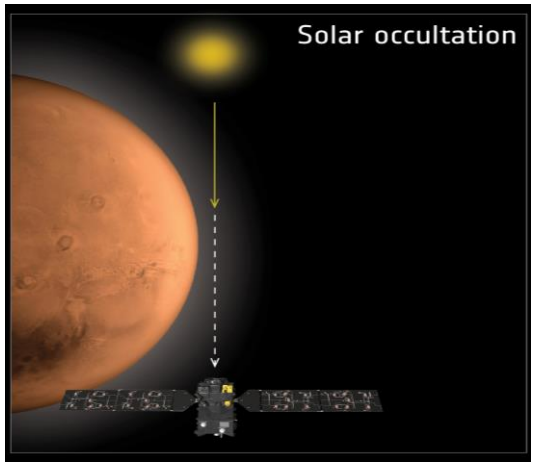


Image: ESA/ExoMars 392257 ID

Comprehensive Two-Dimensional Gas Chromotography (GCxGC)

- Gathering GCxGC data (per sample)

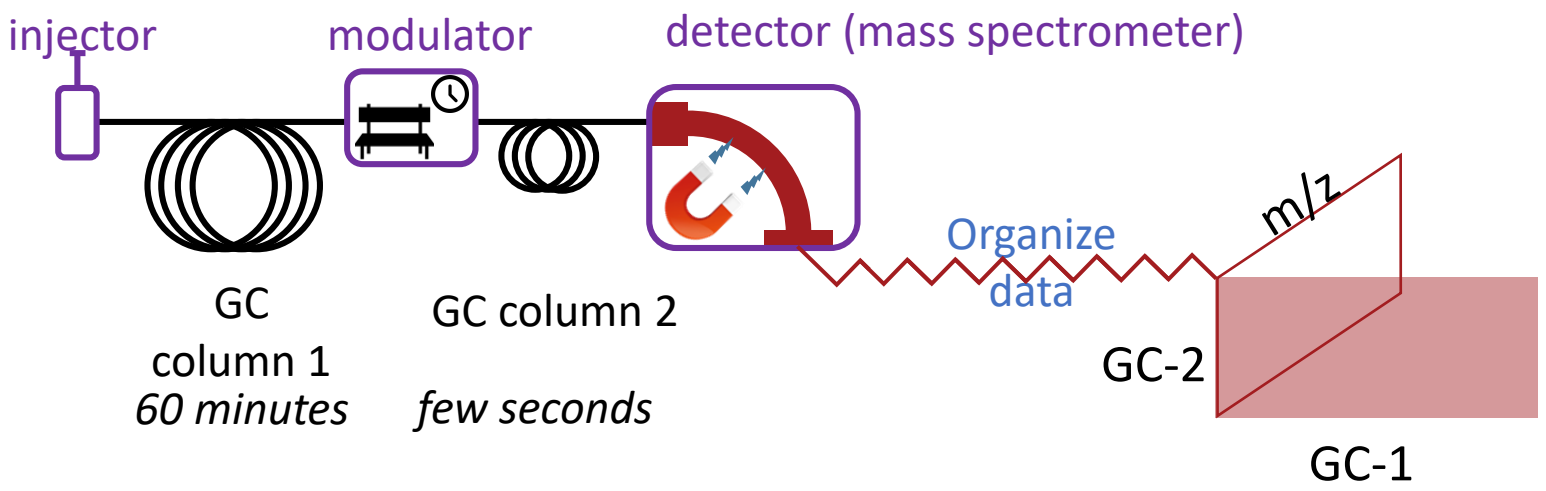
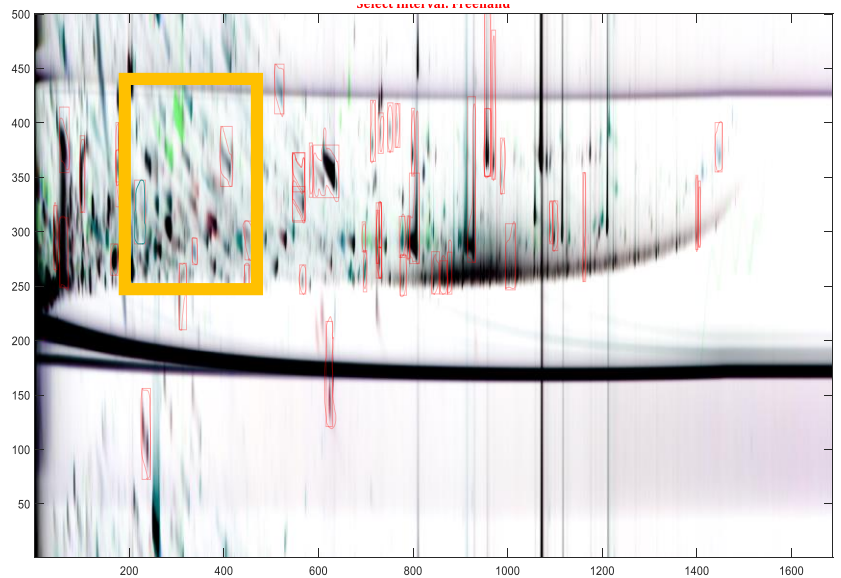


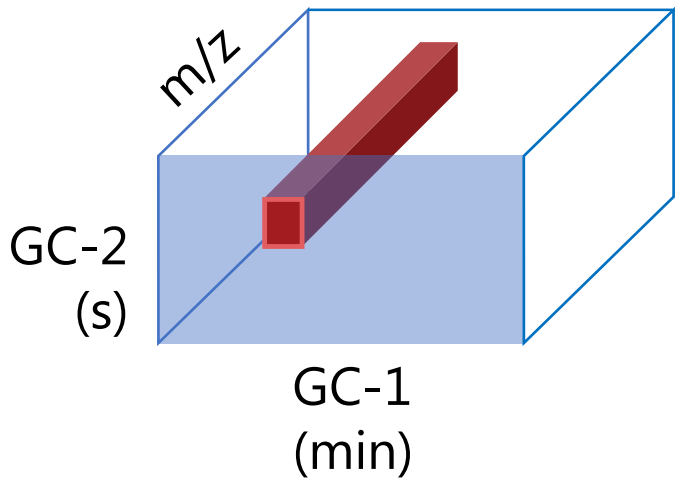
Figure: Dillen Augustijn, CAT, KU FOOD

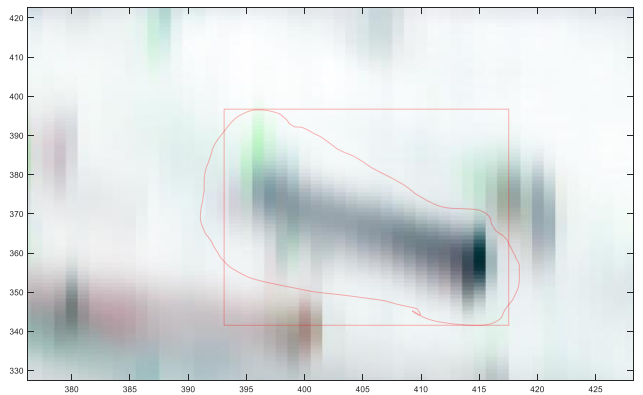
- Issues
  - Each sample is high dimensional and multi-modal (~1600 t_1^R x 800 t_2^R x 600 mz).
  - Possibly thousands of compounds.
  - Lots of nasty problems (sensor saturation, overlapping components, baseline, temporal correlation, shift between samples).
  - Manual identification is prohibitively expensive.

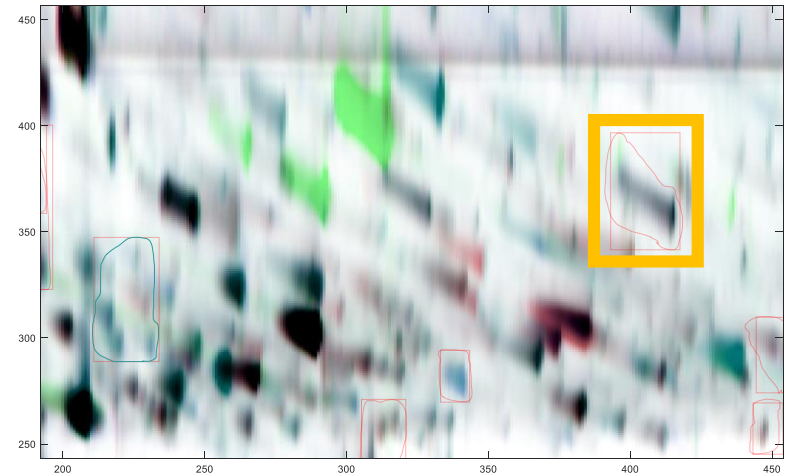
Identifying Chemical Compounds

- Each subarea contains one or more compounds
- Peak identification is non-trivial for most areas
- Areas overlap
- Larger areas → more compounds → harder to model (and computationally expensive)





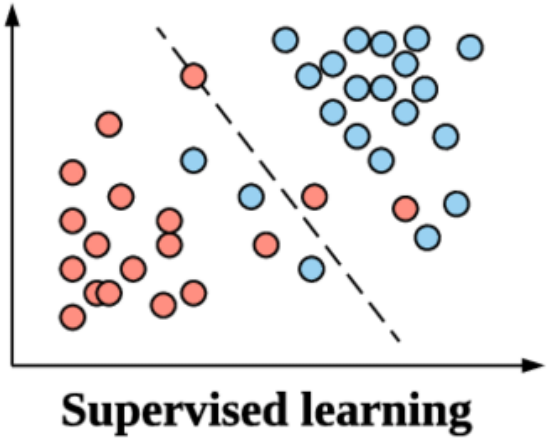




Images are slightly modified from originals by Dillen Augustijn, CAT, KU FOOD

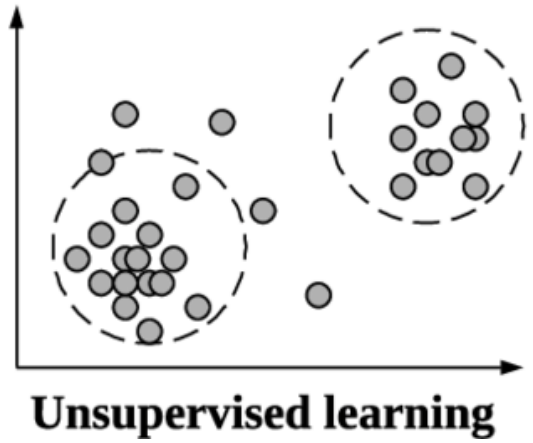
Statistical Science, Machine Learning, Data Science, etc.

- Goal: Learn the parameters $\boldsymbol{\theta}$ under model $f(*)$
  - Supervised: Learn from outcome Y and observed data X,
$$\boldsymbol{Y} = f(\boldsymbol{X}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$
  - Unsupervised: Learn from observed data **X**
$$\boldsymbol{X} = f(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$



- Find $\boldsymbol{\theta}$ by minimizing the error $\boldsymbol{\varepsilon}$.
  - Least squares, Maximum-likelihood

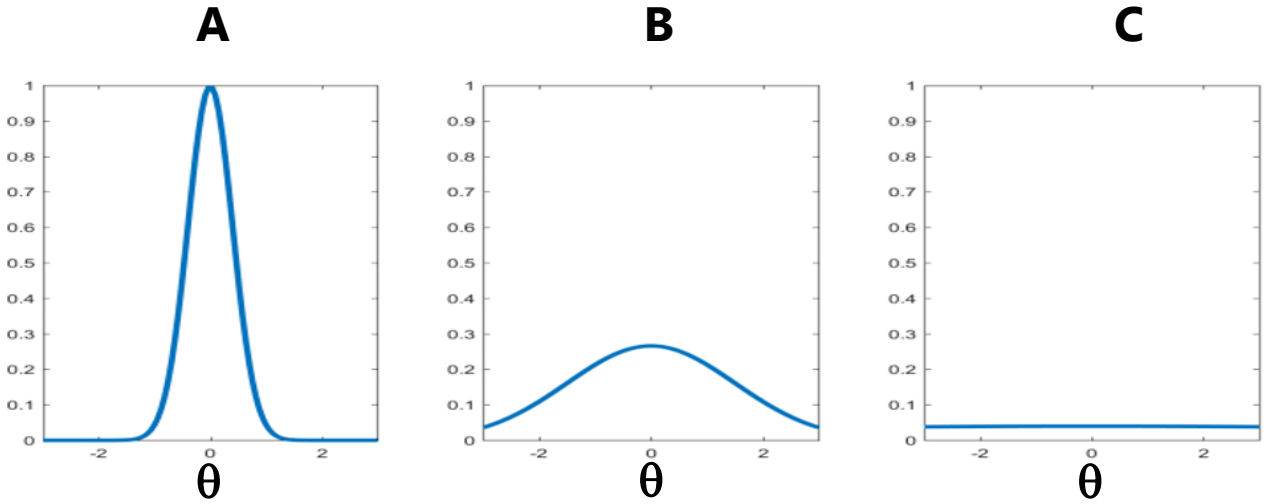
- Alternatively: Characterize $\boldsymbol{\theta}$ via the probability of obtaining the parameters.
$$\boldsymbol{P}(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}, \dots)$$



Parameter Estimation

- Observed data **X**, unknown model parameters θ
- Bayesian Inference
  - Likelihood, P(**X**|θ)
  - Prior knowledge, P(θ)
  - Posterior distribution, P(θ|**X**) = P(**X**|θ) p(θ) / p(**X**)
  - Probability of observing the data, P(**X**) = ∫ P(**X**|θ) P(θ) dθ
- Estimating θ based on
  - Maximum likelihood (ML)
  - Maximum a posteriori (MAP)
  - Full posterior distribution (Bayesian)

Example: A single univariate parameter θ centred at 0. How certain is this estimate?



- Benefits of Bayesian inference
  - Principled way of incorporating prior information
  - Characterize uncertainty via the posterior distribution

Why Bayesian Statistics?

Benefits

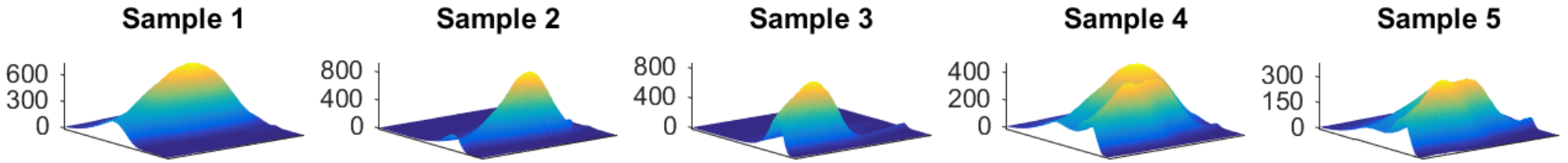
- Inherent penalization of model complexity
  - Robustness to noise and outliers
  - Guards against model over-specification
  - Automatic relevance determination
- Posterior distribution and uncertainty quantification
- A principled way to include prior knowledge
- Prediction on held-out slices or sub-tensors

Issues/Challenges

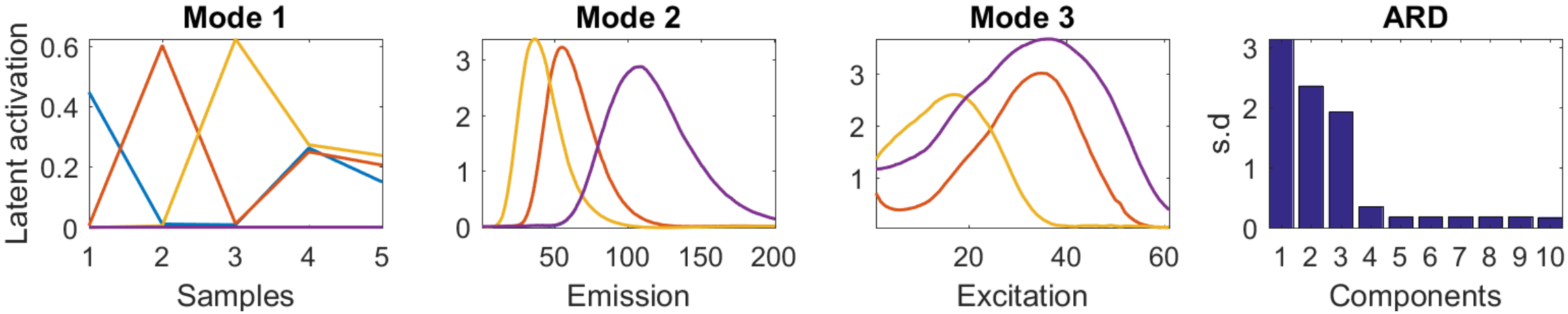
- The posterior distribution is intractable
  - Approximation vs. reliability of posterior estimate
  - Interpretability vs. inference methods
- Model specification and sensitivity to choice of prior
- Conjugate vs. non-conjugate inference
  - Restricts likelihood and prior choices

Automatic Relevance Determination: Fluorescence Spectroscopy

- Mixed samples

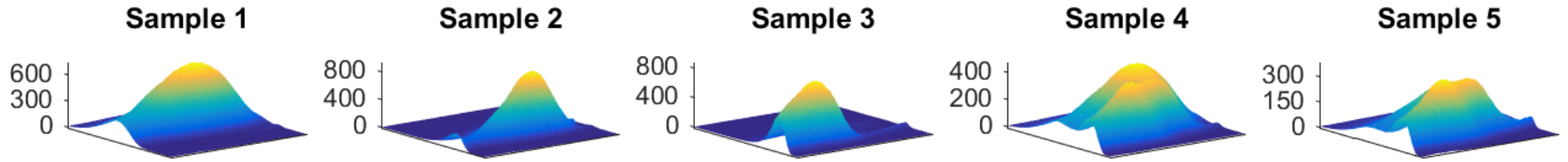


- Unmixing pure spectra and concentration (based on 10 initial components)

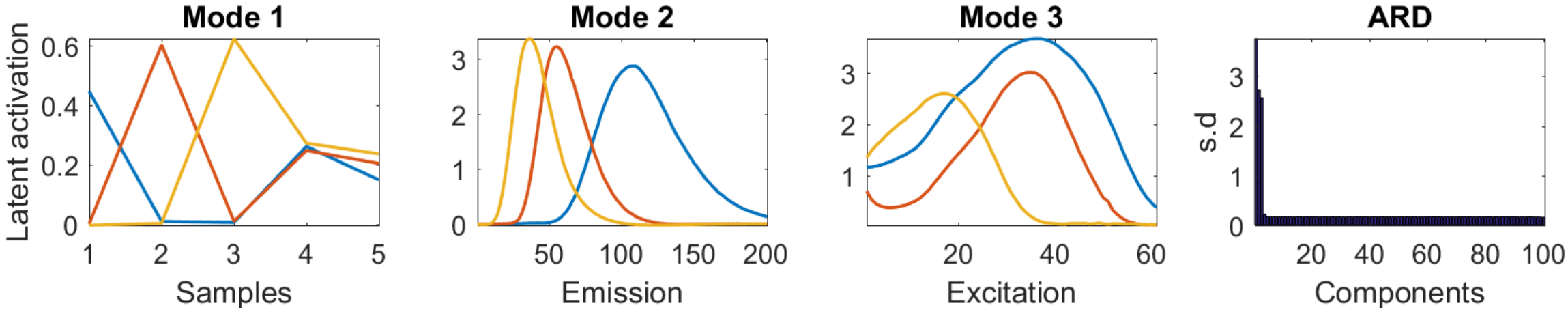


Automatic Relevance Determination: Fluorescence Spectroscopy

- Mixed samples

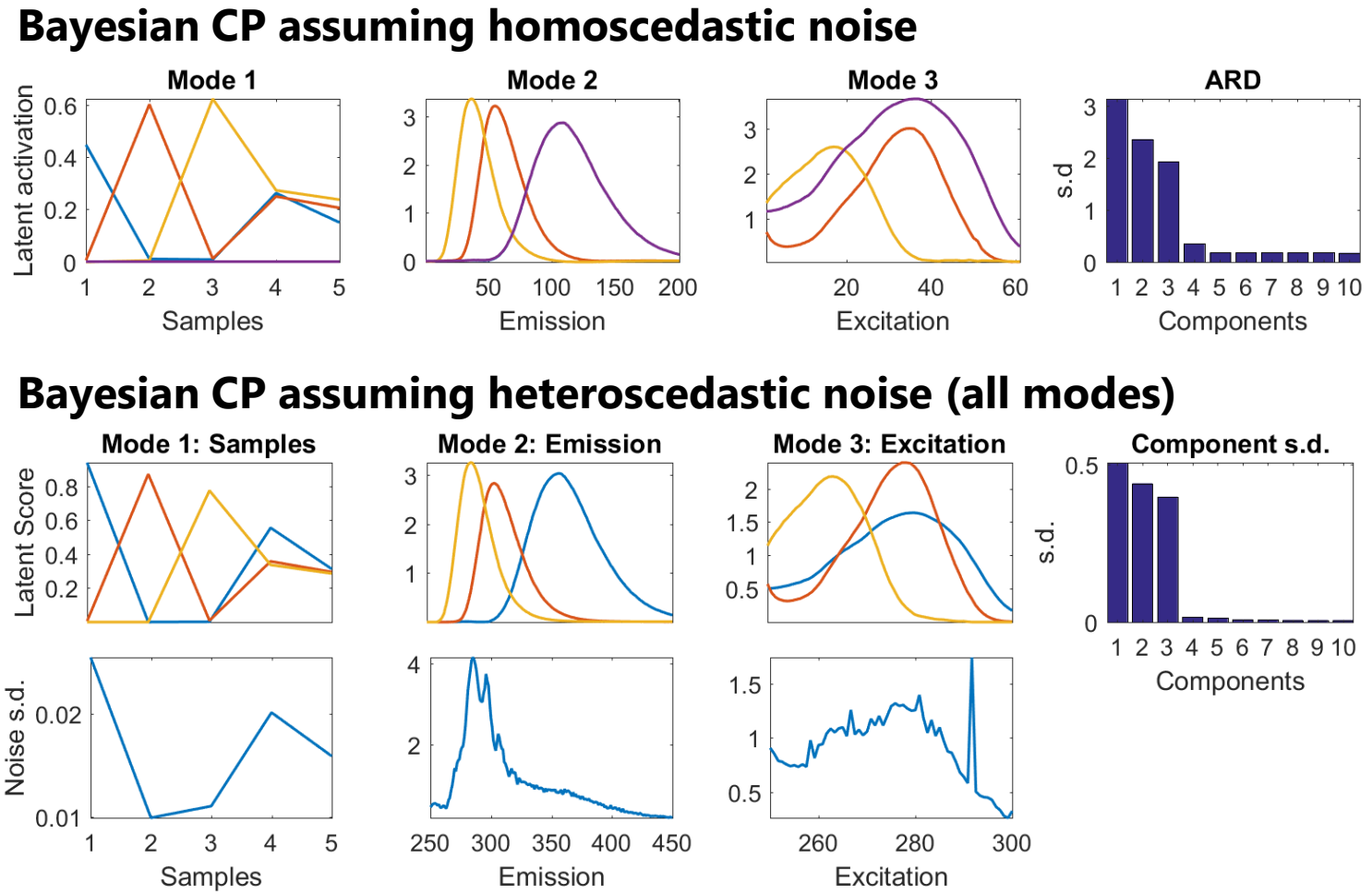


- Unmixing pure spectra and concentration (based on 100 initial components)



Homo- and Heteroscedastic Noise

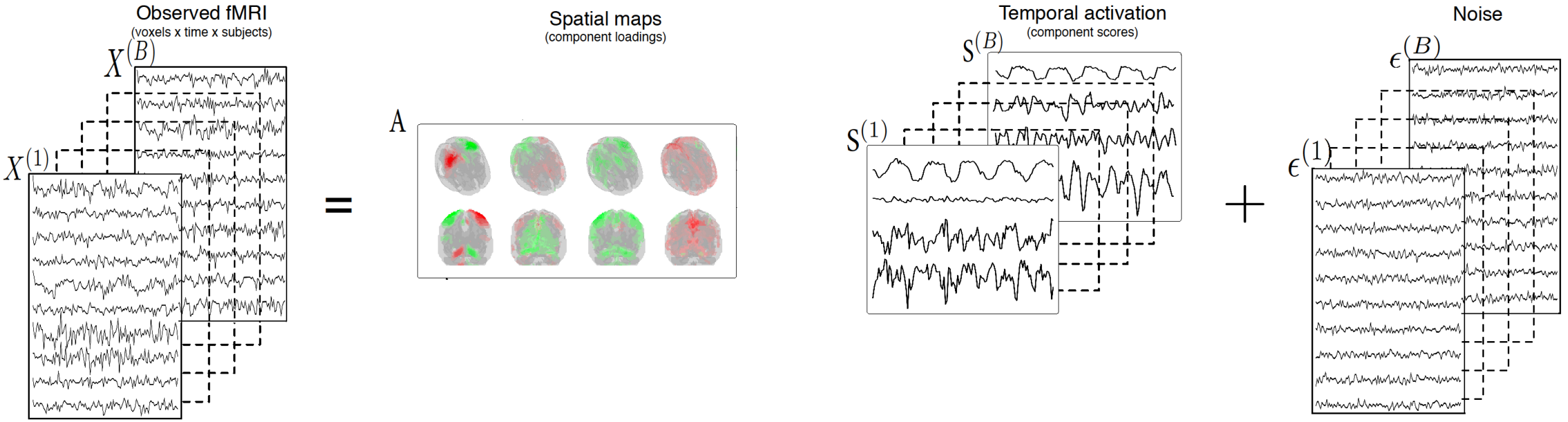
- 3-way fluorescence spectroscopy data
- Noise variance has scale indeterminacy



ARD and Het. Noise Example: Functional MRI

- Functional Magnetic Resonance Imaging
  - B=30 persons/samples, ~60000 voxels, ~300 timepoints
  - Models sample and voxel specific noise ($\boldsymbol{\tau}^{(b)}$) and sparse spatial maps (**A**)

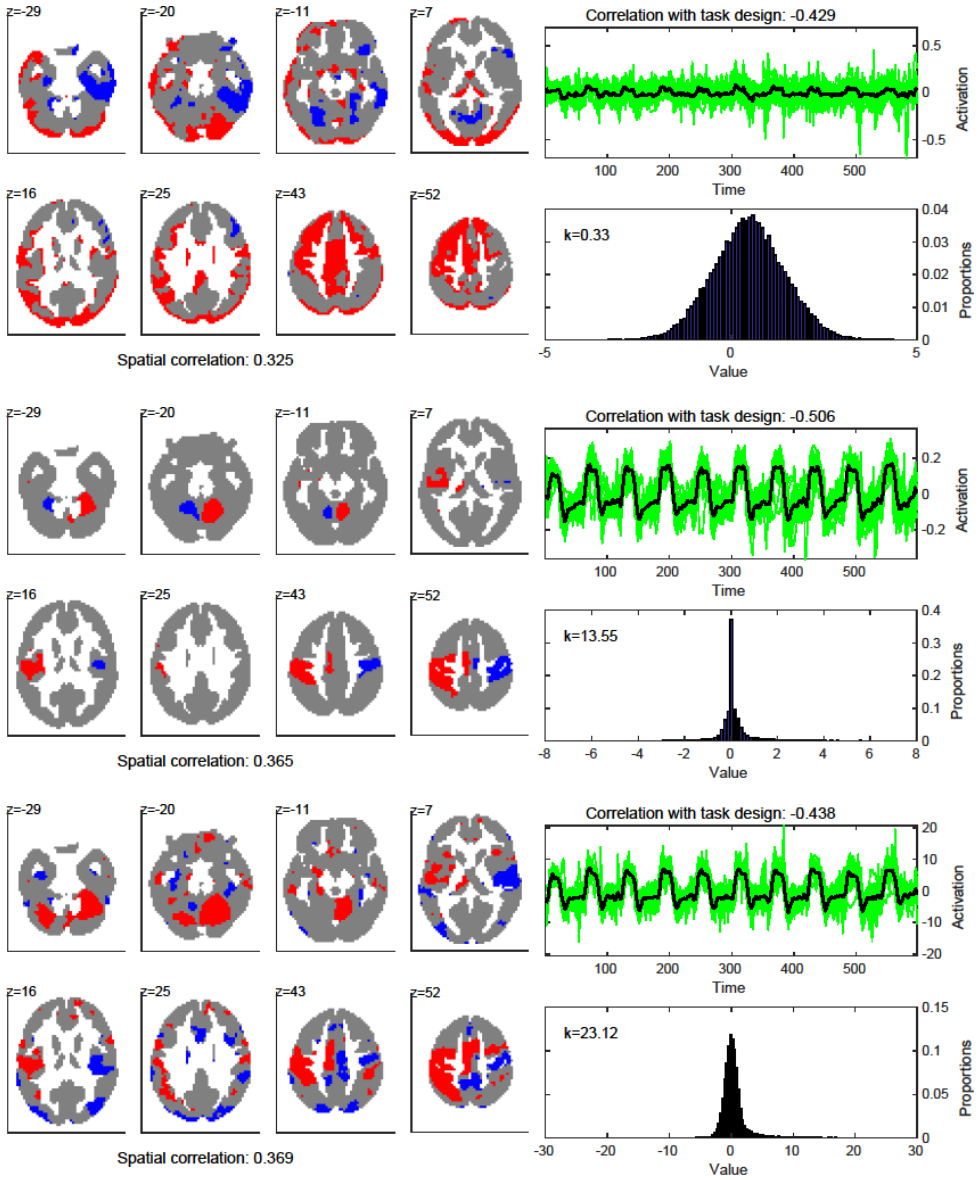
$$\mathcal{L}(\mathbf{X}|\boldsymbol{\theta}) = \prod_{b=1}^{B} \prod_{t=1}^{T^{(b)}} \mathcal{N}\left(\mathbf{x}_t^{(b)} | \mathbf{A}\mathbf{s}_t^{(b)} + \boldsymbol{\mu}^{(b)}, \operatorname{diag}\left(\boldsymbol{\tau}^{(b)}\right)^{-1}\right)$$



Hinrich, J. L., Nielsen, S. F., Riis, N. A., Eriksen, C. T., Frøsig, J., Kristensen, M. D., ... & Mørup, M. (2017, March). Scalable group level probabilistic sparse factor analysis. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on (pp. 6314-6318). IEEE.

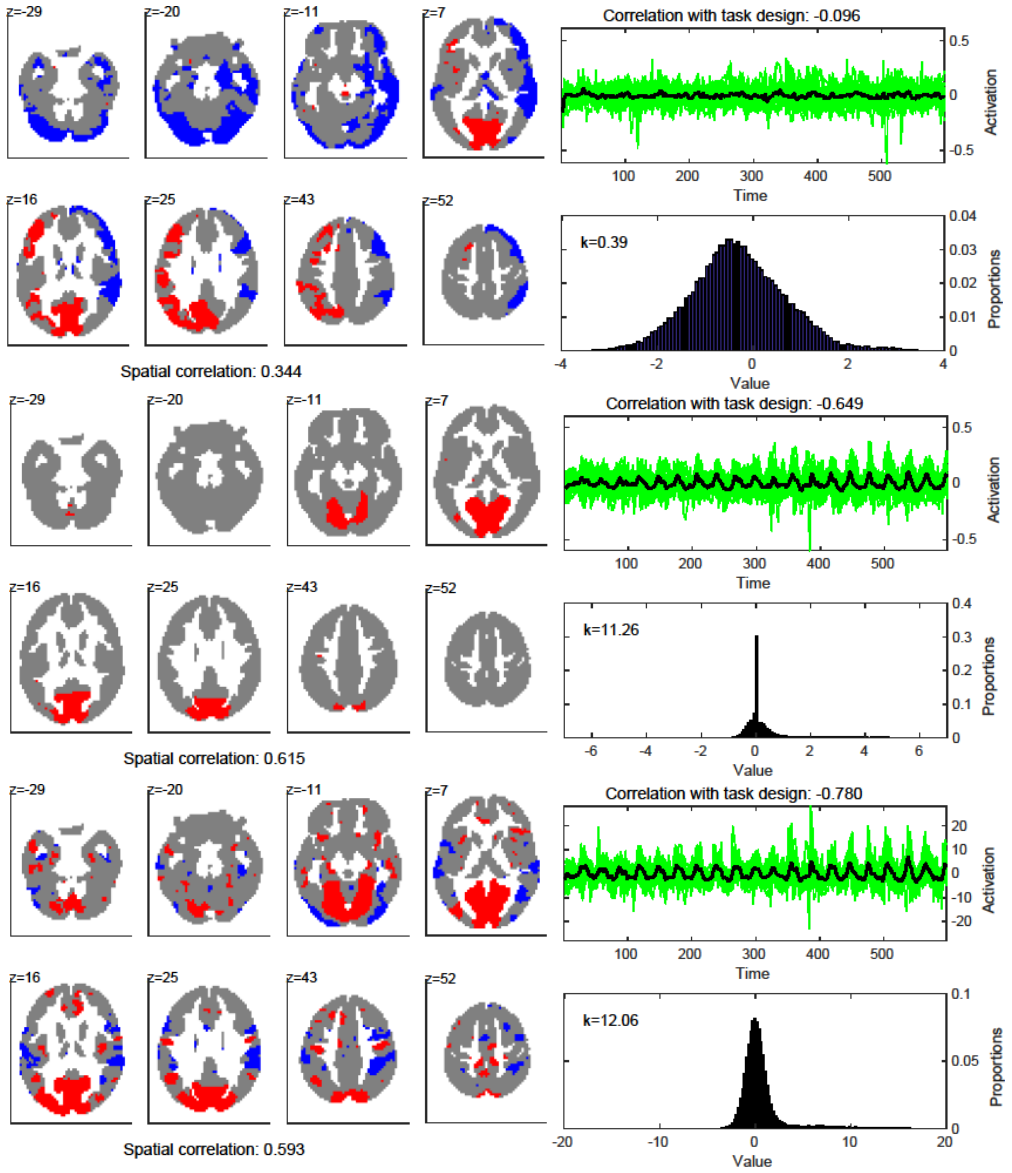
fMRI: Task Activated Component (Motor Cortex Related)

- Bayesian Factor Analysis
  - Hetero. noise
- Bayesian Sparse Factor Analysis
  - Hetero. noise and sparsity
- Independent Component Analysis
  - Group-ICA, **A** consists of independent signals.



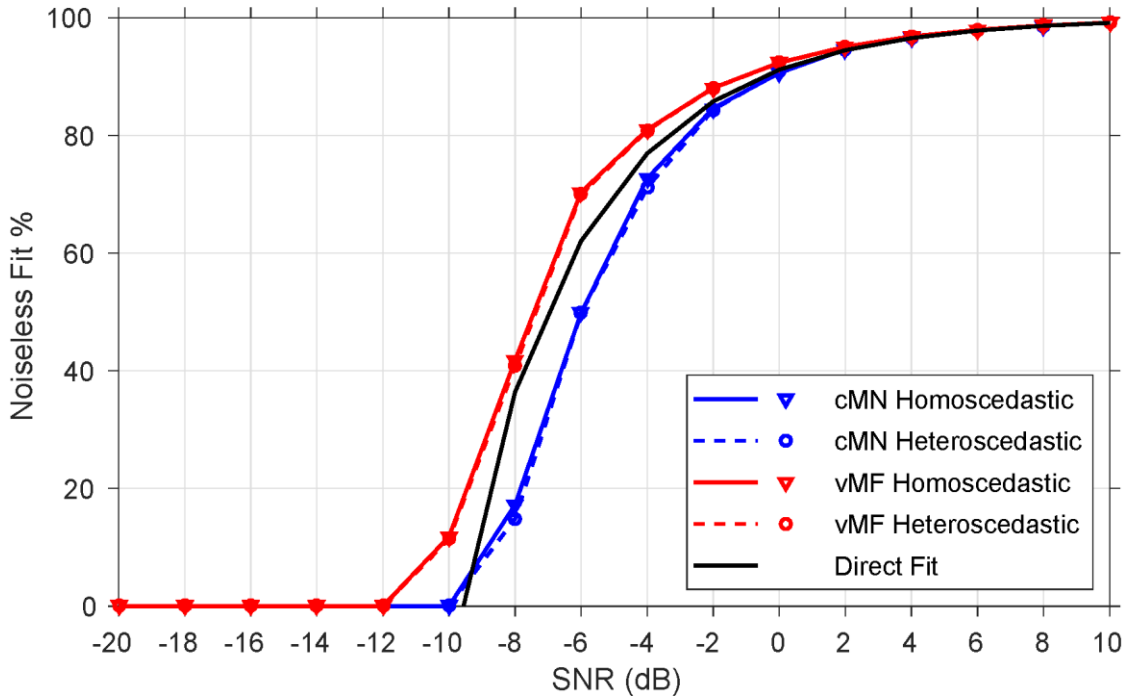
fMRI: Blinking Light (Visual Cortex Related)

- Bayesian Factor Analysis
  - Hetero. noise
- Bayesian Sparse Factor Analysis
  - Hetero. noise and sparsity
- Independent Component Analysis
  - Group-ICA, **A** consists of independent signals.



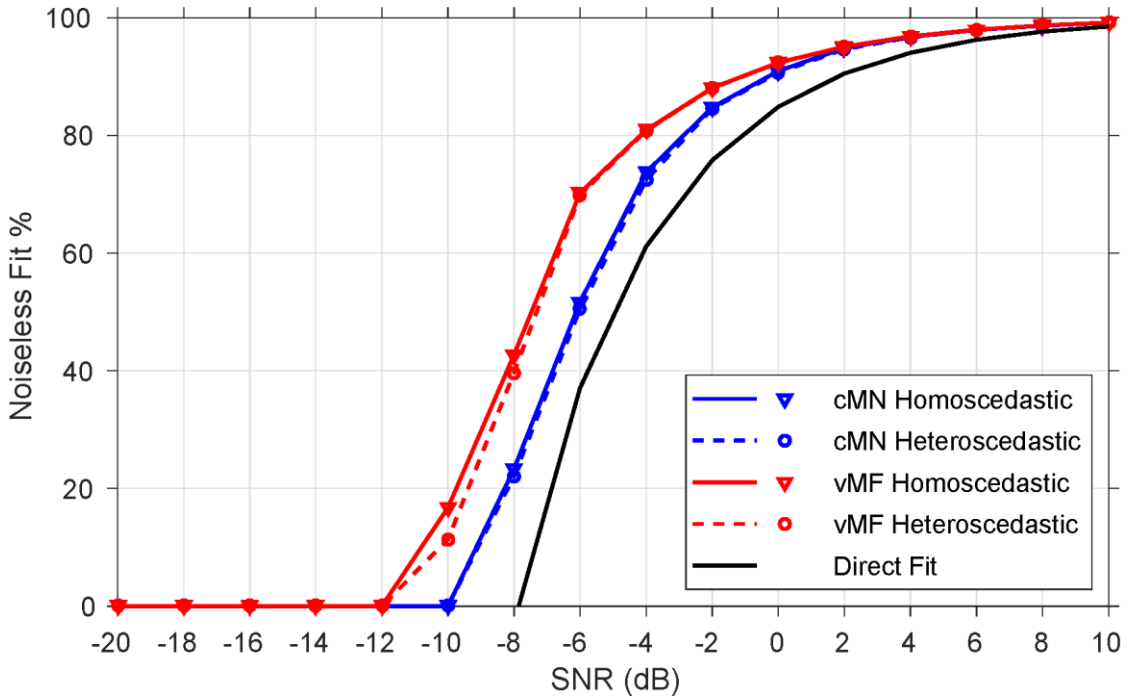
Probabilistic PARAFAC2

- Two Bayesian approaches constrasted to direct fit (MLE).
- Under **homoscedastic** normal noise.



Correct model order (D=D*=4)

- Orthogonality through, either
  - constrained Matrix Normal (cMN)
  - von Mises-Fisher Matrix (vMF)



Incorrect model order (D=6)

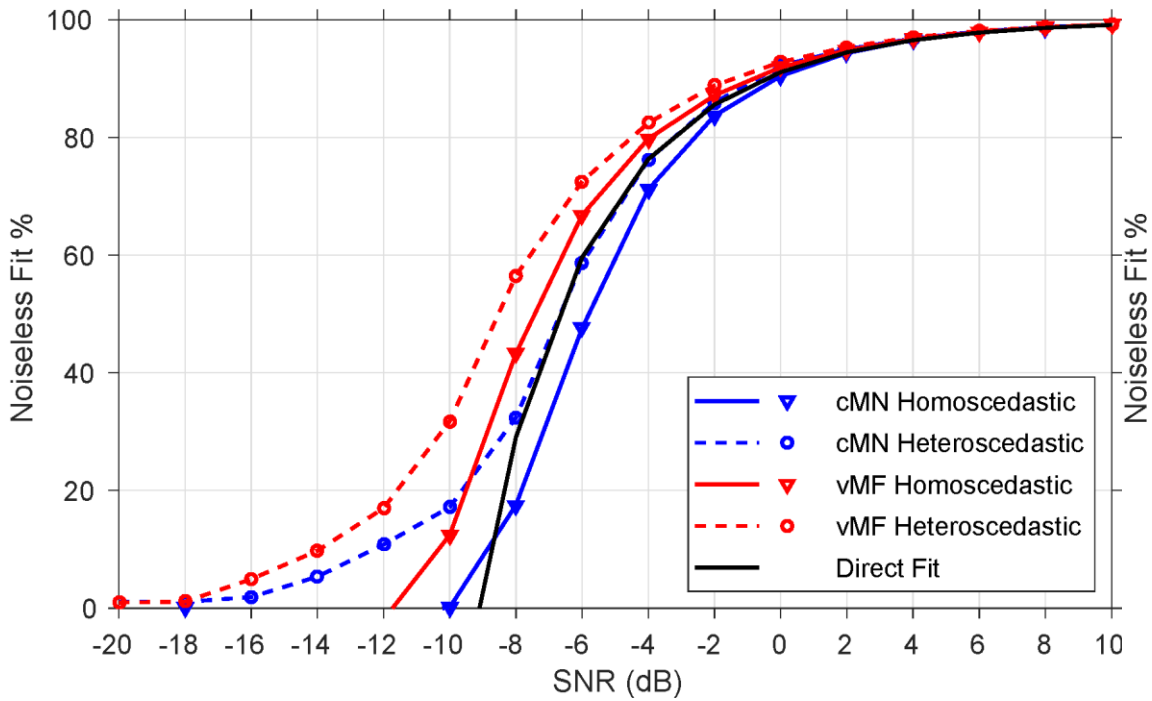
[1] Jørgensen, P. et al. "Probabilistic Parafac2" (Arxiv)

[2] Jørgensen, P. et al. (2019). Analysis of Chromatographic Data using the Probabilistic PARAFAC2. In *33rd Conference on Neural Information Processing Systems*.

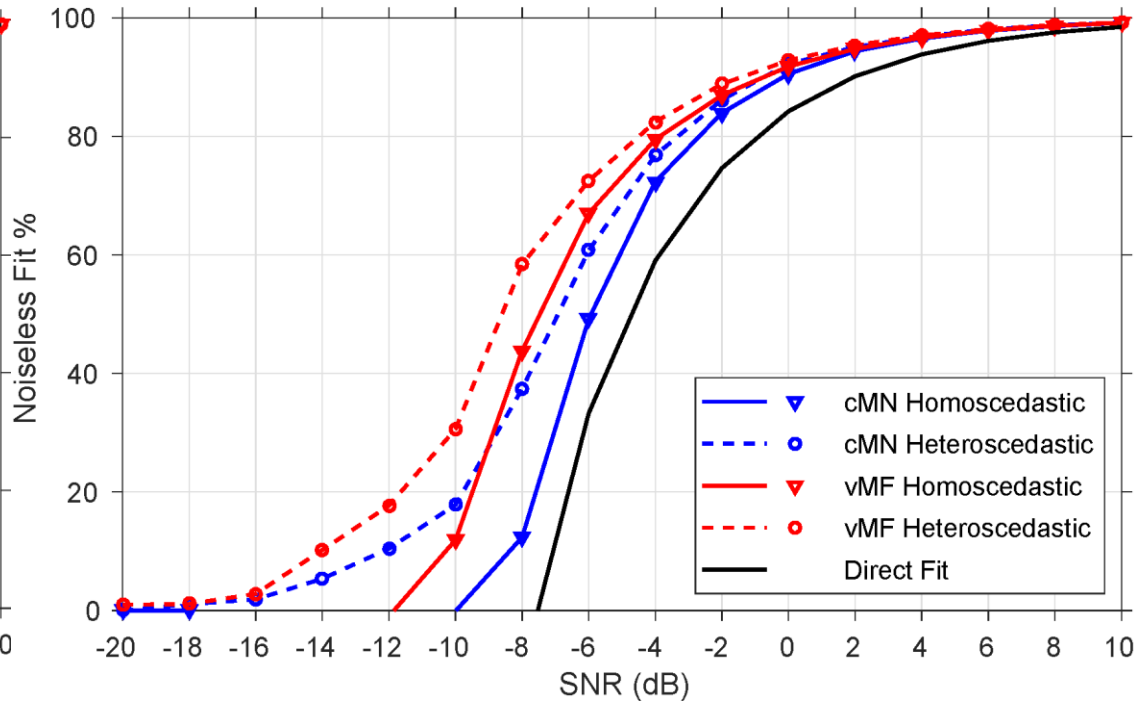
[3] Kiers, H. A., Ten Berge, J. M., & Bro, R. (1999). PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics, 13(3-4), 275-294.

Probabilistic PARAFAC2

- Two Bayesian approaches constrasted to direct fit (MLE).
- Under **heteroscedastic** normal noise.



Correct model order (D=D*=4)



Incorrect model order (D=6)

[1] Jørgensen, P. et al. "Probabilistic Parafac2" (Arxiv)

[2] Jørgensen, P. et. al. (2019). Analysis of Chromatographic Data using the Probabilistic PARAFAC2. In *33rd Conference on Neural Information Processing Systems*.

[3] Kiers, H. A., Ten Berge, J. M., & Bro, R. (1999). PARAFAC2-Part I. A direct fitting algorithm for the PARAFAC2 model. Journal of Chemometrics, 13(3-4), 275-294.

Takeaways and Future Work

Takeaways

- A principled way to include prior knowledge
- Posterior distribution and uncertainty quantification
- New tool for model comparison
- Automatic penalization of model complexity
  - Robustness to noise, outliers, and model over-specification.
  - No two-factor degeneracy

Future work

- Incorporation of chemical knowledge
  - Likely (mass) spectras
  - Elution shape and retention time
  - Sensor saturation
  - Temporal/spatial correlation
- Statistical
  - Improved uncertainty estimation
  - Uniqueness, local optima and global solution
  - Learning causal structures.
  - Interpretability vs. model inference
  - Conjugate vs. non-conjugate inference