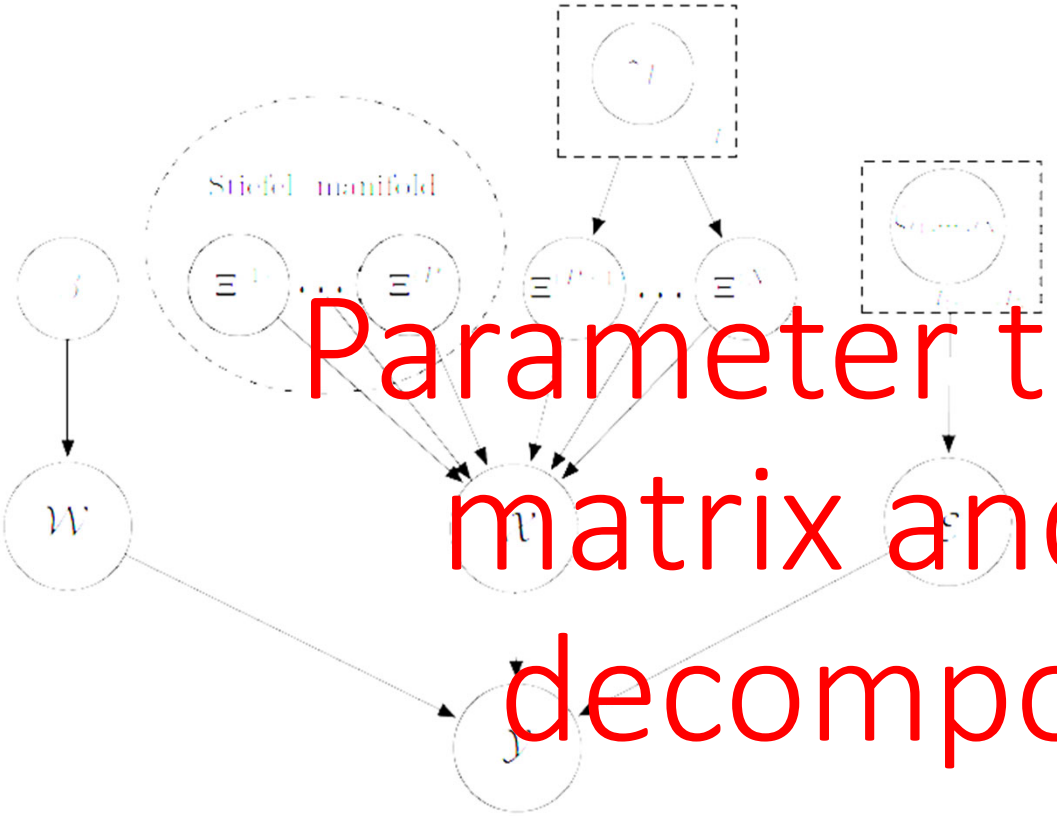
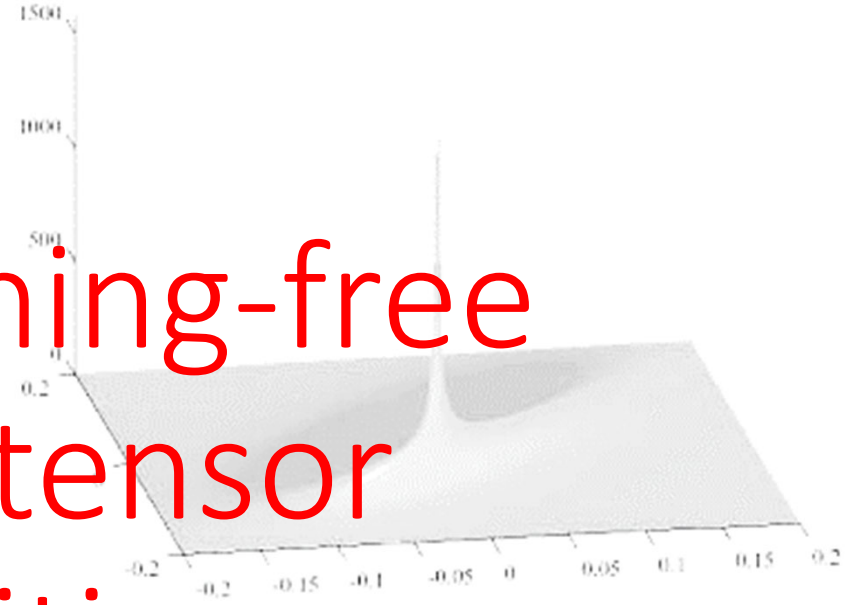
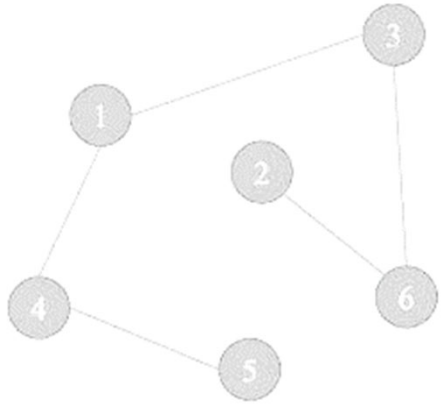




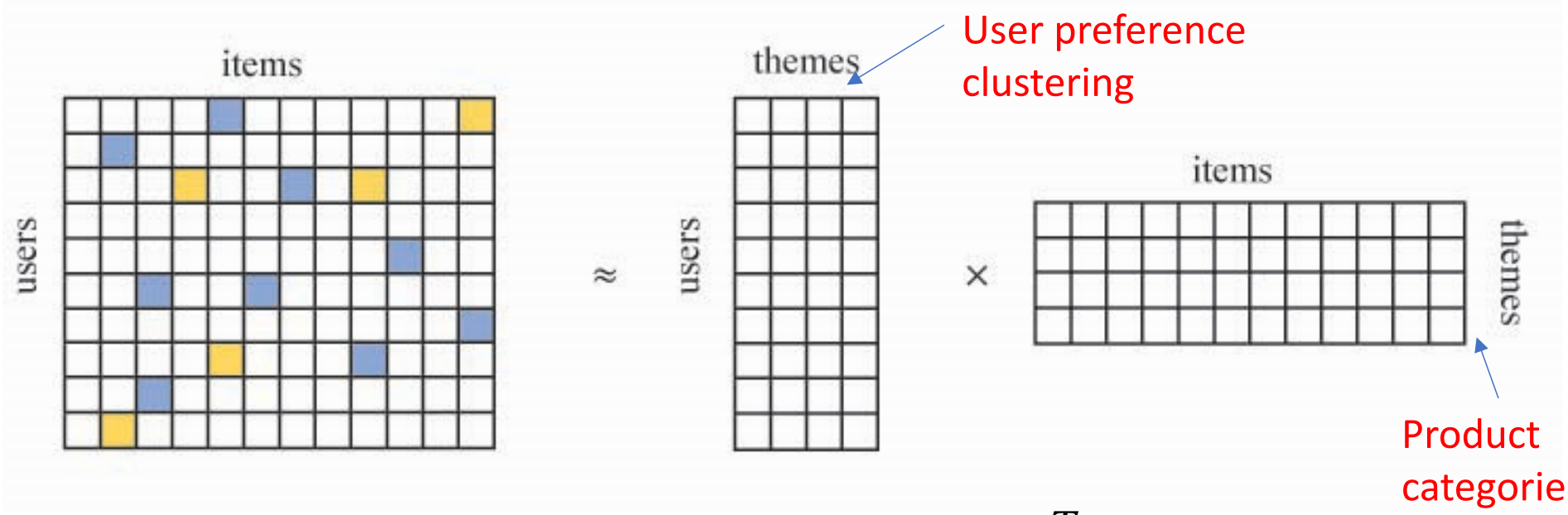
Parameter tuning-free matrix and tensor decompositions

Yik-Chung Wu
The University of Hong Kong
Email: ycwu@eee.hku.hk

(joint works with Lei Cheng, Yangge Chen, Xueke Tong, Le Xu, Zhongtao Chen)



Matrix completion problem



$$\mathbf{M} \quad = \quad \mathbf{U} \times \mathbf{V}^T$$

- Applications: product recommendation systems (e.g., Netflix, Amazon, …), Image completion, Genotype imputation

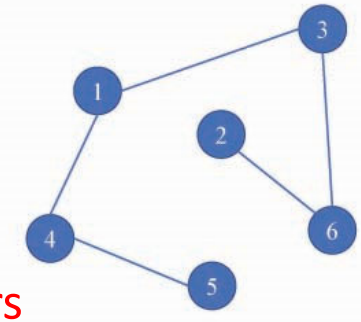
$$\min_{\mathbf{U},\mathbf{V}} \frac{\lambda_u}{2}\|\mathbf{U}\|_F^2 + \frac{\lambda_v}{2}\|\mathbf{V}\|_F^2 + \left\|\mathcal{P}_\Omega(\mathbf{M}-\mathbf{U}\mathbf{V}^T)\right\|_F^2$$

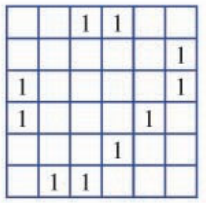
- Can be solved by alternating optimization
- The catch: need to determine appropriate setting for λ_u , λ_v , and number of columns in $\mathbf{U}$ and $\mathbf{V}$ (matrix rank)
- Current practice: trial-and-error
- Disadvantages:
  - Need to run the algorithm many times (e.g., each parameter has 10 options $\longrightarrow$ 1000 combinations)
  - Need to re-tune the parameters for different data or settings (missing ratio, signal-to-noise ratio, recommendation system vs image)
  - May not be able to tune as ground truth data may not be available in practice

Graph regularization

- Missing ratio is large in practical applications
- Domain knowledge / side information adopted to enhance the model learning
  - similarity among users,
  - similarity of movie within genres, same actors,
  - smooth transition in neighboring pixels
- E.g., In recommendation system, each row (user) of $\mathbf{M}$ can be represented as a node, link can be established based on similarity on age, gender, education level, ….

Graph information among users





Adjacency matrix $\mathbf{A}$

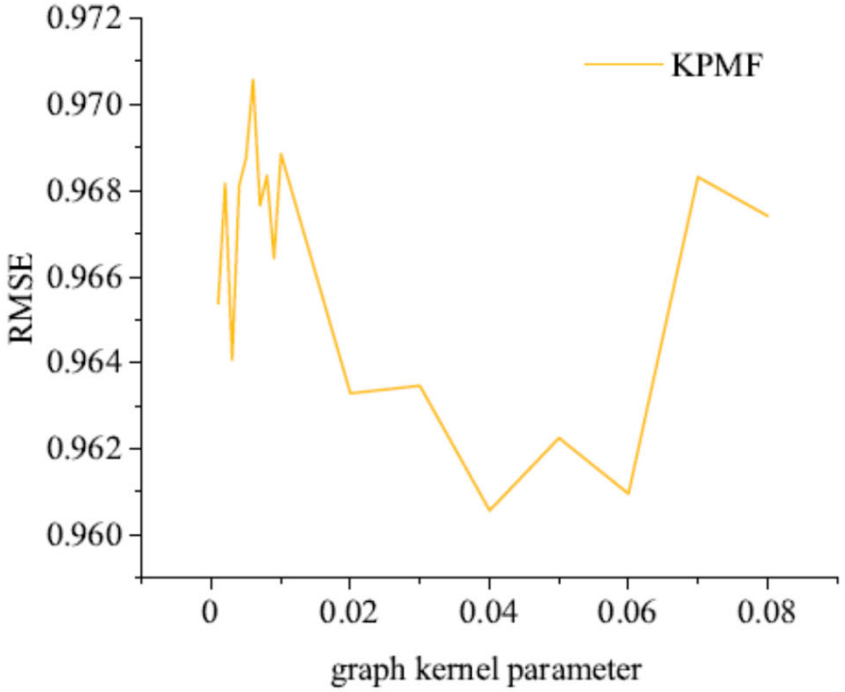
- In image completion, $\mathbf{A}_{ij}=\exp(-|i-j|^2/\theta^2)$
- Graph regularization term:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{A}_{ij}(\mathbf{x}_i-\mathbf{x}_j)^2=\mathrm{Tr}(\mathbf{X}^T\mathbf{L}_r\mathbf{X})$$

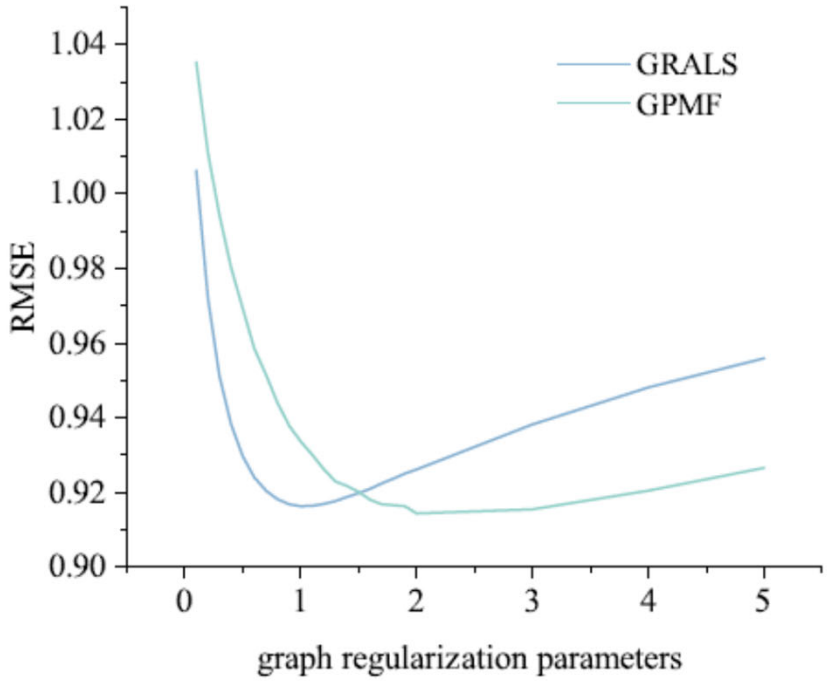
- graph Laplacian matrix: $\mathbf{L}_r=\mathrm{Diag}(\Sigma_j\mathbf{A}_{1j}, \Sigma_j\mathbf{A}_{2j}, \ldots, \Sigma_j\mathbf{A}_{nj}) - \mathbf{A}$
- Graph regularized matrix factorization

$$\min_{\mathbf{U},\mathbf{V}}\frac{\lambda_u}{2}\|\mathbf{U}\|_F^2+\frac{\lambda_v}{2}\|\mathbf{V}\|_F^2+\frac{\tau}{2}\left\|\mathcal{P}_\Omega(\mathbf{M}-\mathbf{U}\mathbf{V}^T)\right\|_F^2$$
$$+\left\{\mathrm{Tr}(\mathbf{U}^T\mathbf{L}_r\mathbf{U})+\mathrm{Tr}(\mathbf{V}^T\mathbf{L}_c\mathbf{V})\right\}$$

- Even more parameters to be tuned



(a) RMSE vs. graph kernel parameter



(b) RMSE vs. graph regularization parameters

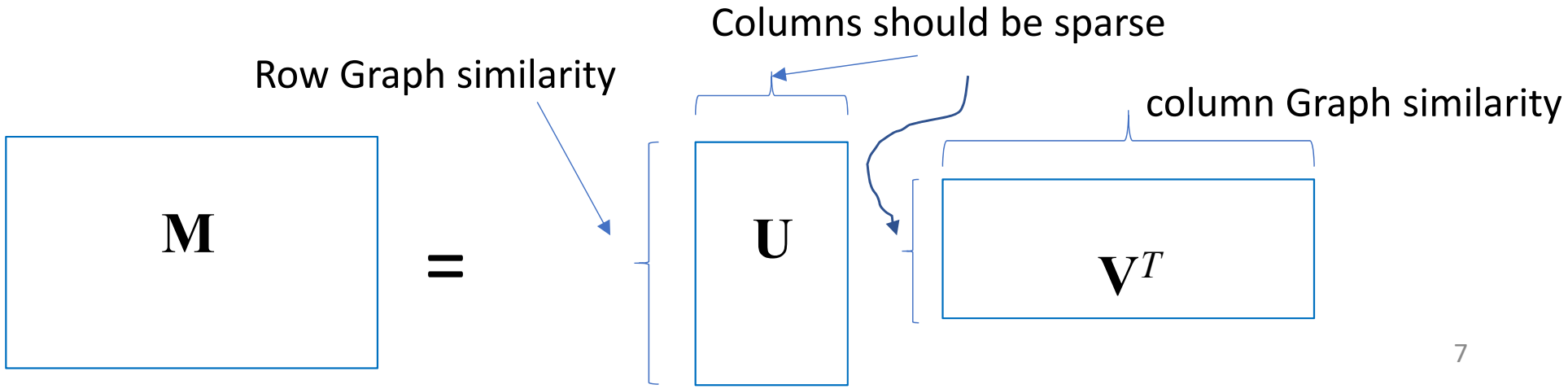
(c) graph regularization parameters = 0.05

(d) graph regularization parameters = 0.1

(e) graph regularization parameters = 0.5

Bayesian approach

- Treat the parameters to be estimated (including regularization parameters) as random variables
- Assign appropriate prior density functions to the random variables
- The observations constitute the likelihood function
- Compute the posterior distributions of different random variables using Bayes' rule



Control the variance of each column of $\mathbf{U}$

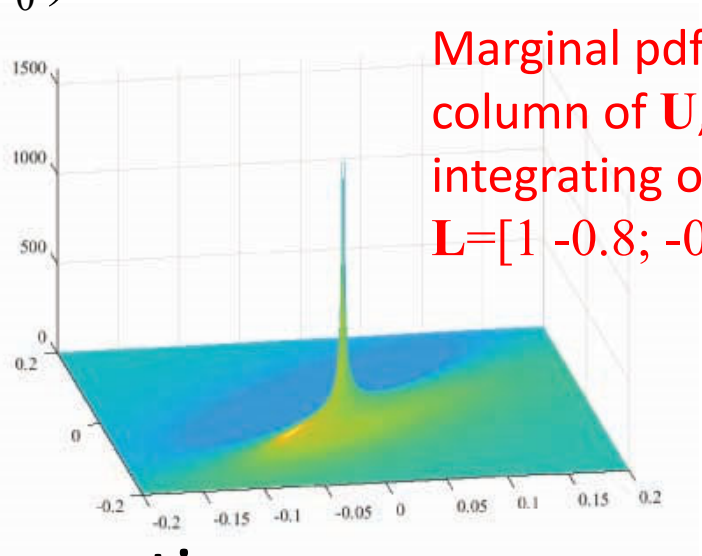
Graph regularization on rows of $\mathbf{U}$

$$p(\mathbf{U}\,|\,\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{U}\,|\,\mathbf{0}, \boldsymbol{\Lambda}^{-1}, \mathbf{L}_r^{-1})$$

$$p(\mathbf{V}\,|\,\boldsymbol{\lambda}) = \mathcal{N}(\mathbf{V}\,|\,\mathbf{0}, \boldsymbol{\Lambda}^{-1}, \mathbf{L}_c^{-1})$$

$$p(\boldsymbol{\lambda}) = \prod_{r=1}^{L} Ga(\lambda_r\,|\,c_0^r, d_0^r)$$

Precision being another random variable. c_0, d_0 are set 10^{-6} to induce non-informative prior



Marginal pdf of a column of $\mathbf{U}$, after integrating out λ. $\mathbf{L}$=[1 -0.8; -0.8, 1]

- The Gaussian-gamma hierarchy is frequently used to induce sparsity on the columns of a matrix
- Likelihood function from observations

$$p(\mathbf{M}\,|\,\mathbf{U}, \mathbf{V}, \tau) = \prod_{(i,j)\in\boldsymbol{\Omega}} \mathcal{N}(\mathbf{M}_{ij}\,|\,(\mathbf{U}\mathbf{V}^T)_{ij}, \tau^{-1})$$

$$p(\tau) = Ga(\tau\,|\,a_0, b_0)$$

- Joint distribution ($\boldsymbol{\Theta}=\{\mathbf{U},\mathbf{V},\boldsymbol{\lambda},\tau\}$):

$$p(\mathbf{M},\boldsymbol{\Theta}) = p(\mathbf{M}\,|\,\mathbf{U},\mathbf{V},\tau)\,p(\mathbf{U}\,|\,\boldsymbol{\lambda})\,p(\mathbf{V}\,|\,\boldsymbol{\lambda})\,p(\boldsymbol{\lambda})\,p(\tau)$$

- The posterior distribution $p(\boldsymbol{\Theta}\,|\,\mathbf{M}) = \dfrac{p(\mathbf{M},\boldsymbol{\Theta})}{\int p(\mathbf{M},\boldsymbol{\Theta})\,d\boldsymbol{\Theta}}$

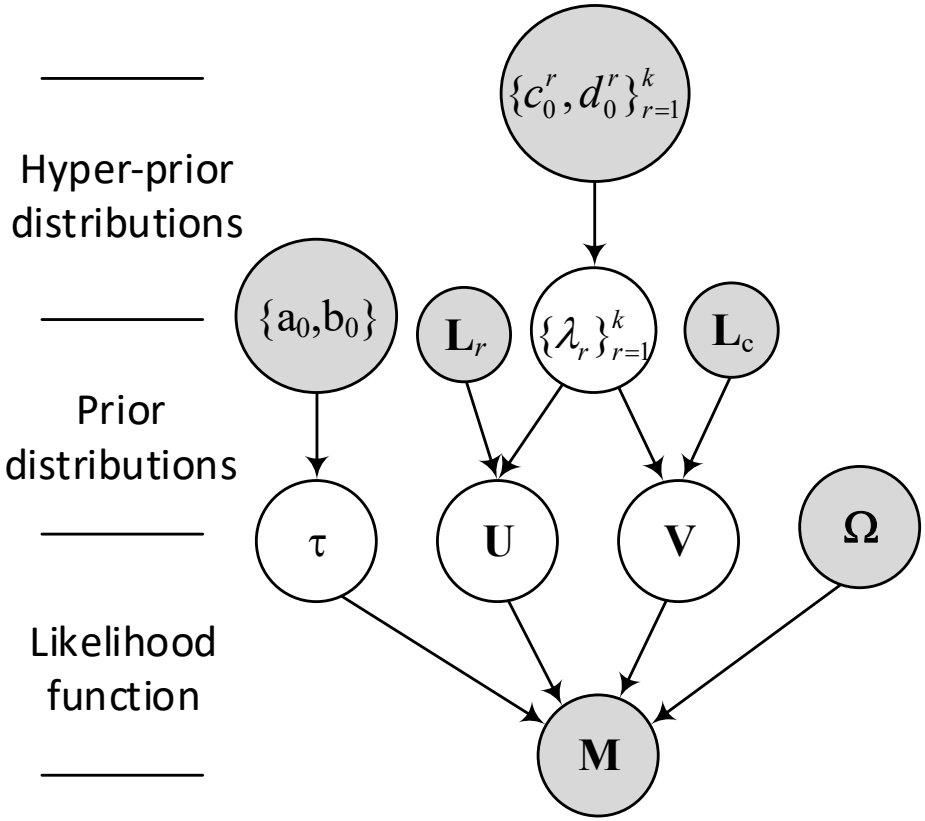
- However, the integration in the denominator is intractable ⇒ sampling methods or variational inference
- Variational inference: Seek $q(\boldsymbol{\Theta})$ that minimizes

$$\mathrm{KL}(q(\boldsymbol{\Theta})\,||\,p(\boldsymbol{\Theta}\,|\,\mathbf{M})) = \int q(\boldsymbol{\Theta})\ln\left(\frac{q(\boldsymbol{\Theta})}{p(\boldsymbol{\Theta}\,|\,\mathbf{M})}\right)d\boldsymbol{\Theta}$$

- Under mean-field assumption $q(\boldsymbol{\Theta}) = \Pi_j \, q(\boldsymbol{\Theta}_j)$

$$\ln q(\boldsymbol{\Theta}_j) = \mathbb{E}_{q(\boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_j)}[\ln p(\mathbf{M}, \boldsymbol{\Theta})] + const$$

- Iterative update the variational distributions
- Closed-form update expressions can be established if the upper level distribution is conjugate to the lower level distribution for a directly connected variable pair



- It turns out that conjugacy is satisfied if we view each column of $\mathbf{U}$ and $\mathbf{V}$ as a unit
- Take the mean-field

$$q(\Theta) = \prod_{i=1}^{L} q(\mathbf{u}_i) \prod_{i=1}^{L} q(\mathbf{v}_i) q(\boldsymbol{\lambda}) q(\tau)$$

- $q(\mathbf{u}_i)$, $q(\mathbf{v}_i)$ are Gaussian distributed vectors, $q(\boldsymbol{\lambda}_r)$, $q(\tau)$ are gamma distributions
- The algorithm looks like this:
- $q(\mathbf{u}_i) = \mathcal{N}(\mathbf{u}_i \mid \hat{\boldsymbol{\mu}}_i^u, \hat{\boldsymbol{\Sigma}}_i^u)$

$$\hat{\boldsymbol{\mu}}_i^u = < \tau > \hat{\boldsymbol{\Sigma}}_i^u (\boldsymbol{\Omega} \circ (\mathbf{M} - \sum_{r \neq i} < \mathbf{u}_r \mathbf{v}_r^T >)) < \mathbf{v}_i >$$

$$\hat{\boldsymbol{\Sigma}}_i^u = (< \tau > \mathrm{diag}(\boldsymbol{\Omega} < \mathbf{v}_i \circ \mathbf{v}_i >) + < \lambda_i > \mathbf{L}_r)^{-1}$$

- $q(\mathbf{v}_i) = \mathcal{N}(\mathbf{v}_i \mid \hat{\boldsymbol{\mu}}_i^v, \hat{\boldsymbol{\Sigma}}_i^v)$

$$\hat{\boldsymbol{\mu}}_i^v = <\tau> \hat{\boldsymbol{\Sigma}}_i^v (\boldsymbol{\Omega} \circ (\mathbf{M} - \sum\nolimits_{r \neq i} <\mathbf{u}_r \mathbf{v}_r^T>))^T <\mathbf{u}_i>$$

$$\hat{\boldsymbol{\Sigma}}_i^v = (<\tau> \mathrm{diag}(\boldsymbol{\Omega}^T <\mathbf{u}_i \circ \mathbf{u}_i>) + <\lambda_i> \mathbf{L}_c)^{-1}$$

- $q(\boldsymbol{\lambda}) = \prod_{r=1}^{L} Ga(\lambda_r \mid \hat{c}^r, \hat{d}^r)$

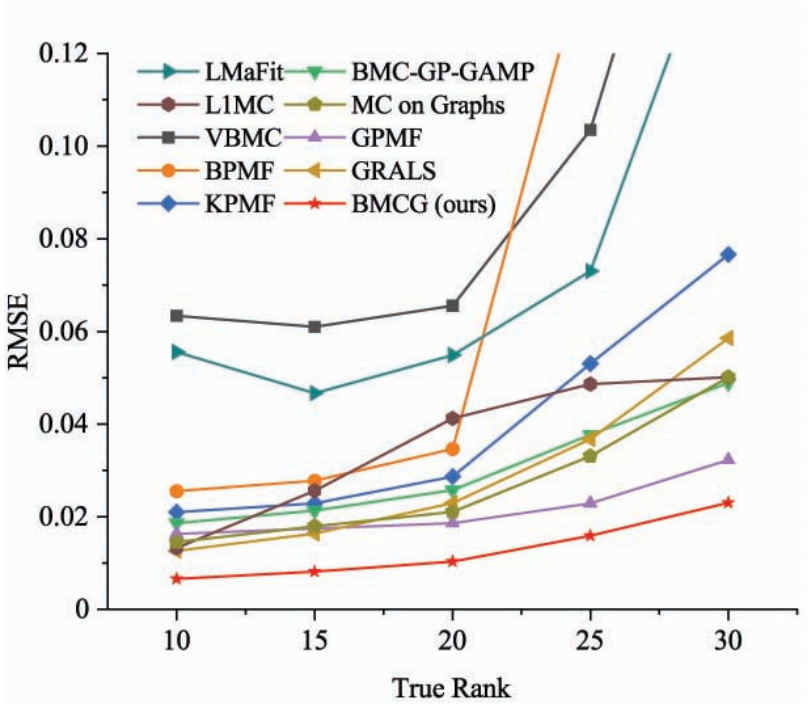
$$\hat{c}^r = c_0^r + (m+n)/2, \quad \hat{d}^r = d_0^r + (<\mathbf{u}_r^T \mathbf{L}_r \mathbf{u}_r> + <\mathbf{v}_r^T \mathbf{L}_r \mathbf{v}_r>)/2$$

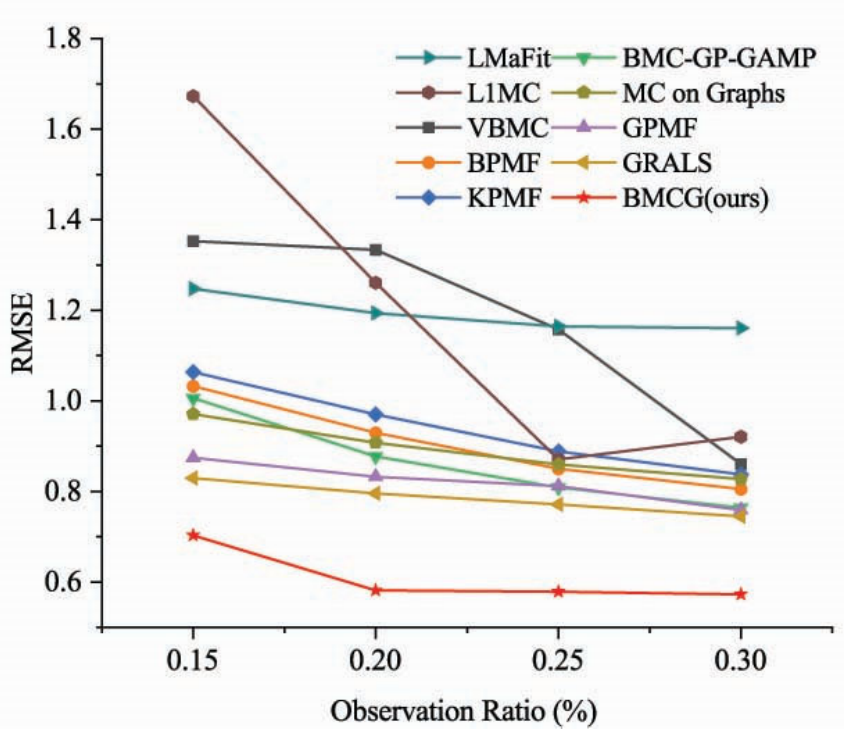
- $q(\tau) = Ga(\tau \mid \hat{a}, \hat{b})$

$$\hat{a} = |\boldsymbol{\Omega}|/2 + a_0, \quad \hat{b} = <||\boldsymbol{\Omega} \circ (\mathbf{M} - \mathbf{U}\mathbf{V}^T)||_F^2>/2 + b_0$$

- The estimates of various parameters are taken as the means of the corresponding $q(.)$ after convergence

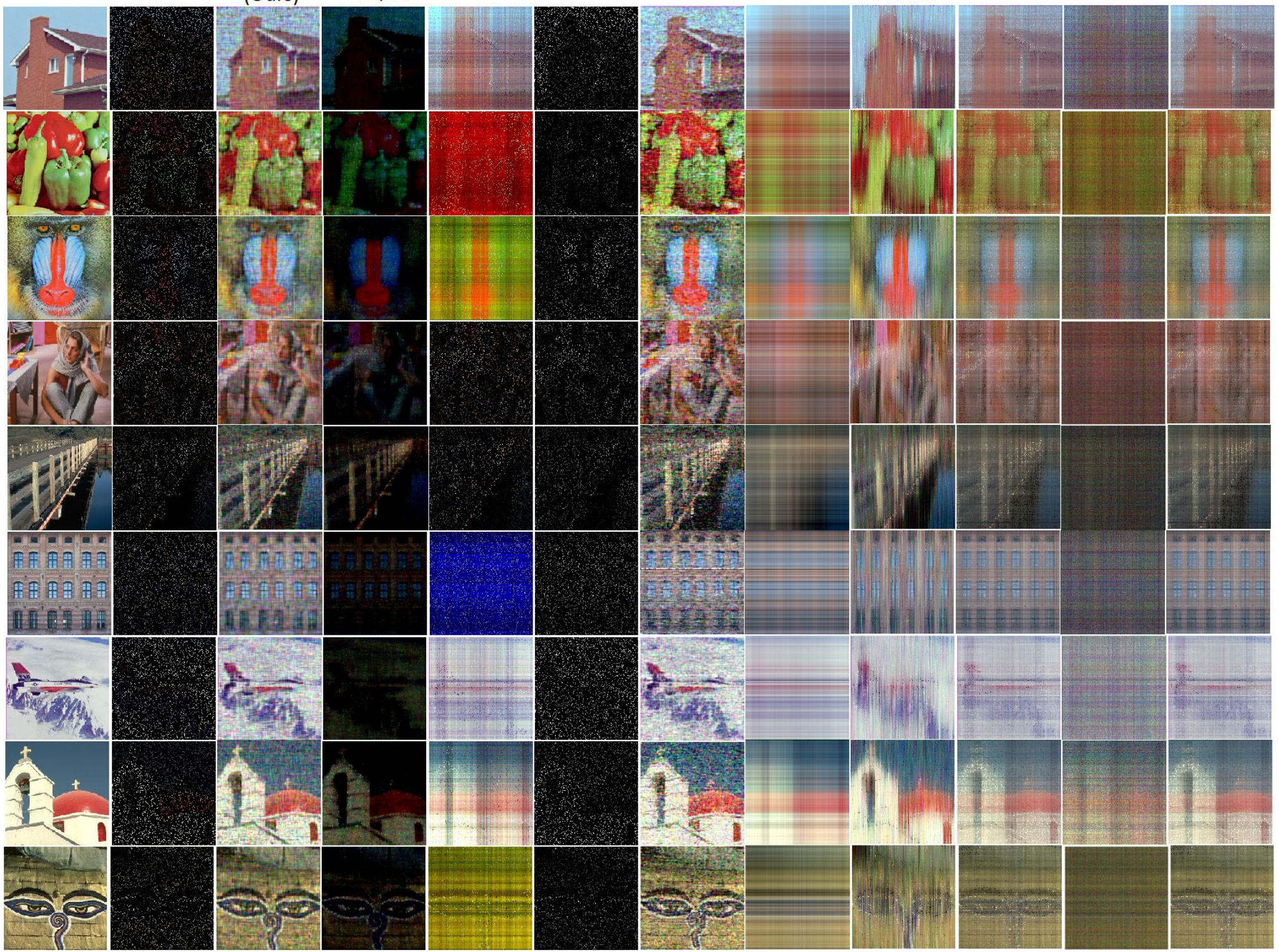
- Physical meaning of learned results :
  - $\|\mathbf{u}_r\|^2+\|\mathbf{v}_r\|^2$ represents the energy of the r^{th} component $\longrightarrow$ used to select the rank
  - $\mathbb{E}[\boldsymbol{\lambda}_r]/\mathbb{E}[\tau]$ represents inverse of "signal-to-noise ratio" of the r^{th} component $\longrightarrow$ regularization is performed component-wise rather than on the whole $\mathbf{U}$ or $\mathbf{V}$





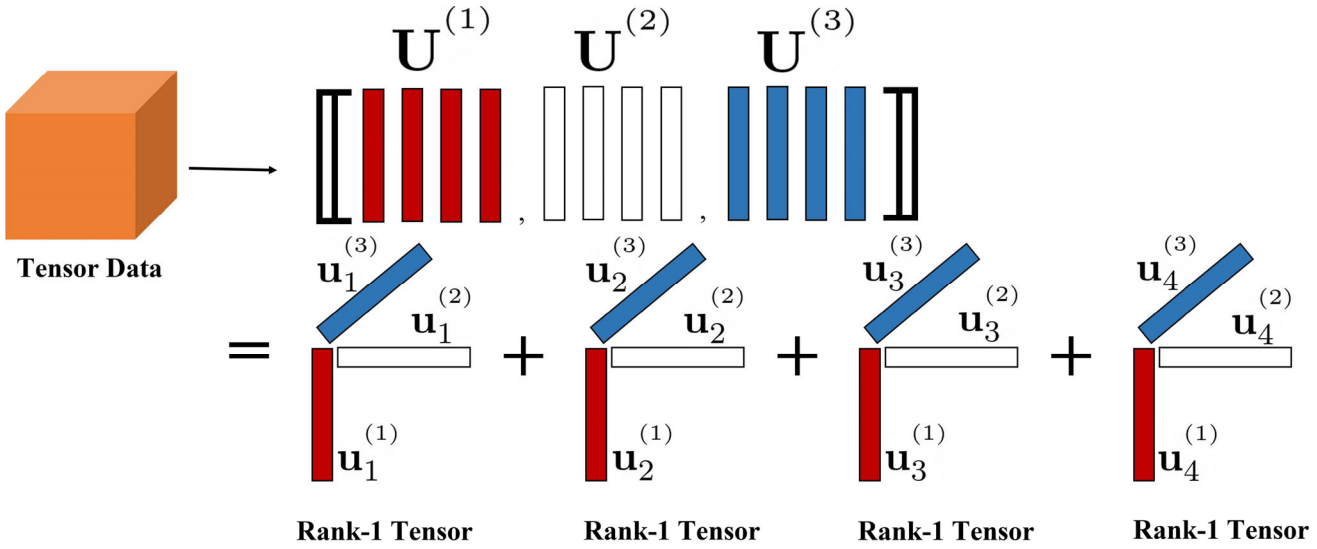
500×500 graph-structured synthetic matrix, observation ratio = 0.2, SNR = 10 dB

Netflix-like rating data matrix 100×200


Original
Observed
BMCG (Ours)
MC on Graphs
L1MC
LMaFit
GRALS
GPMF
BMC-GP-GAMP
KPMF
BPMF
VBMC

From Matrix to Tensor

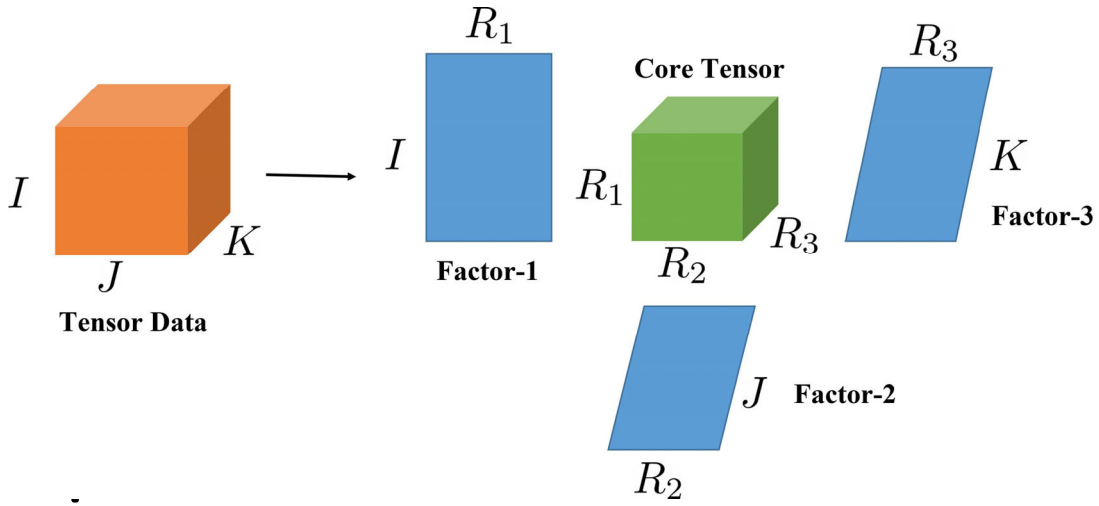
- Tensor are three or more dimensional data
- There are many tensor decomposition formats
- Canonical Polyadic Decomposition (CPD):



Number of components is the CPD rank

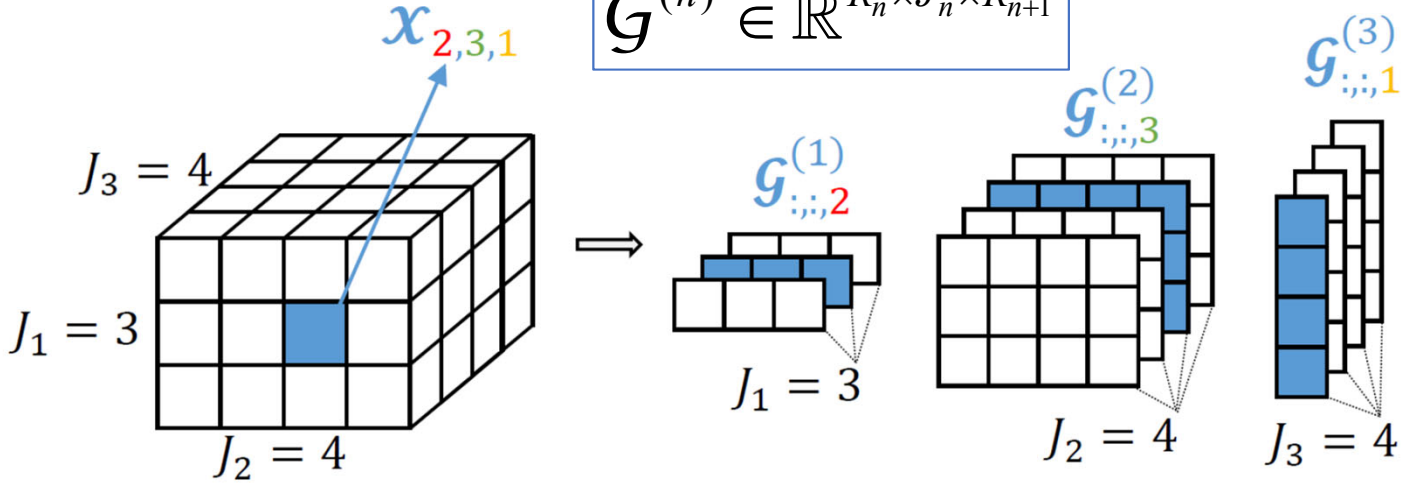
- Tucker decomposition:

(R_1, R_2, R_3) are called multilinear rank



- Tensor train:

$$\mathcal{G}^{(n)} \in \mathbb{R}^{R_n \times J_n \times R_{n+1}}$$



- Other formats: Tensor ring, T-SVD

CPD

- Optimization-based formulation:

$$\min_{\{\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R}\}_{n=1}^{N}, R} \left\| \mathcal{Y} - [\![\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}]\!] \right\|_F^2$$

- Standard algorithm: ALS

- Bayesian formulation (sparsity in columns of $\mathbf{U}^{(n)}$):

$$p(\{\mathbf{U}^{(n)}\}_{n=1}^{N} \mid \{\gamma_l\}_{l=1}^{L}) = \prod_{l=1}^{L} p(\{\mathbf{U}_{:,l}^{(n)}\}_{n=1}^{N} \mid \gamma_l) = \prod_{l=1}^{L} \prod_{n=1}^{N} \mathcal{N}(\mathbf{U}_{:,l}^{(n)} \mid \mathbf{0}_{I_n \times 1}, \gamma_l^{-1} \mathbf{I}_{I_n})$$

Each rank-one component is independent with a different precision γ_l

$$p(\{\gamma_l\}_{l=1}^{L} \mid \{c_l^0, d_l^0\}_{l=1}^{L}) = \prod_{l=1}^{L} p(\gamma_l \mid c_l^0, d_l^0) = \prod_{l=1}^{L} Ga(\gamma_l \mid c_l^0, d_l^0)$$

γ_l's are independent gamma distributed

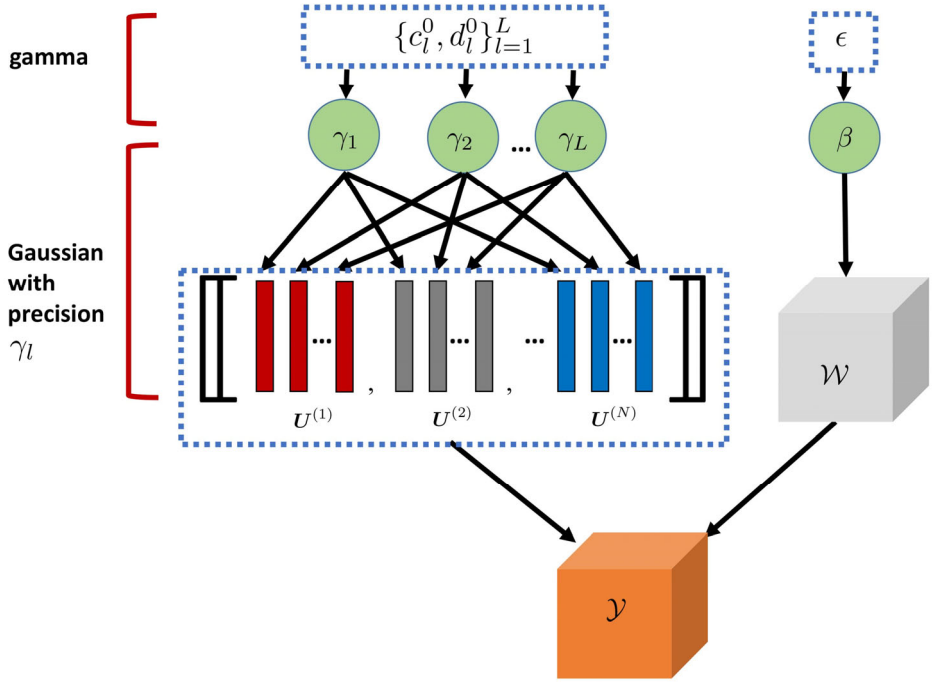
Gaussian-gamma hierarchy gives sparsity in the columns of $\mathbf{U}^{(n)}$

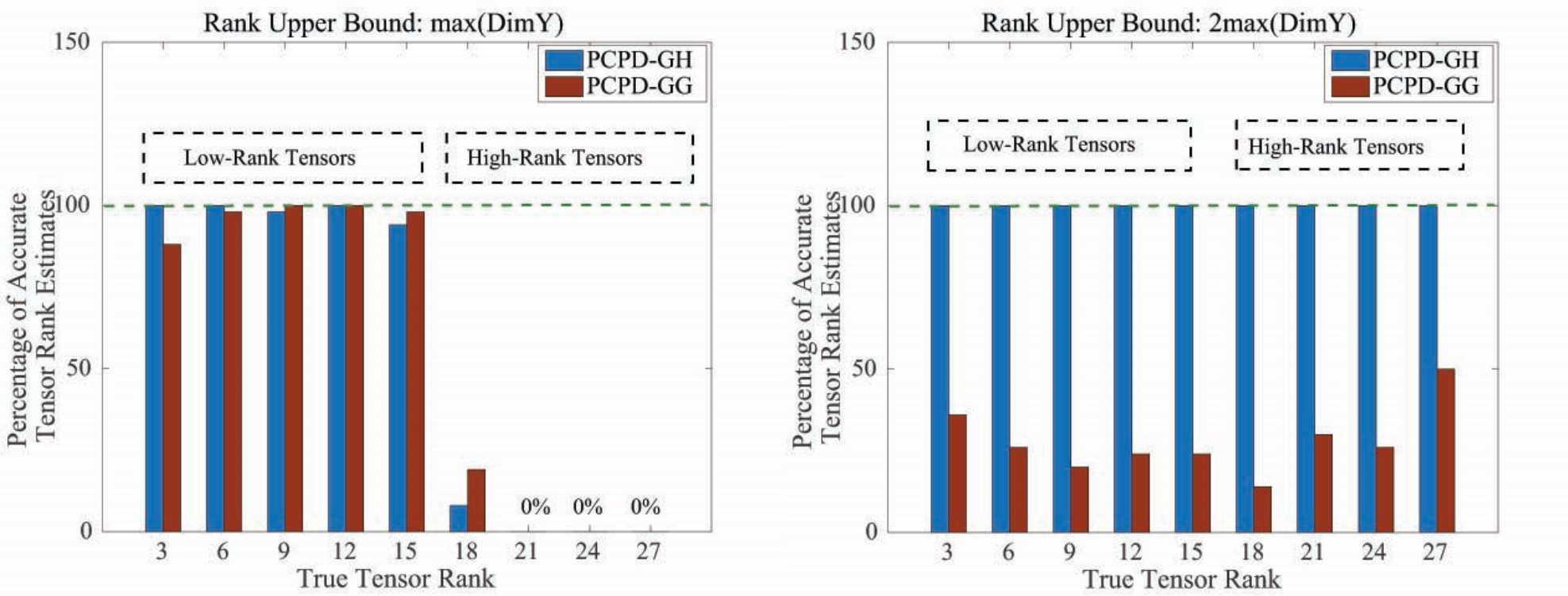
- Likelihood function

$$p(\mathcal{Y} \mid \{\mathbf{U}^{(n)}\}_{n=1}^{N}, \beta) \propto \exp\left(-\frac{\beta}{2} \left\| \mathcal{Y} - [\![\mathbf{U}^{(1)}, \mathbf{U}^{(2)}, \ldots, \mathbf{U}^{(N)}]\!] \right\|_F^2 \right)$$

$$p(\beta \mid \varepsilon) = Ga(\beta \mid \varepsilon, \varepsilon)$$

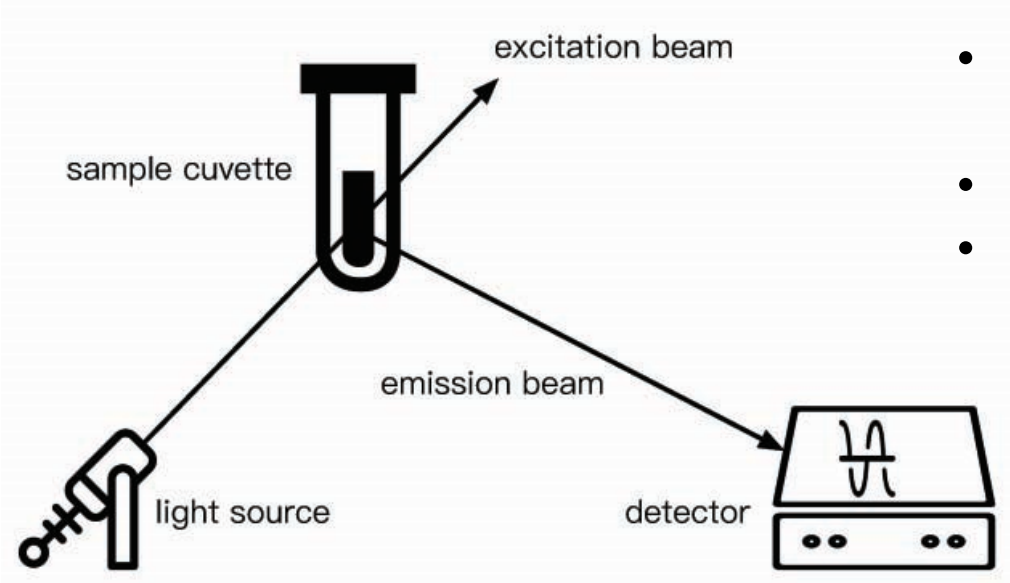
- Conjugacy exists between directly connected variables $\longrightarrow$ iterative closed-form variational distribution update available
- A more advanced sparsity enhancing prior for columns of $\mathbf{U}^{(n)}$ is generalized hyperbolic (GH) prior





- 3D tensor with dimension $30 \times 30 \times 30$, SNR=10dB
- GG prior is good for low-rank tensor
- GH prior can adapt to wide range of sparsity levels

• Application: Fluorescence Data Analytics



- A sample contains a mixture of R species (R is unknown)
- Excitation beam consists of I wavelengths
- Emission beam consists of J wavelengths

Concentration of the r^{th} specie

$$x_{ij} = \sum_{r=1}^{R} a_r(\lambda_i) b_r(\lambda_j) c_r$$

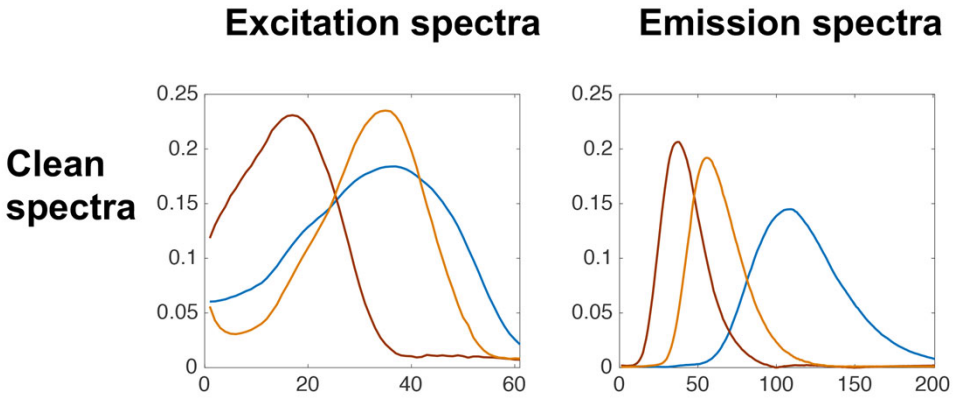
Excitation value of the r^{th} specie at wavelength λ_i

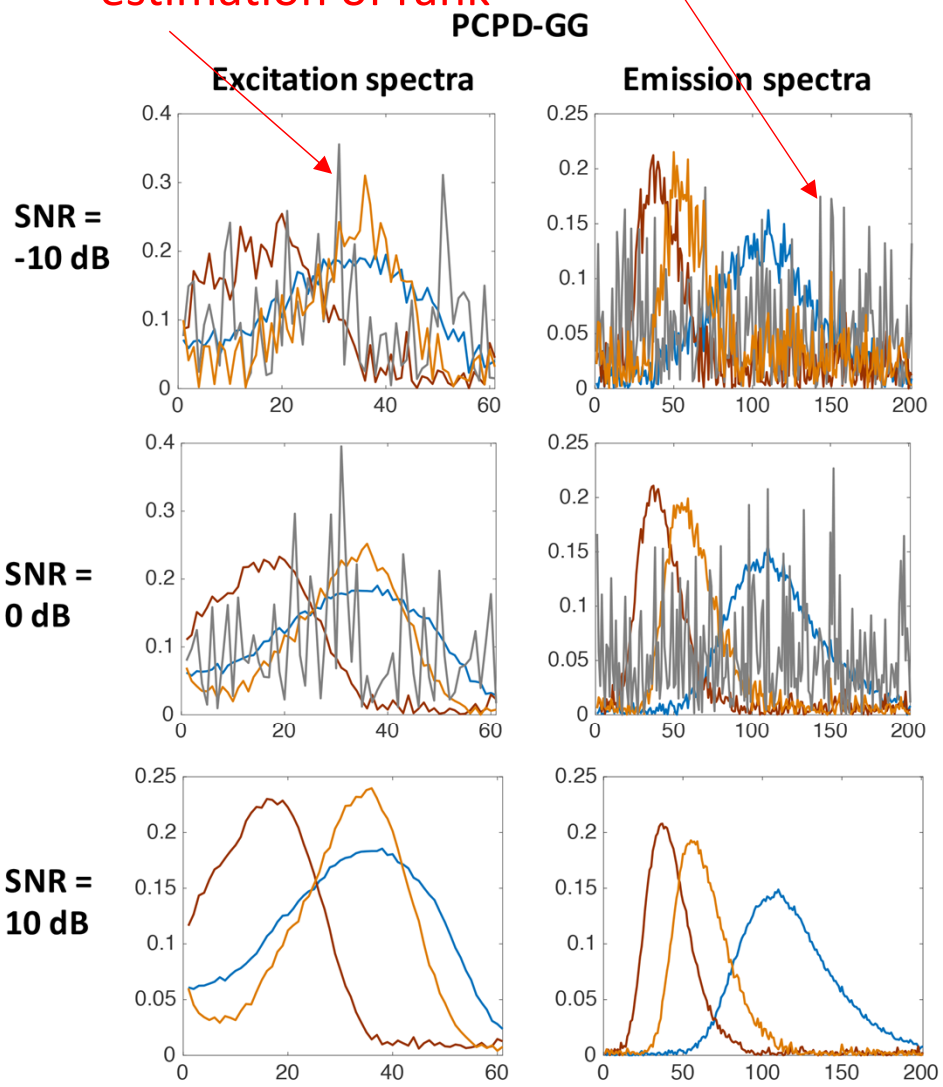
Emission value of the r^{th} specie at wavelength λ_j

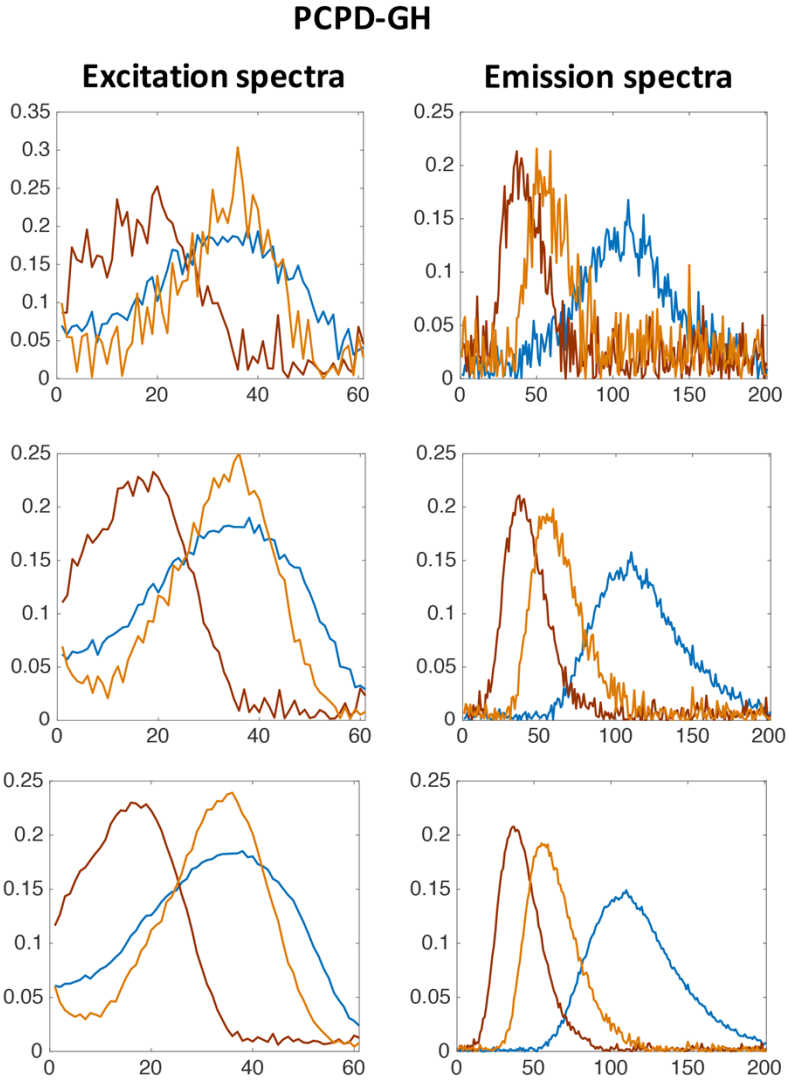
- If we have with the same chemical species but with different concentration of each specie, the observed data is a 3D tensor and obeys CPD
- Tasks:
  - Estimate how many chemical species in the samples (revealed by R)
  - Determine what kind of species in the samples (revealed by $a_r(\lambda_i)$, $b_r(\lambda_j)$)

- Amino acids fluorescence data
- Size $61 \times 201 \times 5$, true rank =3
- Clean spectra recovered by ALS assuming we know R

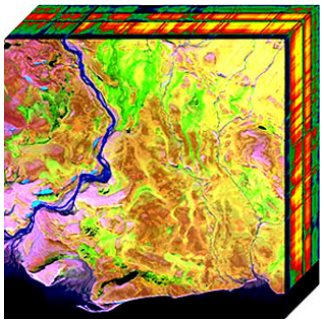
Ghost component due to wrong estimation of rank





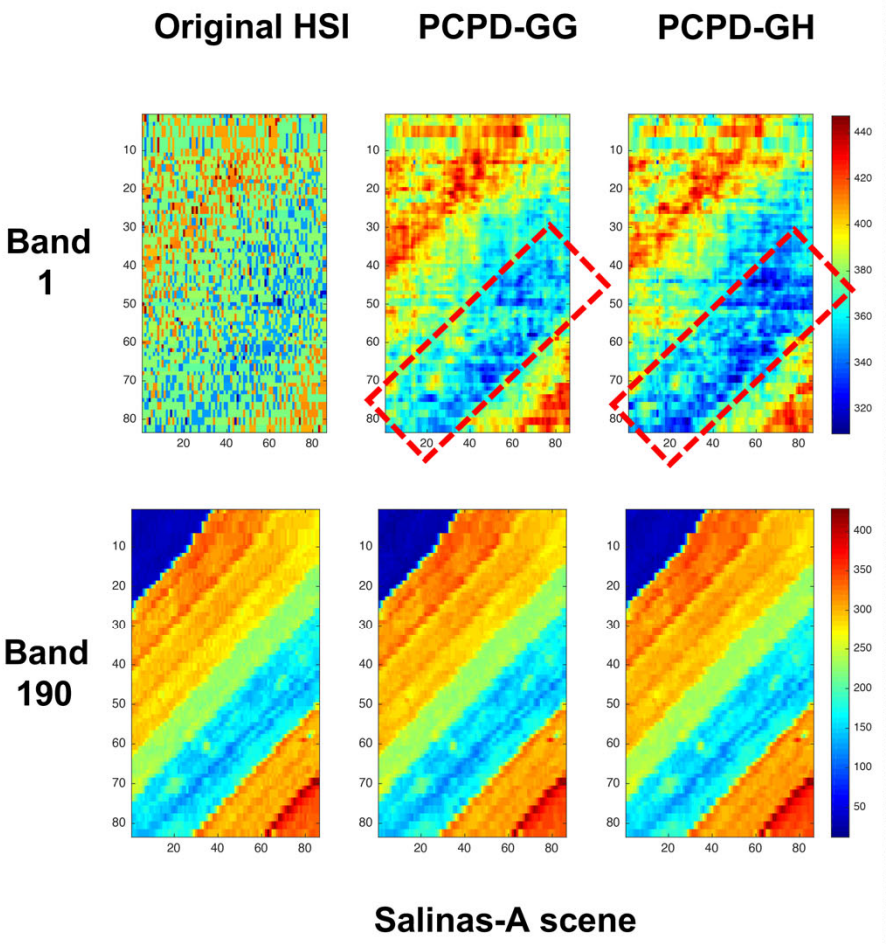


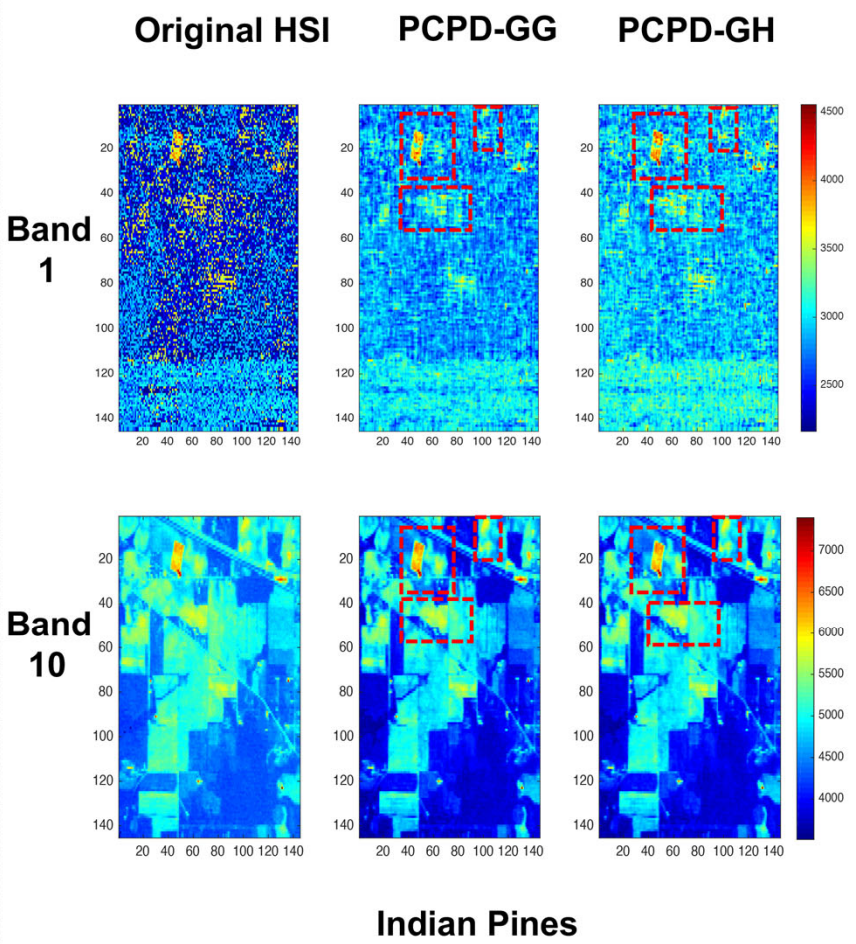
• Application: Hypersepctral Images Denoising



- Measure spectrum of light in each pixel
- Remote sensing: finding objects, identifying minerals

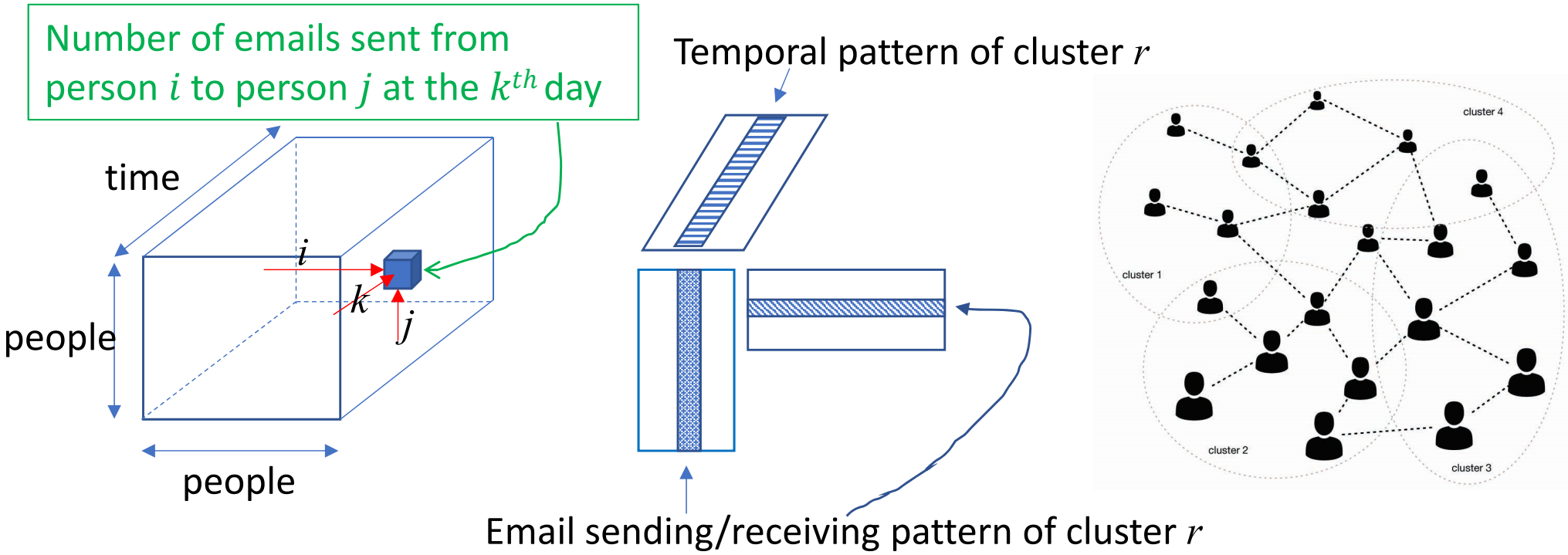
Algorithm		PCPD-GG		PCPD-GH	
Dataset	Rank Upper Bound	SNR Output (dB)	Estimated Tensor Rank	SNR Output (dB)	Estimated Tensor Rank
Salinas-A	$\max\{J_n\}_{n=1}^{N}$	43.7374	137	**44.0519**	143
	$2\max\{J_n\}_{n=1}^{N}$	46.7221	257	**46.7846**	260
Indian Pines	$\max\{J_n\}_{n=1}^{N}$	30.4207	169	**30.5541**	178
	$2\max\{J_n\}_{n=1}^{N}$	31.9047	317	**32.0612**	335





CPD with Non-negative Constraint

- In some applications, the factor matrices of CPD should contain non-negative values only
- For example: 3D email data set



- Rank of CPD is the number of cluster
- Factor matrices should contain non-negative values

- Optimization-based formulation:

$$\min_{\mathbf{A},\mathbf{B},\mathbf{C}} \left\| \mathcal{Y} - [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] \right\|_F^2$$

$$\text{s.t.} \quad \mathbf{A} \geq \mathbf{0}_{I \times R}, \mathbf{B} \geq \mathbf{0}_{J \times R}, \mathbf{C} \geq \mathbf{0}_{K \times R}$$

Still ALS but with an additional projection step

- Challenge of Bayesian modeling: the usual Gaussian-gamma prior supports both positive and negative values
- Truncated Gaussian-gamma prior:

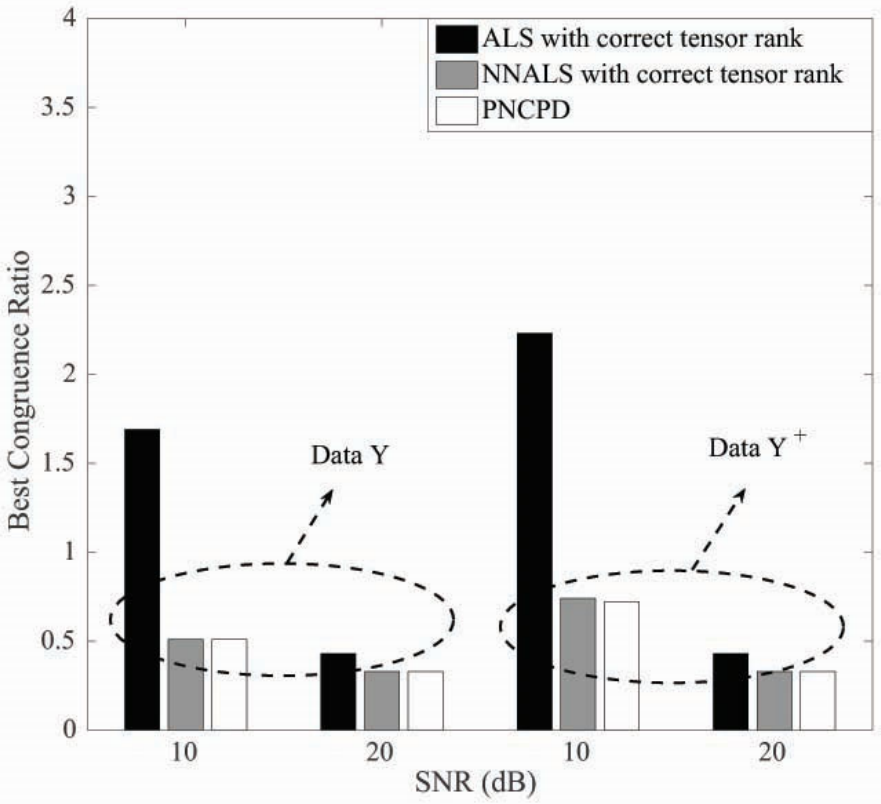
$$p(\{\boldsymbol{\Xi}^{(n)}\}_{n=1}^{N} \mid \{\gamma_l\}_{l=1}^{L}) = \prod_{n=1}^{N} \prod_{l=1}^{L} \frac{\mathcal{N}(\boldsymbol{\Xi}_{:,l}^{(n)} \mid \mathbf{0}_{J_n \times 1}, \gamma_l^{-1} \mathbf{I}_L)}{\int_{\mathbf{0}}^{\infty} \mathcal{N}(\boldsymbol{\Xi}_{:,l}^{(n)} \mid \mathbf{0}_{J_n \times 1}, \gamma_l^{-1} \mathbf{I}_L) d\boldsymbol{\Xi}_{:,l}^{(n)}} U(\boldsymbol{\Xi}_{:,l}^{(n)} \geq \mathbf{0}_{J_n \times R})$$

$$p(\{\gamma_l\}_{l=1}^{L} \mid \{c_l^0, d_l^0\}_{l=1}^{L}) = \prod_{l=1}^{L} Ga(\gamma_l \mid c_l^0, d_l^0)$$

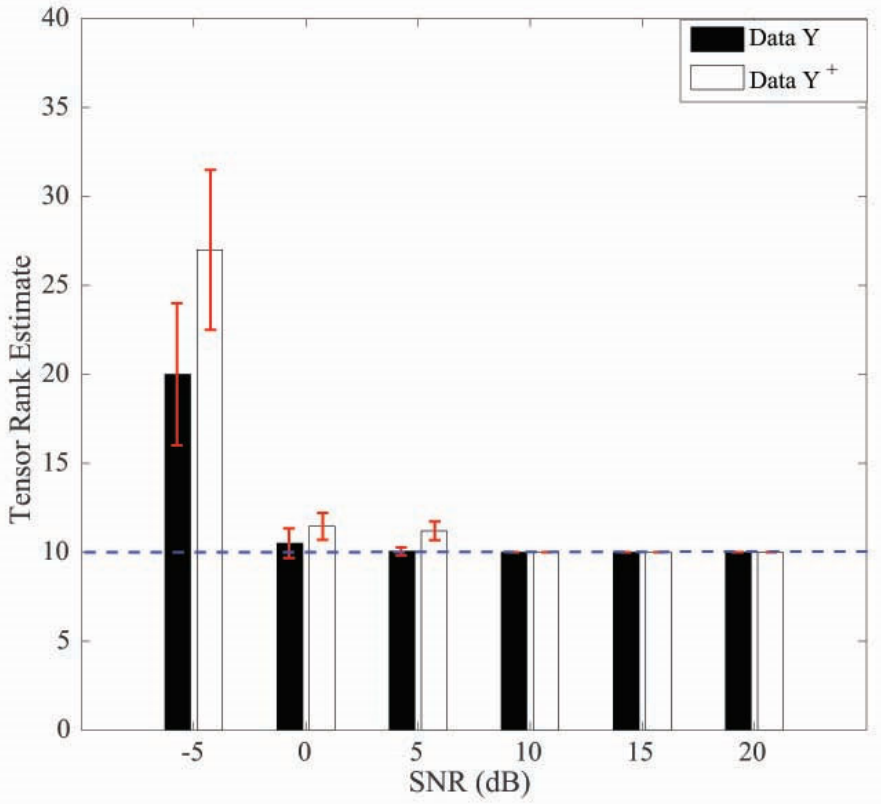
Still sparsity promoting and satisfy the conjugacy

- In variational inference, each variational distribution update depends on the moments of other variational distributions
- Moments of truncated Gaussian is hard to obtain
- In additional the usual mean field, we employ $q(\boldsymbol{\Xi}^{(k)}) = \delta(\boldsymbol{\Xi}^{(k)} - \hat{\boldsymbol{\Xi}}^{(k)})$
- This corresponds to the variational EM, where $\boldsymbol{\Xi}^{(k)}$ is obtained as point estimate (by optimization), while other parameters are updated in their variational distributions

- Synthetic data: size $100 \times 100 \times 100$, true rank=10

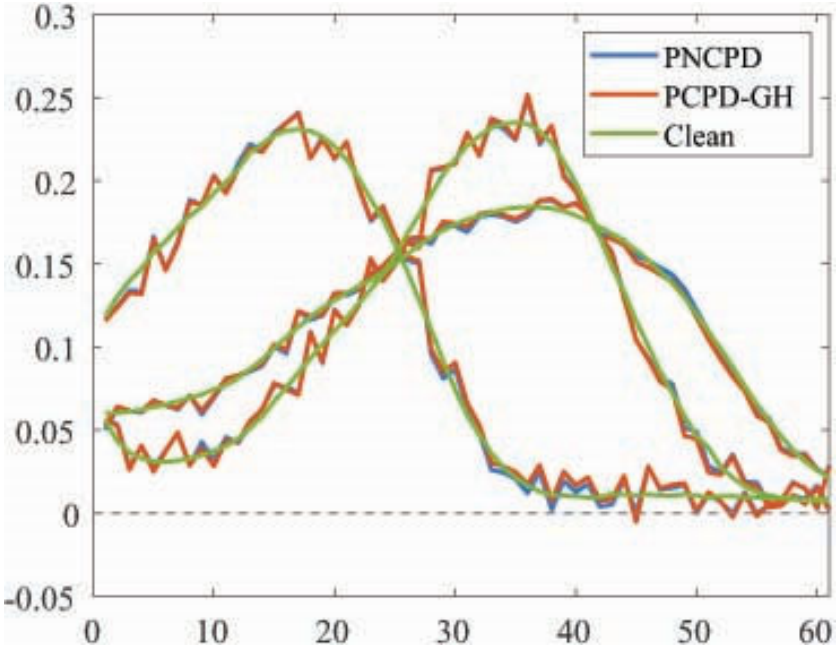


Factor matrices recovery relative error

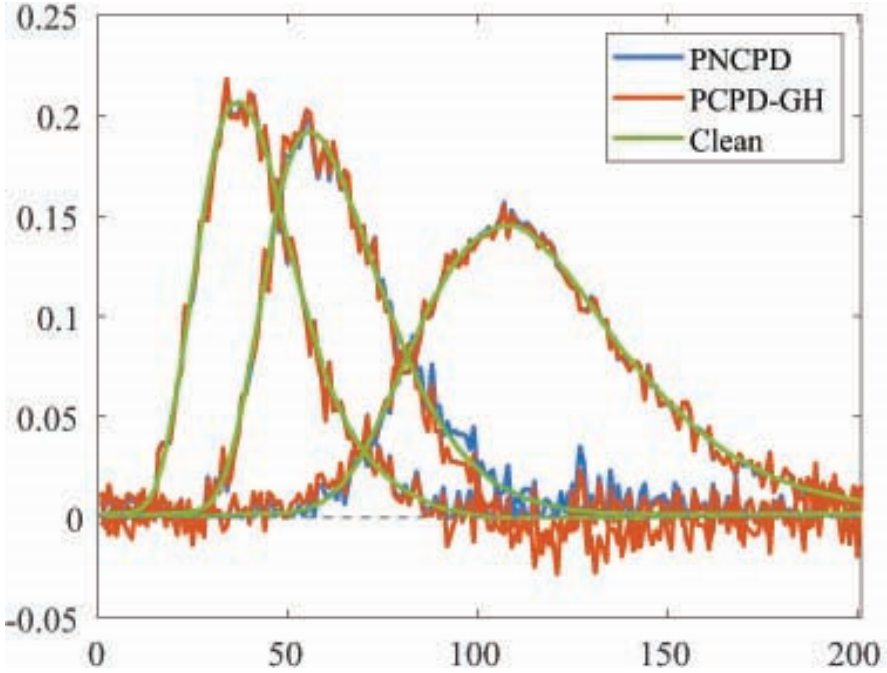


Tensor rank estimation accuracy

- Revisiting amino acids fluorescence data

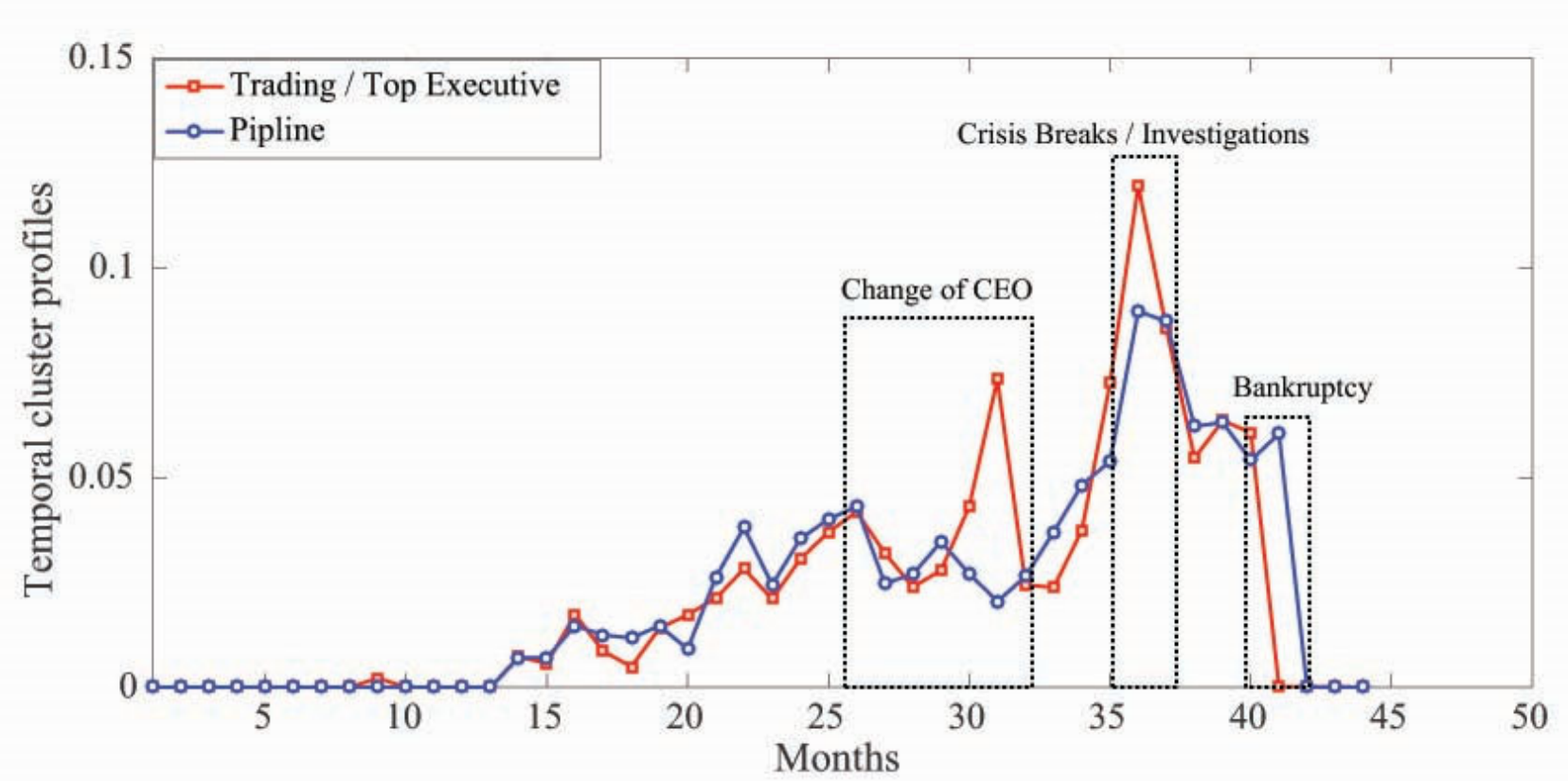


Recovered excitation spectra



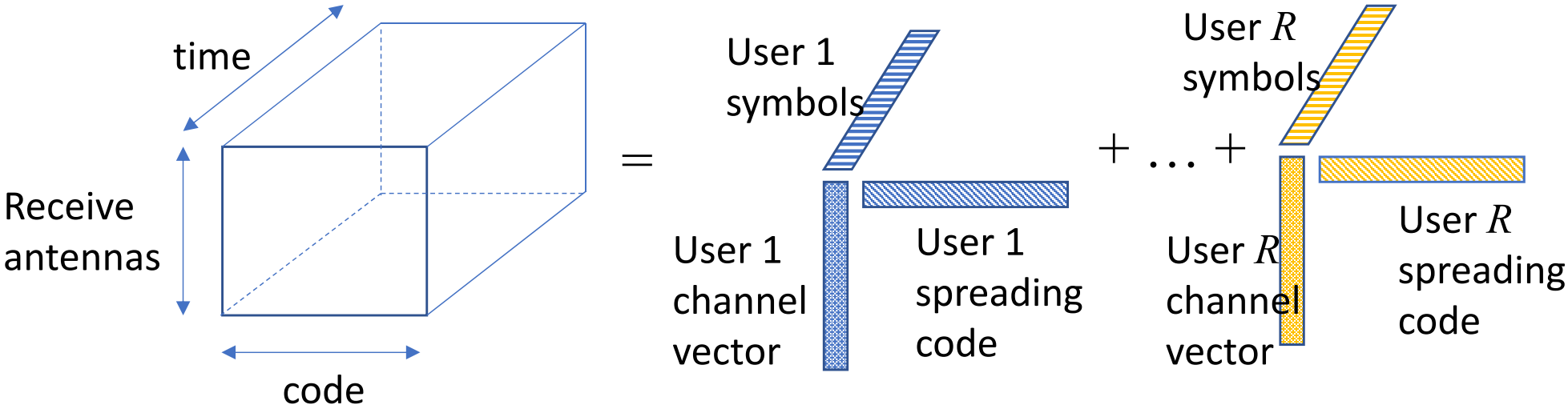
Recovered emission spectra

- ENRON E-mail Data Mining (184 people, 44 months)
- Bayesian non-negative CPD identifies 4 clusters
- Vanilla Bayesian CPD identifies 5 clusters



Orthogonal Constraint and outlier modeling

- Blind Data Detection for DS-CDMA Systems



$$\mathcal{Y} = \sum_{r=1}^{R} \mathbf{H}_{:,r} \circ \mathbf{C}_{:,r} \circ \mathbf{S}_{:,r} + \mathcal{W} = [\![\mathbf{H}, \mathbf{C}, \mathbf{S}]\!] + \mathcal{W}$$

Symbols are uncorrelated among users ⇒ $\mathbf{S}$ is approximately orthogonal (i.e., $\mathbf{S}^H\mathbf{S}=\mathbf{I}$)

- Linear Image Coding

Objective function =

$$\| \mathcal{Y} - \sum_{r=1}^{R} \mathbf{U}_{:,r} \circ \mathbf{V}_{:,r} \circ \mathbf{D}_{:,r} \|_F^2$$

K images each of size $M \times Z$

$$\min_{\mathbf{U},\mathbf{V},\{d_r(k)\}_{r=1}^{R}} \sum_{k=1}^{K} \left\| \mathbf{B}(k) - \mathbf{U}\,\mathrm{diag}\{d_1(k),\ldots,d_R(k)\}\mathbf{V}^T \right\|_F^2$$

$$\text{s.t.} \quad \mathbf{U}^H\mathbf{U} = \mathbf{I}_R, \quad \mathbf{V}^H\mathbf{V} = \mathbf{I}_R$$

$\mathbf{U}$, $\mathbf{V}$ are common orthogonal basis for all images

- General problem formulation:

Existence of outliers

$$\min_{\{\boldsymbol{\Xi}^{(n)}\}_{n=1}^{N},\mathcal{E}} \beta \| \mathcal{Y} - [\![\boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}, \ldots, \boldsymbol{\Xi}^{(N)}]\!] - \mathcal{E} \|_F^2 + \sum_{l=1}^{L} \gamma_l \left(\sum_{n=1}^{N} \boldsymbol{\Xi}_{:,l}^{(n)H} \boldsymbol{\Xi}_{:,l}^{(n)} \right)$$

$$\text{s.t.} \quad \boldsymbol{\Xi}^{(n)H}\boldsymbol{\Xi}^{(n)} = \mathbf{I}_L, \quad n = 1,2,\ldots,P < N$$

For complexity control

Allow factor matrices to take complex values

Not all factor matrices are orthogonal

- Challenges: True rank is unknown, regularization parameters β, $\{\gamma_l\}_{l=1,\ldots,L}$ need to be tuned
- Probabilistic model
  - Log-likelihood:

$$p(\mathcal{Y} \mid \boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}, \ldots, \boldsymbol{\Xi}^{(N)}, \mathcal{E}, \beta) \propto \exp\left(-\beta \| \mathcal{Y} - [\![\boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}, \ldots, \boldsymbol{\Xi}^{(N)}]\!] - \mathcal{E} \|_F^2 \right)$$

  - Prior distributions:

Indicator function for orthogonal manifold: uniformly distributed in the manifold

$$p(\boldsymbol{\Xi}^{(1)}, \boldsymbol{\Xi}^{(2)}, \ldots, \boldsymbol{\Xi}^{(P)}) \propto \prod_{n=1}^{P} \mathbb{I}_{V_L(\mathbb{C}^{I_n})}(\boldsymbol{\Xi}^{(n)})$$

$$p(\boldsymbol{\Xi}^{(P+1)}, \ldots, \boldsymbol{\Xi}^{(N)} \mid \{\gamma_l\}_{l=1}^{L}) = \prod_{n=P+1}^{N} \prod_{l=1}^{L} \mathcal{CN}(\boldsymbol{\Xi}_{:,l}^{(n)} \mid \mathbf{0}, \gamma_l^{-1} \mathbf{I}_L)$$

  - Hyper-prior distributions:

$$p(\beta) \propto Ga(\beta \mid \varepsilon, \varepsilon), \quad p(\gamma_l) \propto Ga(\gamma_l \mid \varepsilon, \varepsilon)$$

  - Outlier prior distribution:

$$\mathcal{E}_{i_1,\ldots,i_N} \sim \mathcal{CN}(\mathcal{E}_{i_1,\ldots,i_N} \mid 0, \xi_{i_1,\ldots,i_N}^{-1}), \quad \xi_{i_1,\ldots,i_N} \sim Ga(\xi_{i_1,\ldots,i_N} \mid \varepsilon, \varepsilon)$$

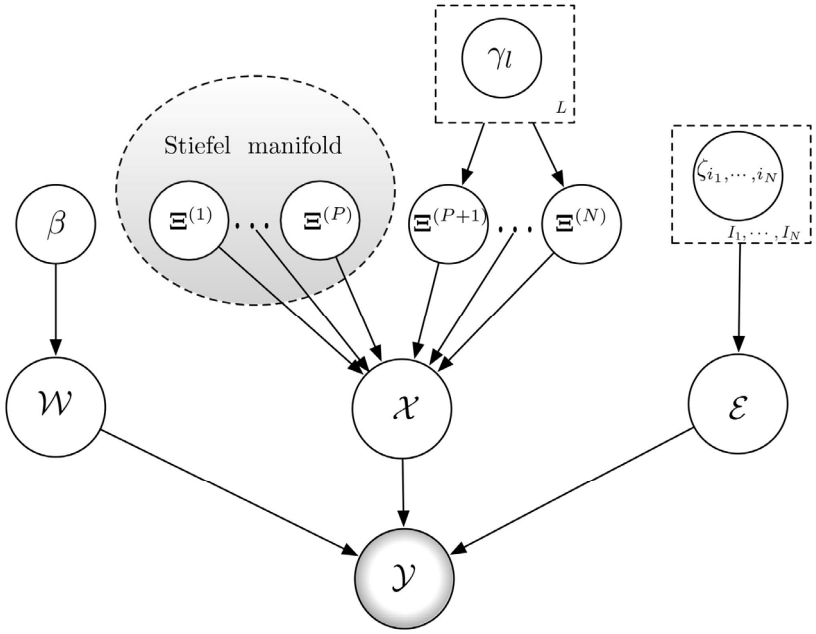
This Gaussian-gamma construction induces Sparsity

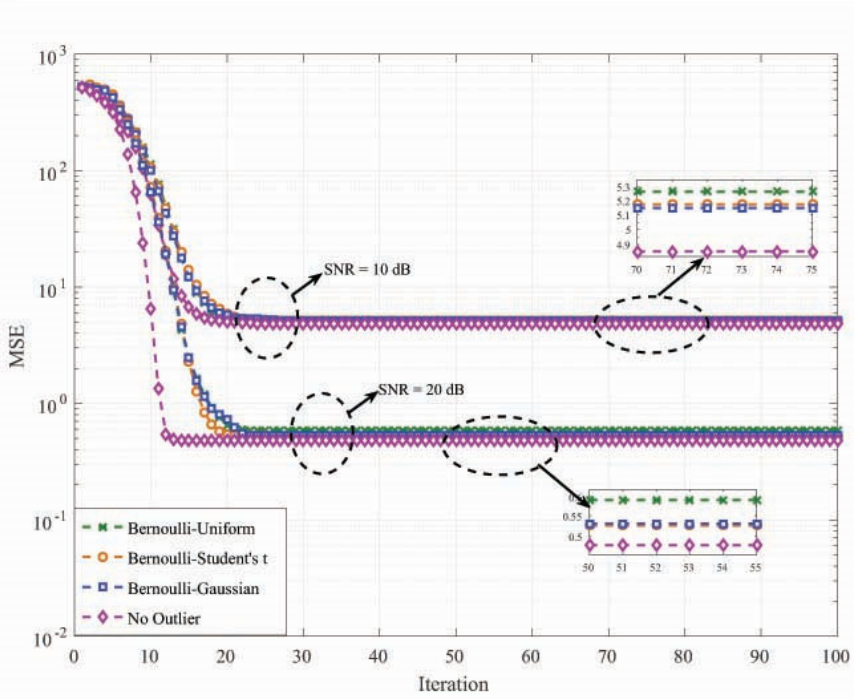
- To handle the orthogonal constraint, we employ

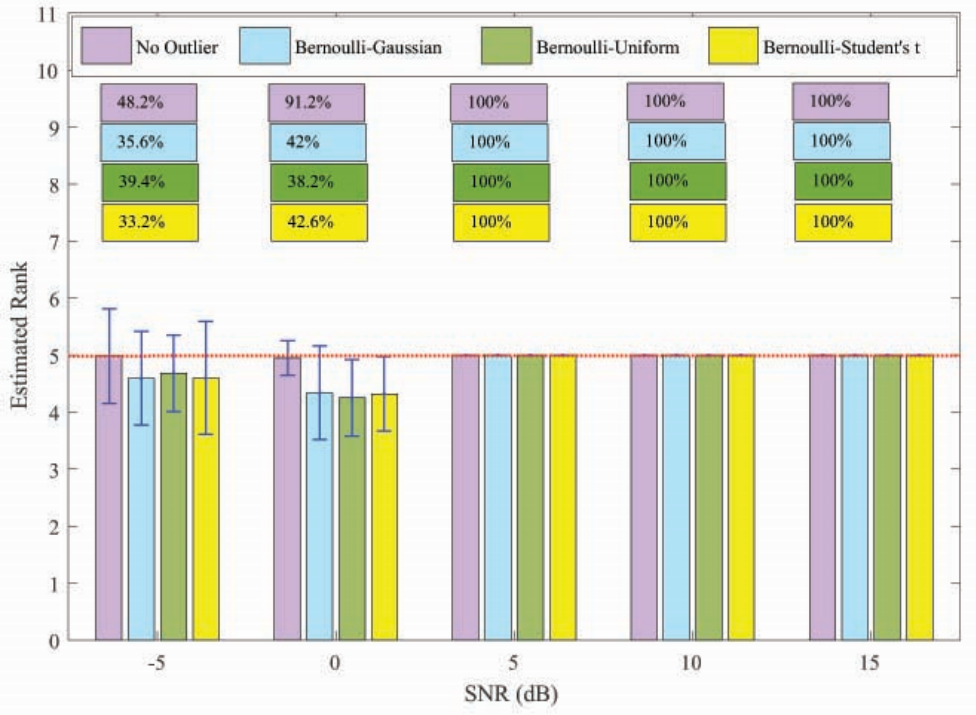
$$q(\boldsymbol{\Xi}^{(k)}) = \delta(\boldsymbol{\Xi}^{(k)} - \hat{\boldsymbol{\Xi}}^{(k)})$$

in additional to the usual mean-field approximation in variational inference

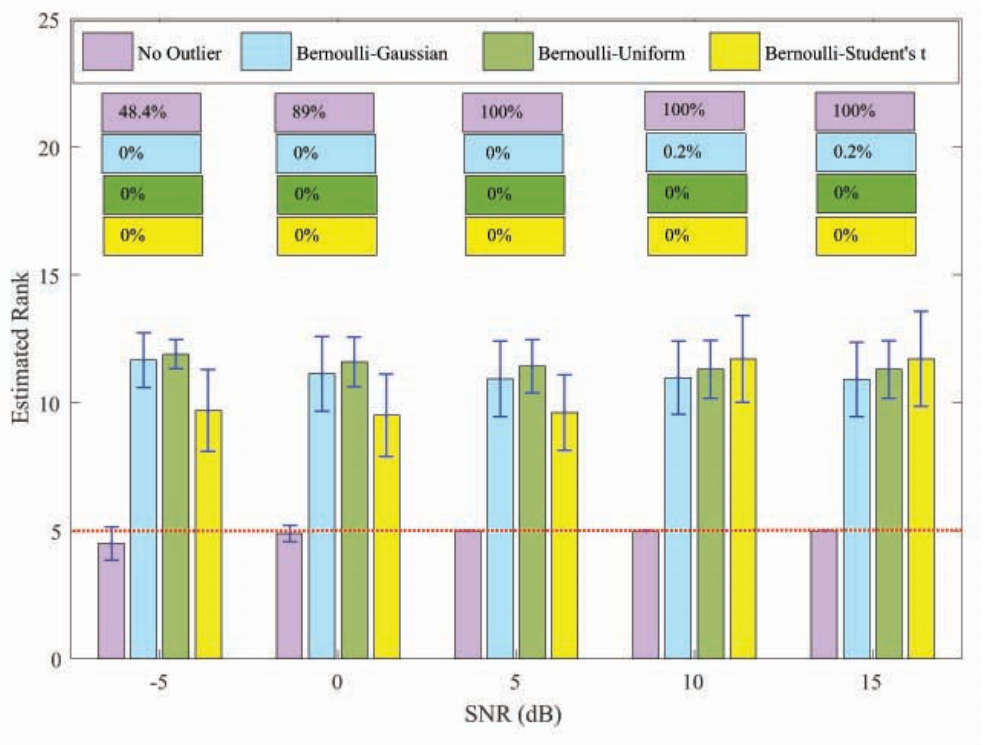
- Different variational distributions are updated iteratively, and closed-form solution is available in each update step
- Synthetic data: $12 \times 12 \times 12$, true rank=5, one orthogonal factor, test on three outlier models





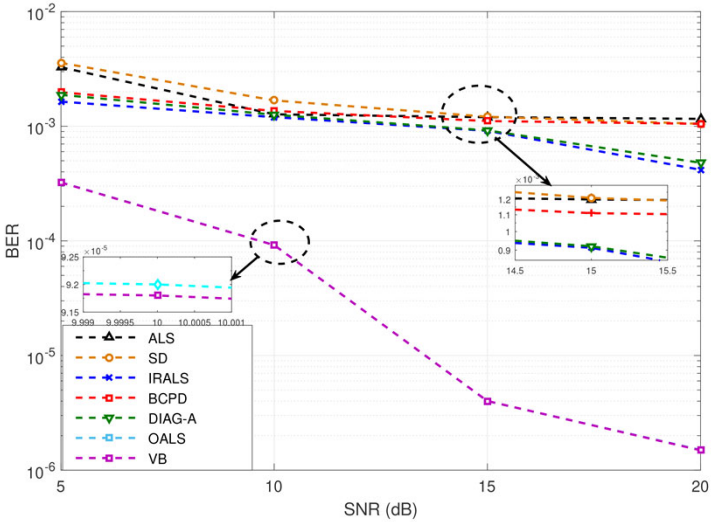


Bayesian CPD model including outlier modeling

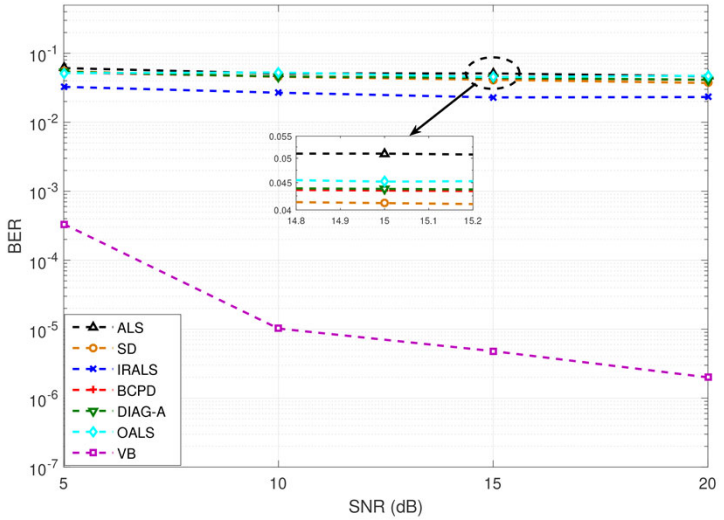


Vanilla Bayesian CPD model

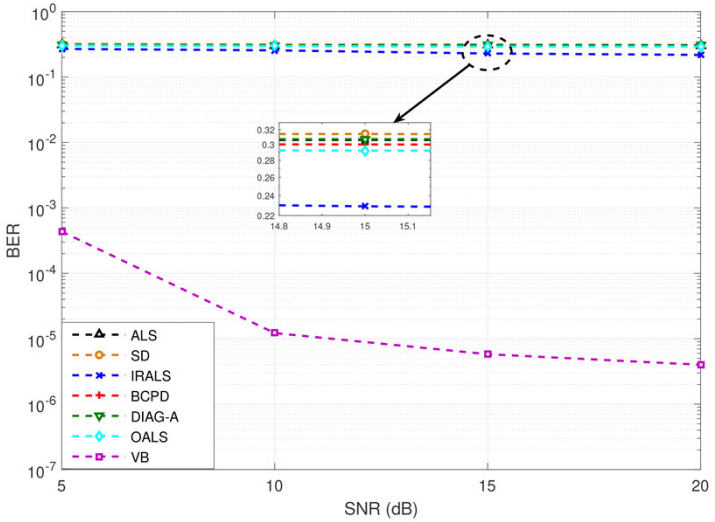
- CDMA example: $8 \times 6 \times 100$ complex-valued received tensor, R=5 users



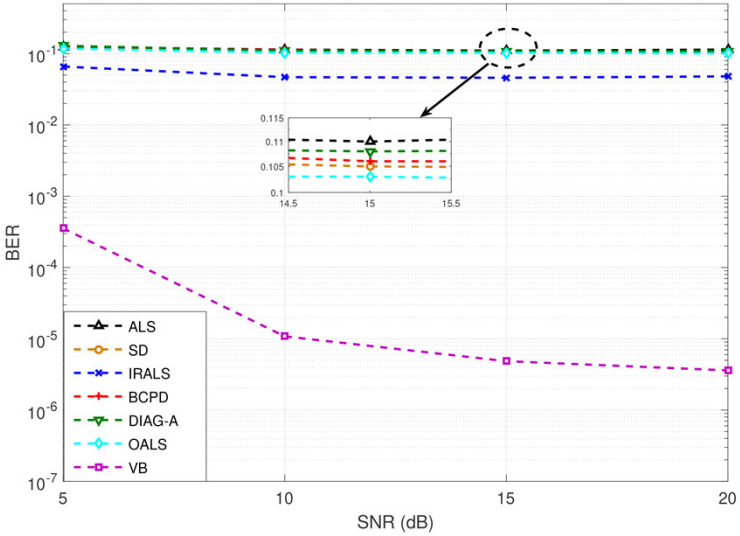
No outlier



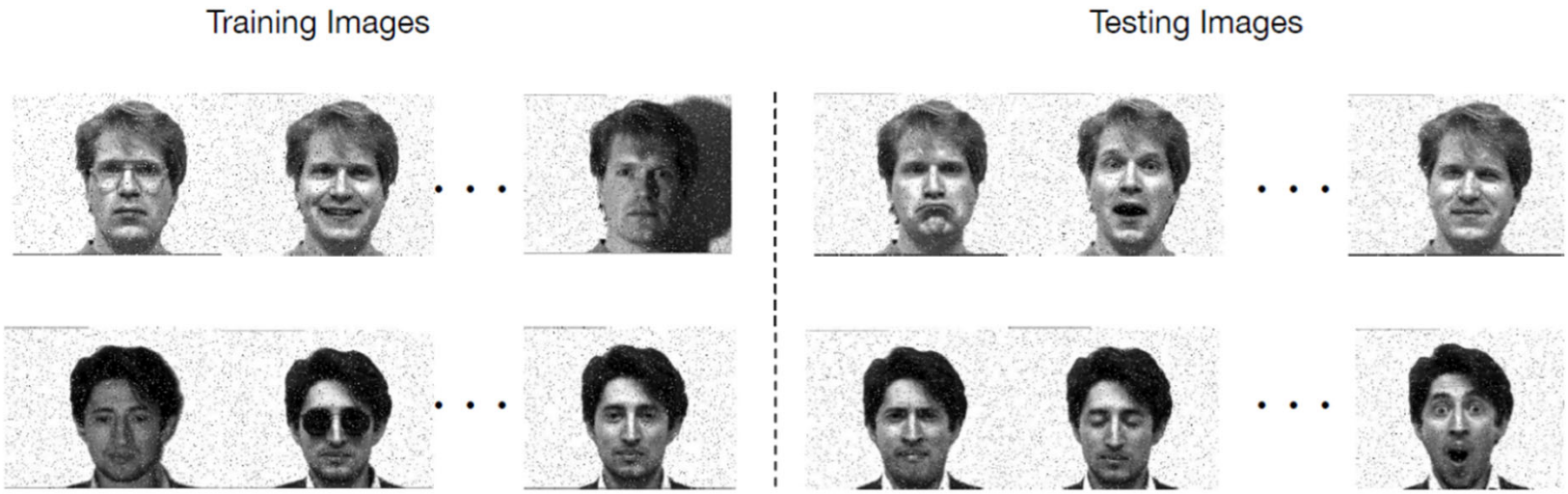
Bernoulli-Gaussian



Bernoulli-Uniform



Bernoulli-Student's t



- The task: to determine the person in photo is person A or B
- Use tensor decomposition for feature extraction and support vector machine (SVM) for classification
- Only requires 6 images from each person for training

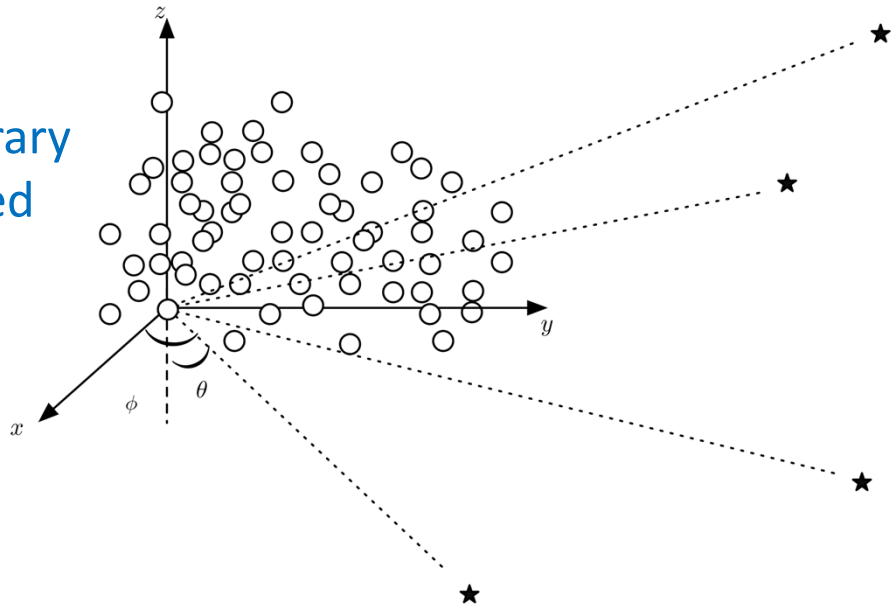
Algorithm	No Outlier		Bernoulli-Gaussian		Bernoulli-Uniform		Bernoulli-Student's t	
	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)	Classification Error	CPD Time (s)
ALS	9%	1.9635	51%	46.3041	47%	63.6765	48%	58.1047
SD	12%	0.5736	49%	0.6287	44%	0.6181	49%	0.6594
IRALS	9%	4.3527	43%	15.8361	27%	16.6151	36%	17.5181
BCPD	2%	4.1546	53%	20.7338	35%	17.9896	48%	17.1151
DIAG-A	11%	3.7384	50%	28.9961	41%	30.6317	51%	22.5127
OALS	11%	1.0174	58%	40.2731	34%	38.1754	47%	24.0162
VB	2%	2.4827	10%	2.7806	6%	2.4912	7%	2.8895

CPD with missing values

- Application: Subspace estimation

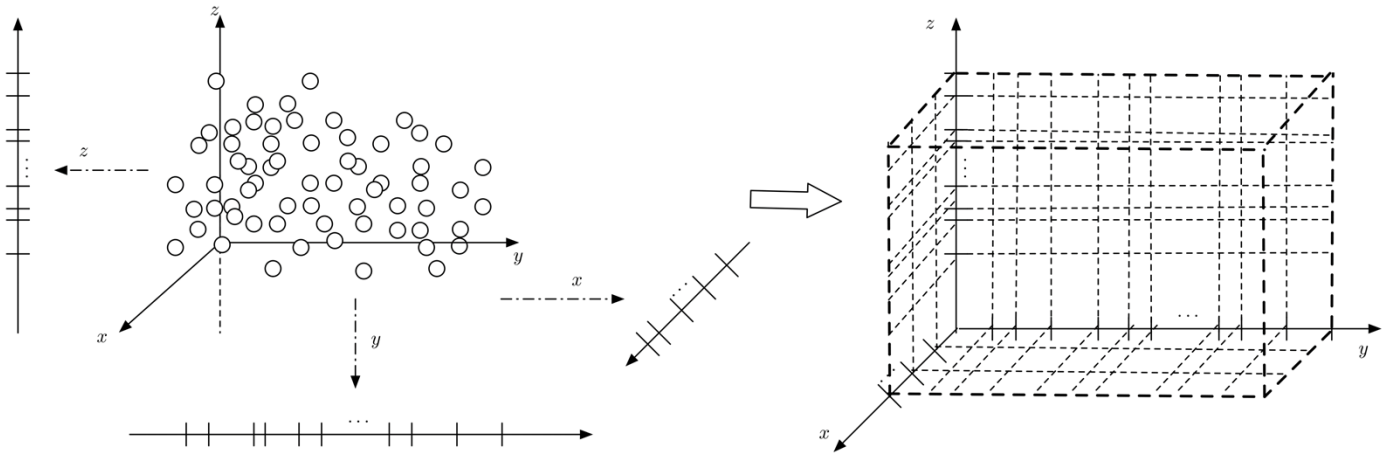
Unknown number of signals impinging on the array

Arbitrary shaped array



Related to elevation angle θ_r and azimuth angle ϕ_r

$$y_{x_m,y_m,z_m}(n) = \sum_{r=1}^{R} \xi_r(n) \exp\left\{ j[x_m u_r + y_m v_r + z_m p_r] \right\} + w_{x_m,y_m,z_m}(n)$$



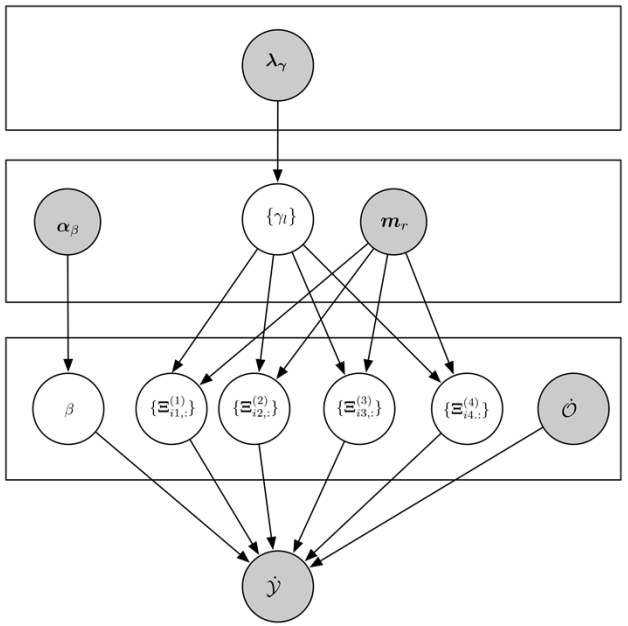
View the arbitrary array as a cuboid array but many missing elements

$$\dot{\mathcal{Y}} = \dot{O} \odot \left(\underbrace{\sum_{r=1}^{R} \mathbf{a}(u_r) \circ \mathbf{a}(v_r) \circ \mathbf{a}(p_r) \circ \mathbf{a}(\xi_r)}_{[\![\mathbf{A}[\mathbf{u}], \mathbf{A}[\mathbf{v}], \mathbf{A}[\mathbf{p}], \mathbf{A}[\xi]]\!]} + \dot{\mathcal{W}} \right)$$

The estimated rank is the number of sources

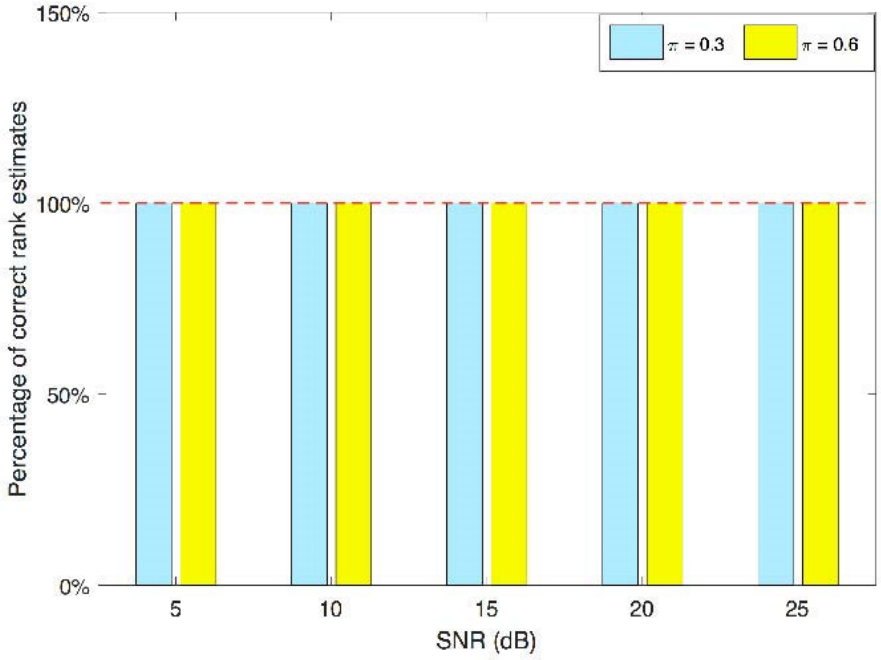
This is a CPD with missing elements

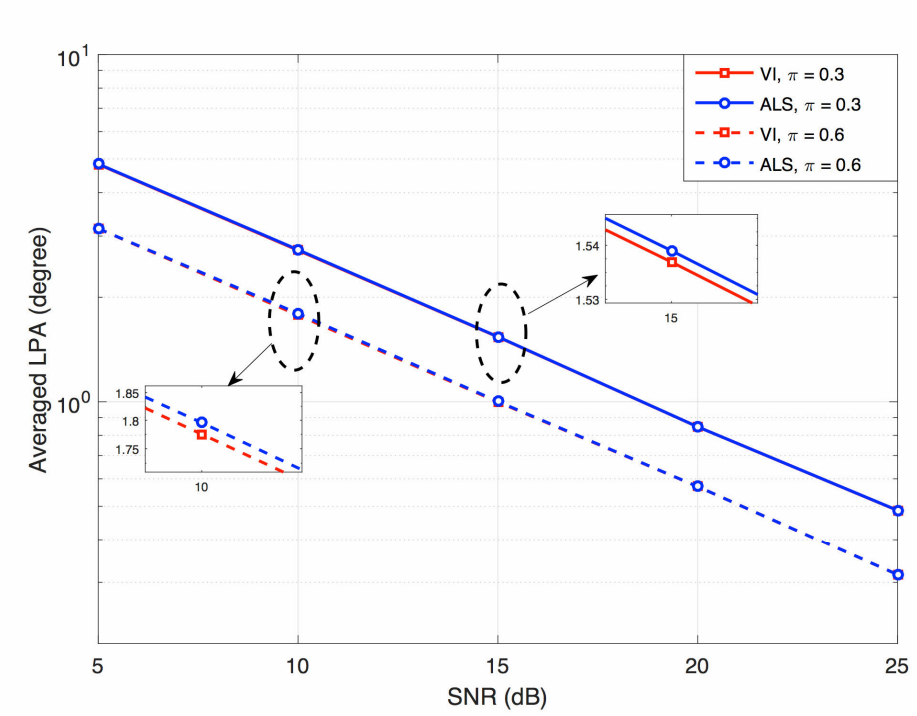
Once we obtain these factor matrices, we can recover the elevation angle θ_r and azimuth angle ϕ_r



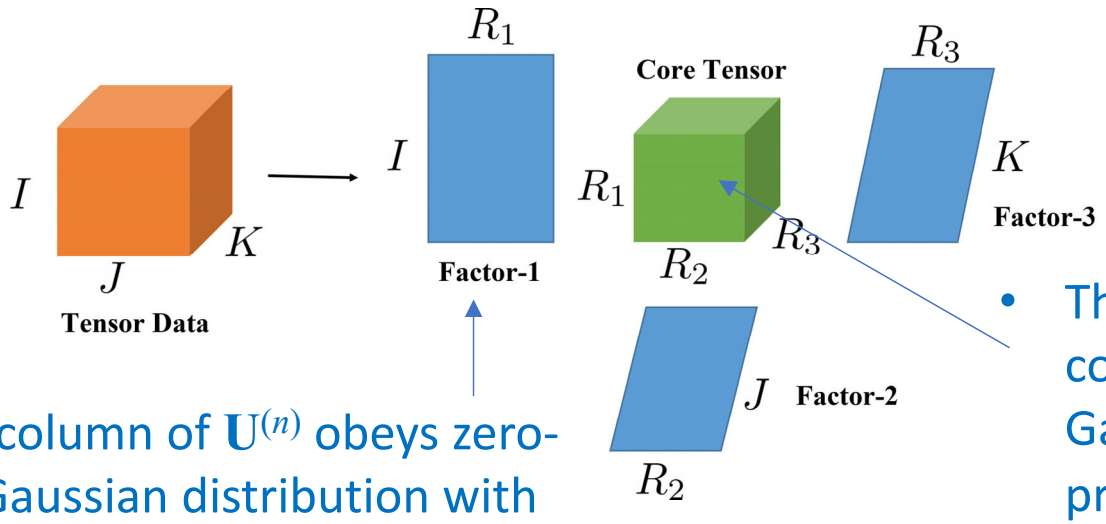
- Again use Gaussian-gamma prior for the columns of $\mathbf{A}[\mathbf{u}]$, $\mathbf{A}[\mathbf{v}]$, $\mathbf{A}[\mathbf{p}]$, $\mathbf{A}[\xi]$

- Cuboid array $8 \times 10 \times 12$, time dimension 100, occupy ratio: 0.3 or 0.6, 3 sources



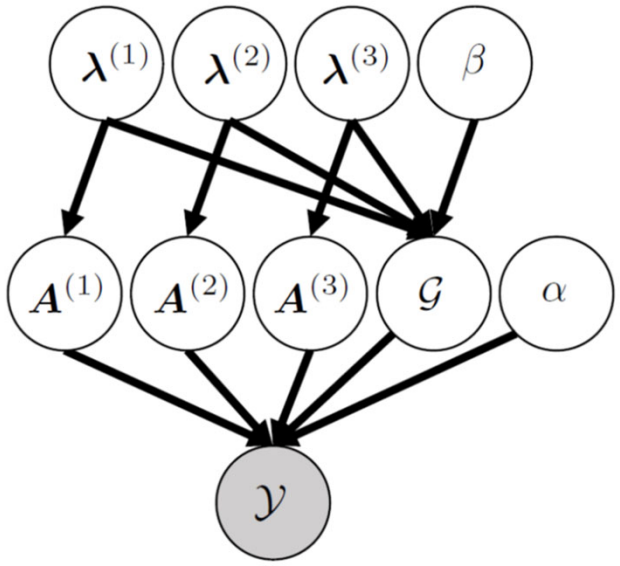


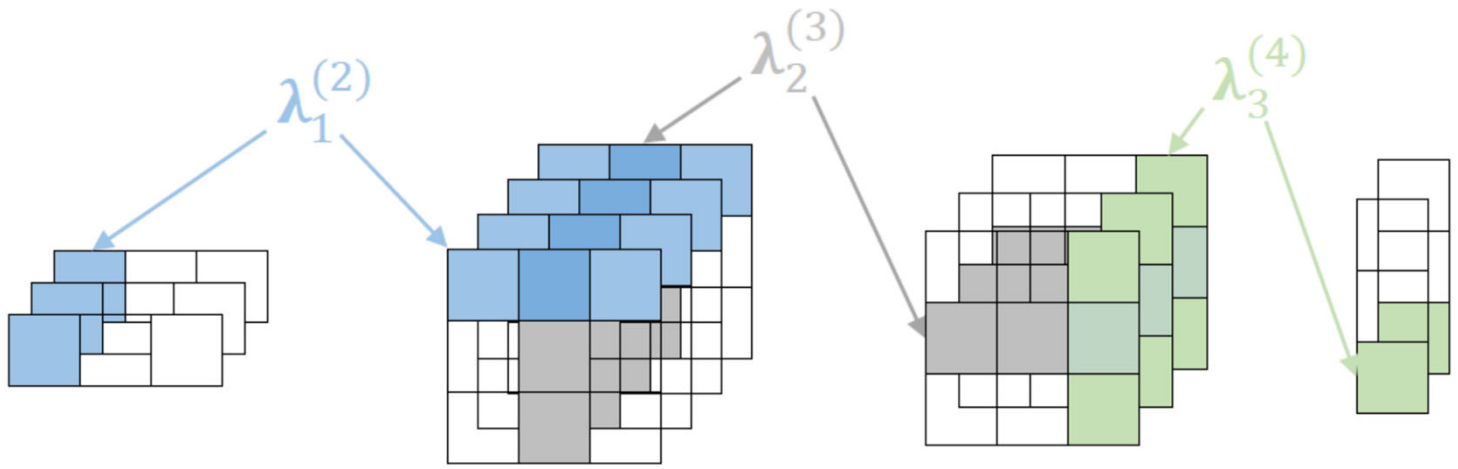
Beyond CPD



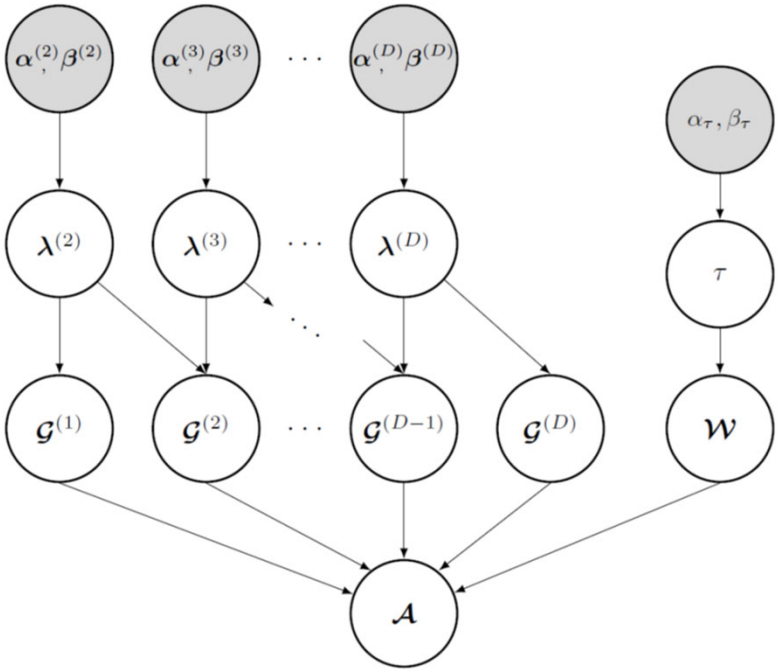
- The r^{th} column of $\mathbf{U}^{(n)}$ obeys zero-mean Gaussian distribution with precision $\lambda_r^{(n)}$, which itself obeys gamma distribution
- $\lambda_r^{(n)}$ does not share among factor matrices such that ranks in different factor matrices are different

- The $(r_1, r_2, \ldots, r_N)^{th}$ element of core tensor obeys zero mean Gaussian distribution with precision $\beta \prod_{n=1}^{N} \lambda_{r_n}^{n}$
- β obeys another gamma distribution





- The $(k,l)^{th}$ fiber of the d^{th} TT core is modeled as a Gaussian vector with zero mean and precision $\lambda_k^{(d)}\lambda_l^{(d+1)}$
- $\lambda_k^{(d)}$ obeys gamma distribution
- This make sure the TT adjacent TT core dimensions are compatible for multiplication
- Conjugacy can be established in terms of TT core fiber



- Upcoming book:
- Lei Cheng, Zhongtao Chen and Yik-Chung Wu, *Bayesian Tensor Decomposition for Signal Processing and Machine Learning*, Springer (expected date: early 2023)
- Matlab codes for various Bayesian matrix and tensor decompositions:
- https://github.com/leicheng-tensor/Reproducible-Bayesian-Tensor-Matrix-Machine-Learning-SOTA

References

- Bayesian matrix factorization:

[1] Yangge Chen, Lei Cheng, and Yik-Chung Wu, ``Bayesian Low-rank Matrix Completion with Dual-graph Embedding: Prior Analysis and Tuning-free Inference," *Signal Processing*. Volume 204, 2023, https://doi.org/10.1016/j.sigpro.2022.108826

- Bayesian CPD:

[2] Lei Cheng, Yik-Chung Wu, and H. Vincent Poor, ``Probabilistic Tensor Canonical Polyadic Decomposition With Orthogonal Factors," *IEEE Trans. on Signal Processing*, Vol. 65, no. 3, pp. 663-676, Feb 2017.

[3] Lei Cheng, Yik-Chung Wu, and H. Vincent Poor, ``Scaling Probabilistic Tensor Canonical Polyadic Decomposition to Massive Data," *IEEE Trans. on Signal Processing*, Vol. 66, no. 21, pp. 5534-5548, Nov 2018.

[4] Lei Cheng, Xueke Tong, Shuai Wang, Yik-Chung Wu, and H. Vincent Poor, ``Learning Nonnegative Factors from Tensor Data: Probabilistic Modeling and Inference Algorithm," in *IEEE Trans. on Signal Processing*, vol. 68, pp. 1792-1806, 2020, doi: 10.1109/TSP.2020.2975353.

[5] Lei Cheng, Zhongtao Chen, Qingjiang Shi, Yik-Chung Wu, and Sergios Theodoridis, ``Towards Flexible Sparsity-Aware Modeling: Automatic Tensor Rank Learning using The Generalized Hyperbolic Prior," *IEEE Trans. on Signal Processing*, vol. 70, pp. 1834-1849, 2022, doi: 10.1109/TSP.2022.3164200

[6] Zhongtao Chen, Lei Cheng, and Yik-Chung Wu, ``Accelerating Probabilistic Tensor Canonical Polyadic Decomposition with Nonnegative Factors: An Inexact BCD Approach," in *Signal Processing*

- Bayesian Tensor Train:

[7] Le Xu, Lei Cheng, Ngai Wong, and Yik-Chung Wu, ``Overfitting Avoidance in Tensor Train Factorization and Completion: Prior Analysis and Inference," Proceedings of the *IEEE International Conference on Data Mining* (ICDM) 2021.

- Bayesian tensor in wireless channel estimation:

[8] Lei Cheng, Yik-Chung Wu, Jianzhong (Charlie) Zhang, and Lingjia Liu, ``Subspace Identification for DOA Estimation in Massive / Full-dimension MIMO System: Bad Data Mitigation and Automatic Source Enumeration,'' *IEEE Trans. on Signal Processing,* Vol. 63, no. 22, pp. 5897-5909, Nov 2015.

[9] Lei Cheng, Chengwen Xing, and Yik-Chung Wu, ``Irregular Array Manifold Aided Channel Estimation in Massive MIMO Communications," *IEEE Journal of Selected Topics in Signal Processing*, Vol. 13, no. 5, pp. 974-988, Sep 2019.