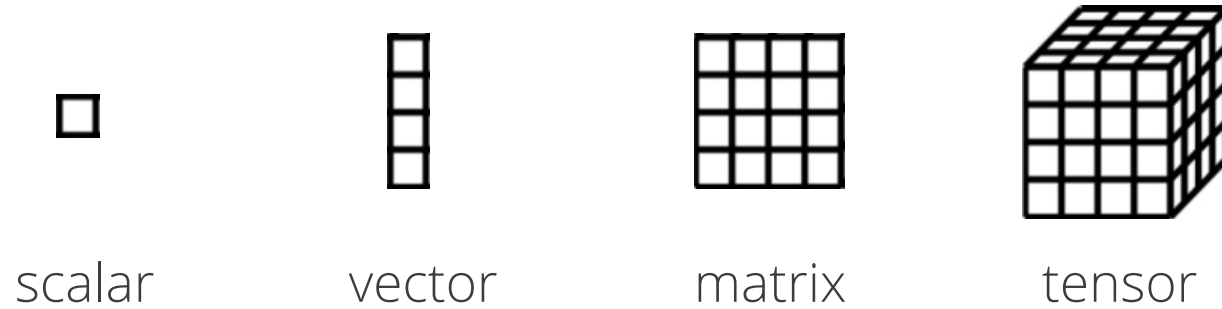# Sandia National Laboratories

Exceptional service in the national interest

THE POISSON CANONICAL POLYADIC TENSOR MODEL AS A LATENT-VARIABLE MODEL

Carlos Llosa, Daniel M. Dunlavy, Richard B. Lehoucq,
Oscar Lopez, Arvind Prasadan

December 16 2023, CMStatistics Berlin

U.S. DEPARTMENT OF ENERGY

NNSA
National Nuclear Security Administration

OVERVIEW

scalar   vector   matrix   tensor

- **Goal:** Understand key relationships in tensor data

- **Current approach**
  - Low-rank tensor models (canonical polyadic, Tucker, tensor train, …)
  - Parameter inference via maximum likelihood estimation

- **Our approach**
  - Latent-variable model formulation
  - Parameter inference via complete-data loglikelihood
    - EM algorithms for maximum likelihood estimation
    - Fisher information matrix

# FROM DENSE-CONTINUOUS TO SPARSE-DISCRETE TENSOR DATA ANALYSIS

Netflix Prize [1]

- Sparse data: only ~1% of entries are observed
- Ranking data (1-5)
- Winner algorithm used matrix factorization techniques
- This led to increased interest in non-Gaussian matrix factorization

ICEWS Database [2]

- Countries as receivers and senders
- Events such as threats or aid
- Count data: number of times an event happens from a receiver to a sender
- Sparse data, low-count data
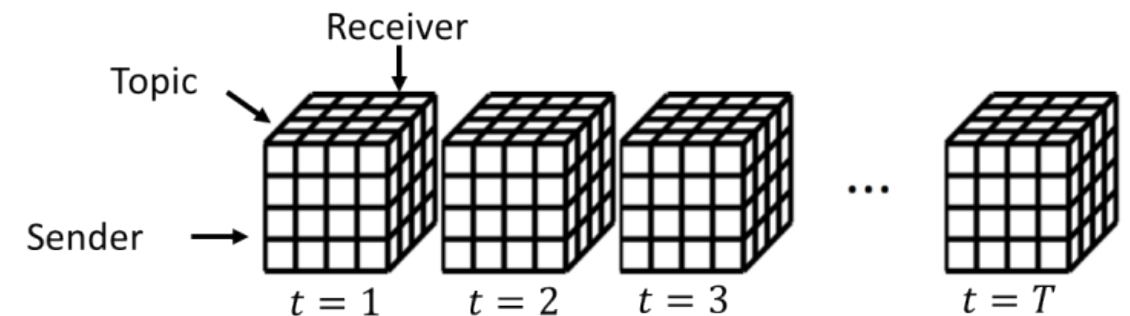
[1] Bennett and Lanning, *The Netflix Prize*, Proc. of KDD Cup and Workshop, 2007.
[2] O'Brien, *Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research*, ISR, 2010.

# POISSON CANONICAL POLYADIC (PCP) TENSOR MODEL [3][4]

The count tensor follows a Poisson distribution element-wise

$$\mathcal{X} \sim \text{Poisson}(\mathcal{M}) \iff \mathcal{X}_{i,j,k} \overset{\text{indep.}}{\sim} \text{Poisson}(\mathcal{M}_{i,j,k})$$

$$\ell(\boldsymbol{\theta}) = \sum_{i,j,k} \left[\mathcal{X}_{i,j,k} * \log(\mathcal{M}_{i,j,k}) - \mathcal{M}_{i,j,k} \right] + \text{constant}$$

We present results for 3-way tensors, but our work generalizes to arbitrary D-way tensors

The parameter tensor is imposed a Canonical Polyadic tensor structure

Rank-1 tensor   **Rank summation**
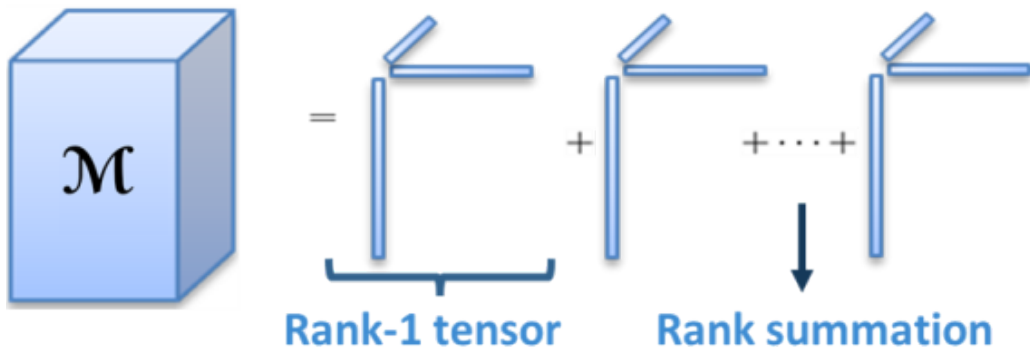
$$\boldsymbol{\theta} = [\text{vec}(A)'\,\text{vec}(B)'\,\text{vec}(C)']'$$

$$\mathcal{M}_{i,j,k} = \sum_{r=1}^{R} A_{i,r} B_{j,r} C_{k,r}$$

[3] Lee and Seung, *Algorithms for Non-negative Matrix Factorization*, NeurIPS 2000.
[4] Chi and Kolda, *On Tensors, Sparsity, and Nonnegative Factorizations*, SIMAX 2012.

PCP TENSOR MODEL: CHALLENGES AND APPROACHES

$$\ell(\boldsymbol{\theta}) = \sum_{i,j,k} \left[\boldsymbol{\mathcal{X}}_{i,j,k} * \log(\boldsymbol{\mathcal{M}}_{i,j,k}) - \boldsymbol{\mathcal{M}}_{i,j,k} \right] + \text{constant}$$

$$\boldsymbol{\mathcal{M}}_{i,j,k} = \sum_{r=1}^{R} A_{i,r} B_{j,r} C_{k,r}$$

Optimization Approach [3,4]

How to efficiently optimize the loglikelihood?
- MM optimization [3,4]

Their MM algorithms are actually EM algorithms!

- Higher order methods

Our Fisher Info can be used for Fisher scoring optimization!

Our Approach

PCP is a latent-variable model!
- EM algorithms
- Fisher information

Evaluate model and fit
- Well-posed statistical problems
- Evaluate convergence of algorithm

Probabilistic Approach [5]

How many entries do I need to recover $\boldsymbol{\mathcal{M}}$?
- Matrix/tensor completion
- Typically an upper bound on MSE

Our Fisher Info can be used for Cramer Rao bounds on MSE!

[3] Lee and Seung, *Algorithms for Non-negative Matrix Factorization*, NeurIPS 2000.
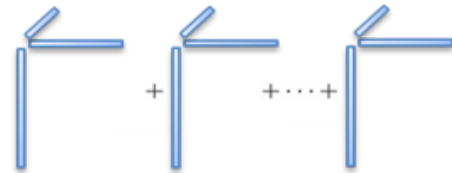[4] Chi and Kolda, *On Tensors, Sparsity, and Nonnegative Factorizations*, SIMAX 2012.
[5] Cao and Xie, *Poisson Matrix Recovery and Completion*, IEEE TSP 2016

FIRST MAIN RESULT: PCP IS A LATENT-VARIABLE MODEL

Random Variable

$$\mathcal{X} = \text{(cube)}$$

$$\mathcal{X} \sim \text{Poisson}\left(\sum_{r=1}^{R} A[:,r] \circ B[:,r] \circ C[:,r]\right)$$

Latent mechanism $\quad \mathcal{X} \overset{d}{=} \sum_{r=1}^{R} \mathcal{Z}_r$

Latent Random Variable

$$\mathcal{Z} = \left[\underset{\mathcal{Z}_1}{\text{(cube)}} \ \underset{\mathcal{Z}_2}{\text{(cube)}} \ \cdots \ \underset{\mathcal{Z}_R}{\text{(cube)}} \right]$$

$$\mathcal{Z}_r \sim \text{Poisson}\left(A[:,r] \circ B[:,r] \circ C[:,r]\right)$$
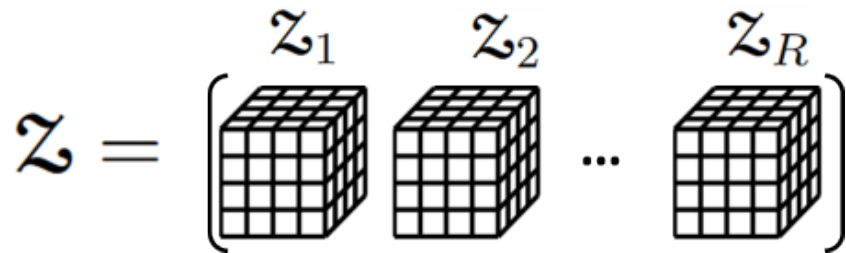
$\mathcal{X}$ follows a rank-R PCP model

$\updownarrow$

$\mathcal{X}$ is the sum of R independent $\mathcal{Z}_r$ each following a rank-1 PCP model

PCP AS A LATENT-VARIABLE MODEL

Observed Data $\boldsymbol{x} = \operatorname{vec}(\mathcal{X})$

$$\mathcal{X} = $$
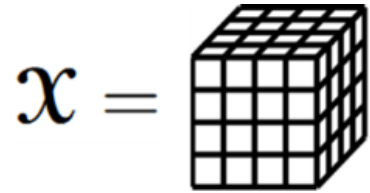
Complete Data $\boldsymbol{z} = \operatorname{vec}(\mathcal{Z})$

$$\mathcal{Z} = \begin{bmatrix} \mathcal{Z}_1 & \mathcal{Z}_2 & \cdots & \mathcal{Z}_R \end{bmatrix}$$

Latent mechanism $\quad \boldsymbol{x} \overset{d}{=} \sum_{r=1}^{R} \boldsymbol{z}_r$

$$\boldsymbol{\theta} = [\operatorname{vec}(A)' \operatorname{vec}(B)' \operatorname{vec}(C)']'$$

Loglikelihood $\qquad x_{i,j,k} \sim \operatorname{Poisson}\left(\sum_r A_{i,r} B_{j,r} C_{k,r}\right)$

$$\ell(\boldsymbol{x}|\boldsymbol{\theta}) := \log p(\boldsymbol{x}|\boldsymbol{\theta})$$
$$= \underbrace{\log p(\boldsymbol{z}|\boldsymbol{\theta})}_{\ell_c(\boldsymbol{z}|\boldsymbol{\theta})} - \underbrace{\log p(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})}_{\ell_m(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}$$

Complete loglikelihood
$$z_{r,i,j,k} \sim \operatorname{Poisson}(A_{i,r} B_{j,r} C_{k,r})$$

Missing loglikelihood
$$(z_{1,i,j,k} \ldots z_{R,i,j,k})|x_{i,j,k} \sim \operatorname{Multinomial}(n = x_{i,j,k}, p_1, \ldots, p_R)$$
$$p_r = \frac{A_{i,r} B_{j,r} C_{k,r}}{\sum_r A_{i,r} B_{j,r} C_{k,r}}$$

SECOND MAIN RESULT: EXISTING MM ALGORITHMS CAN BE DERIVED AS EM ALGORITHMS

E-step $Q(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) := \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}, \bar{\boldsymbol{\theta}}}(\ell_c(\boldsymbol{\theta}))$

$$= \sum_{r,i,j,k} \left[\bar{\mathcal{Z}}_{r,i,j,k} * \log(A_{i,r} B_{j,r} C_{k,r}) - (A_{i,r} B_{j,r} C_{k,r}) \right]$$

$$\bar{\mathcal{z}}_{r,i,j,k} := \mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}, \bar{\boldsymbol{\theta}}}(\mathcal{Z}_{r,i,j,k}) = \mathcal{X}_{i,j,k} \frac{\bar{A}_{i,r} \bar{B}_{j,r} \bar{C}_{k,r}}{\sum_r \bar{A}_{i,r} \bar{B}_{j,r} \bar{C}_{k,r}}$$

⟵ This is the mean of a multinomial

CM-step Here $\boldsymbol{\theta}$ is split into 3 blocks, corresponding to A, B and C.

Solve $\underset{A}{\mathrm{argmax}}\, Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \left\{ \sum_{j,k} \bar{\mathcal{z}}^{(t)}_{r,i,j,k} \right\}_{i,r} = A * \left([\mathcal{X}_{(1)} \oslash (A(C \odot B)')] (C \odot B) \right)$

This is EM algorithm!

This is the update used in [3] and [4]!!

[3] Lee and Seung, *Algorithms for Non-negative Matrix Factorization*, NeurIPS 2000.
[4] Chi and Kolda, *On Tensors, Sparsity, and Nonnegative Factorizations*, SIMAX 2012.

A PATH TOWARDS THE FISHER INFORMATION

Latent mechanism $\qquad x \overset{d}{=} \sum_{r=1}^{R} z_r$

The Missing Information Principle [6]

"observed information" equals the "complete information" minus the "missing information"

$$\mathcal{I}(\boldsymbol{\theta}; \boldsymbol{x}) := -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ell(\boldsymbol{x}|\boldsymbol{\theta})$$

$$= \underbrace{\mathbb{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ell_c(\boldsymbol{z}, \boldsymbol{x}|\boldsymbol{\theta})\Big|\boldsymbol{x}\right]}_{\mathcal{I}_c(\boldsymbol{\theta}; \boldsymbol{x})} - \underbrace{\mathbb{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ell_m(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})\Big|\boldsymbol{x}\right]}_{\mathcal{I}_m(\boldsymbol{\theta}; \boldsymbol{x})}$$

observed information	complete information	missing information
$x_{i,j,k}$	$z_{r,i,j,k}$	$(z_{1,i,j,k} \ldots z_{R,i,j,k})|x_{i,j,k}$
$\mathrm{Poisson}\left(\sum_r A_{i,r}B_{j,r}C_{k,r}\right)$	$\mathrm{Poisson}(A_{i,r}B_{j,r}C_{k,r})$	$\mathrm{Multinomial}(n = x_{i,j,k}, p_1, \ldots, p_R)$

[6] Orchard and Woodbury, *A missing information principle: theory and applications*, Berkeley Symp. on Math. Statist. and Prob., 1972

RANK 1 PCP CASE: FISHER INFORMATION MATRIX

When *R*=1, the complete data is observed $\mathcal{X} = \mathcal{Z}$. No information is lost.

Parameter Vector

$$\boldsymbol{\theta} = [\boldsymbol{a}'\boldsymbol{b}'\boldsymbol{c}']'$$

Model

$$\mathcal{X} \sim \text{Poisson}(\boldsymbol{a} \circ \boldsymbol{b} \circ \boldsymbol{c})$$

Loglikelihood

$$\ell(\boldsymbol{\theta}) = \sum_{i,j,k} [\mathcal{X}_{i,j,k} * \log(\boldsymbol{a}_i \boldsymbol{b}_j \boldsymbol{c}_k) - \boldsymbol{a}_i \boldsymbol{b}_j \boldsymbol{c}_k] + \text{constant}$$

Fisher Information

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{bmatrix} \text{diag}(\boldsymbol{a}^{-1}) & \mathbf{11}' & \mathbf{11}' \\ \mathbf{11}' & \lambda\,\text{diag}(\boldsymbol{b}^{-1}) & \lambda\mathbf{11}' \\ \mathbf{11}' & \lambda\mathbf{11}' & \lambda\,\text{diag}(\boldsymbol{c}^{-1}) \end{bmatrix}$$

- Above we are parameterizing $\boldsymbol{\theta}$ so that $\lambda = \boldsymbol{a}'\mathbf{1}$ and $1 = \boldsymbol{b}'\mathbf{1} = \boldsymbol{c}'\mathbf{1}$. Any parameterization follows from this
- For $\mathcal{X}$ of size $N_1 \times N_2 \times N_3$, $\mathcal{I}(\boldsymbol{\theta})$ is square with $N_1 + N_2 + N_3$ rows and columns
- $\mathcal{I}(\boldsymbol{\theta})$ is singular of rank $N_1 + N_2 + N_3 - 2$
- The FIM is nonsingular if you remove any one entry from $\boldsymbol{b}$ and any one entry from $\boldsymbol{c}$

GENERAL RANK PCP CASE, AND OAKES' THEOREM

- For the general rank case, we have missing information.
- Unlike Gaussian CP, the Fisher information for R=1 is not a special case of general rank
- Direct differentiation of the loglikelihood is challenging
- We can leverage the missing information principle

$$\mathcal{I}(\boldsymbol{\theta}; \boldsymbol{x}) = \mathcal{I}_c(\boldsymbol{\theta}; \boldsymbol{x}) - \mathcal{I}_m(\boldsymbol{\theta}; \boldsymbol{x})$$

Similar to
R=1 case

Challenging

Latent mechanism
$$\boldsymbol{x} \stackrel{d}{=} \sum_{r=1}^{R} \boldsymbol{z}_r$$

- Many techniques exist for obtaining/estimating $\mathcal{I}_m(\boldsymbol{\theta}; \boldsymbol{x})$ from the <u>complete loglikelihood</u>

- Most popular is Louis' method [7], but can only be evaluated at the MLE.

- I use Oakes' method [8], which is more general: $\mathcal{I}_m(\boldsymbol{\theta}; \boldsymbol{x}) = \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \bar{\boldsymbol{\theta}}'} Q(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}) \right]_{\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}}$ $\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}, \bar{\boldsymbol{\theta}}}(\ell_c(\boldsymbol{\theta}))$

[7] Louis, *Finding the Observed Information Matrix when Using the EM Algorithm*, JRSSB 1982
[8] Oakes, *Direct Calculation of the Information Matrix via the EM Algorithm*, JRSSB 1999

GENERAL RANK PCP: FISHER INFORMATION MATRIX

Model

$$\mathcal{X}_{a,b,c} \overset{\text{indep.}}{\sim} \text{Poisson}(\mathcal{M}_{a,b,c})$$

Parameter Vector

$$\boldsymbol{\theta} = [\text{vec}(A)'\text{vec}(B)'\text{vec}(C)']'$$

Loglikelihood

$$\ell(\boldsymbol{\theta}) = \sum_{a,b,c} [\mathcal{X}_{a,b,c} * \log(\mathcal{M}_{a,b,c}) - \mathcal{M}_{a,b,c}]$$

$$\mathcal{M}_{a,b,c} = \sum_{r=1}^{R} A[a,r]B[b,r]C[c,r]$$

Fisher Information

$$\mathcal{I}(\boldsymbol{\theta}) = \left\{ \left\{ G_{k,l}^{r,s} \right\}_{r,s=1,\ldots,R} \right\}_{k,l=1,2,3}$$

$$G_{k,l}^{r,s} = \begin{cases} \text{diagonal matrix} & k = l \\ \text{dense matrix} & k \neq l \end{cases}$$

- $\mathcal{I}(\boldsymbol{\theta})$ is a 3 x 3 block matrix, where each block is itself a R x R block matrix
- Above is for arbitrary parameterization of $\boldsymbol{\theta}$
- For $\mathcal{X}$ of size $N_1 \times N_2 \times N_3$, $\mathcal{I}(\boldsymbol{\theta})$ is square with $R(N_1 + N_2 + N_3)$ rows and columns
- $\mathcal{I}(\boldsymbol{\theta})$ is singular of rank $R(N_1 + N_2 + N_3 - 2)$
- The FIM is nonsingular if you remove one entry from each column of B, and one entry from each column of C

BIAS-VARIANCE TRADE-OFF AND THE CRAMER-RAO LOWER BOUND

Consider a true parameter vector
$$\boldsymbol{\theta}_0 = [\, \text{vec}(A)^\top \; \text{vec}(B)^\top \; \text{vec}(C^\top)]^\top$$
and its estimated counterpart $\hat{\boldsymbol{\theta}}$

Call

1. Bias² $||\mathbb{E}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\theta}_0||^2$

2. Variance: $\mathbb{E}||\hat{\boldsymbol{\theta}} - \mathbb{E}(\hat{\boldsymbol{\theta}})||^2$

3. MSE: $\mathbb{E}||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0||^2$

4. CRLB: $\text{tr}(\mathcal{I}^\dagger(\boldsymbol{\theta}_0))$

Then:

- MSE = Bias²+*Variance*
- Bias² $= 0 \implies$ CRLB $\leq$ Variance

Monte-Carlo Study

- Draw $\mathcal{X}_k^*$ from $\text{Poisson}(\mathcal{M}(\boldsymbol{\theta}_0))$
- Estimate $\hat{\boldsymbol{\theta}}_k^*$ from $\mathcal{X}_k^*$.
- Repeat for some large K
  $k = 1, 2, \ldots, K$

Monte-Carlo approximations:

1. Mean: $\bar{\boldsymbol{\theta}}_{MC} = K^{-1}\sum_k \hat{\boldsymbol{\theta}}_k^*$

2. Bias2: $||\bar{\boldsymbol{\theta}}_{MC} - \boldsymbol{\theta}_0||^2$

3. Variance: $K^{-1}\sum_k ||\hat{\boldsymbol{\theta}}_k^* - \bar{\boldsymbol{\theta}}_{MC}||^2$

4. MSE: $K^{-1}\sum_k ||\hat{\boldsymbol{\theta}}_k^* - \boldsymbol{\theta}_0||^2$

BIAS-VARIANCE TRADE-OFF AND THE CRAMER-RAO LOWER BOUND

Simulation Setup

Visualize the bias-variance trade-off for tensors $\mathcal{M}(\boldsymbol{\theta}_o)$ with varying:

- Entry-wise mean $s = 1, 2, 3, 4$
  $s = \text{mean}(\mathcal{M})$

- Sizes $N = 10, 15, 20$
  $\mathcal{M} \in \mathbb{R}^{N \times N \times N}$

- Rank $R = 1, 2, \ldots, 16$

Cramer-Rao Lower Bound

$$\left(\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}(\hat{\boldsymbol{\theta}}) \right) \mathcal{I}^{\dagger}(\boldsymbol{\theta}) \left(\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}(\hat{\boldsymbol{\theta}}) \right)^{\top} \leq Var(\hat{\boldsymbol{\theta}})$$

Jacobian identity

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}(\hat{\boldsymbol{\theta}}) = \text{Cov}\left(\hat{\boldsymbol{\theta}} \,, \, \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) \right)$$

Score simplification

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{x}; \boldsymbol{\theta}) = \begin{bmatrix} \text{vec}[\mathcal{S}_{(1)}(C \odot B)] \\ \text{vec}[\mathcal{S}_{(2)}(C \odot A)] \\ \text{vec}[\mathcal{S}_{(3)}(B \odot A)] \end{bmatrix}$$

$$\mathcal{S}_{\boldsymbol{i}} = \mathcal{X}_{\boldsymbol{i}} / \mathcal{M}_{\boldsymbol{i}} - 1$$
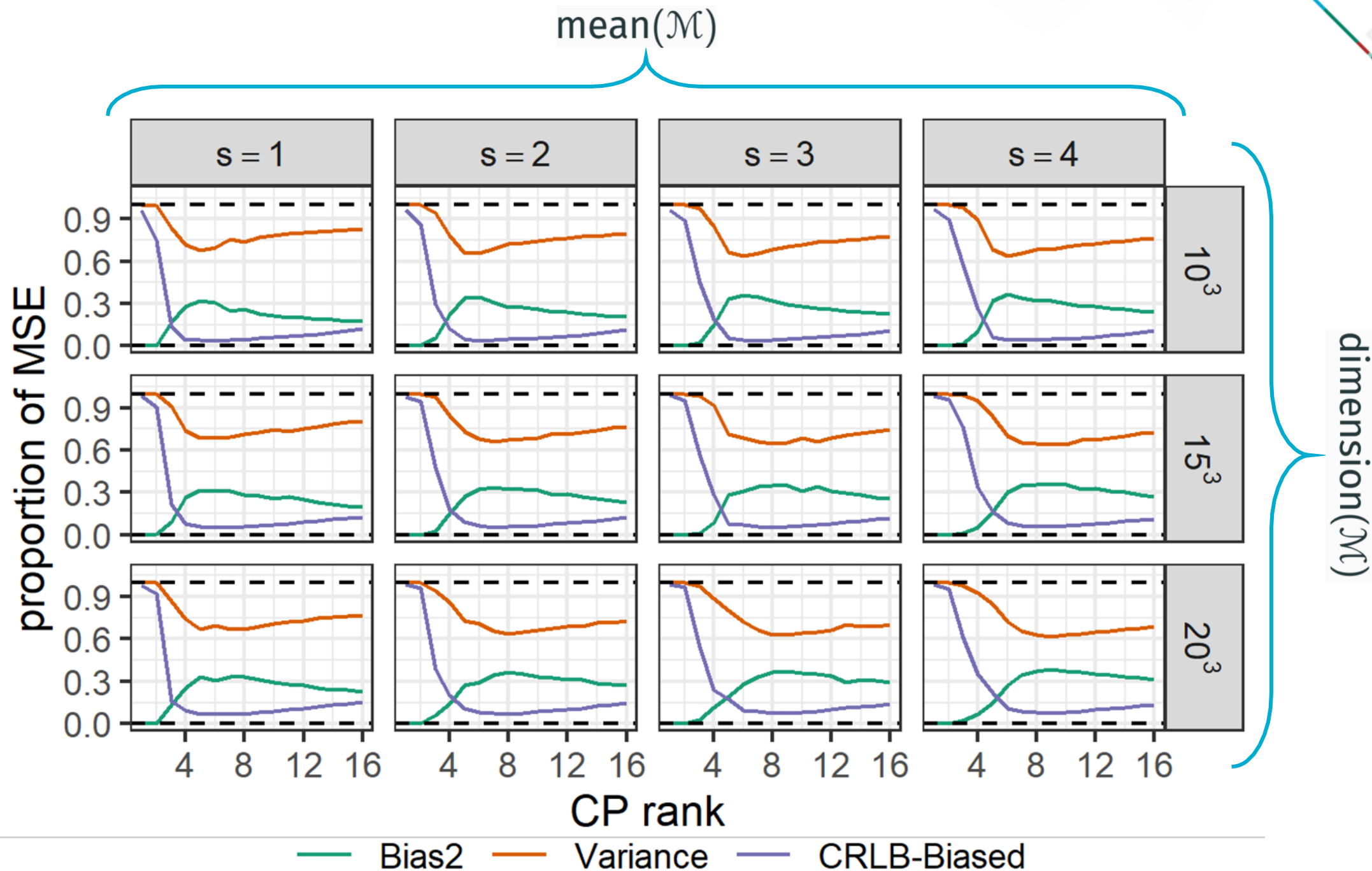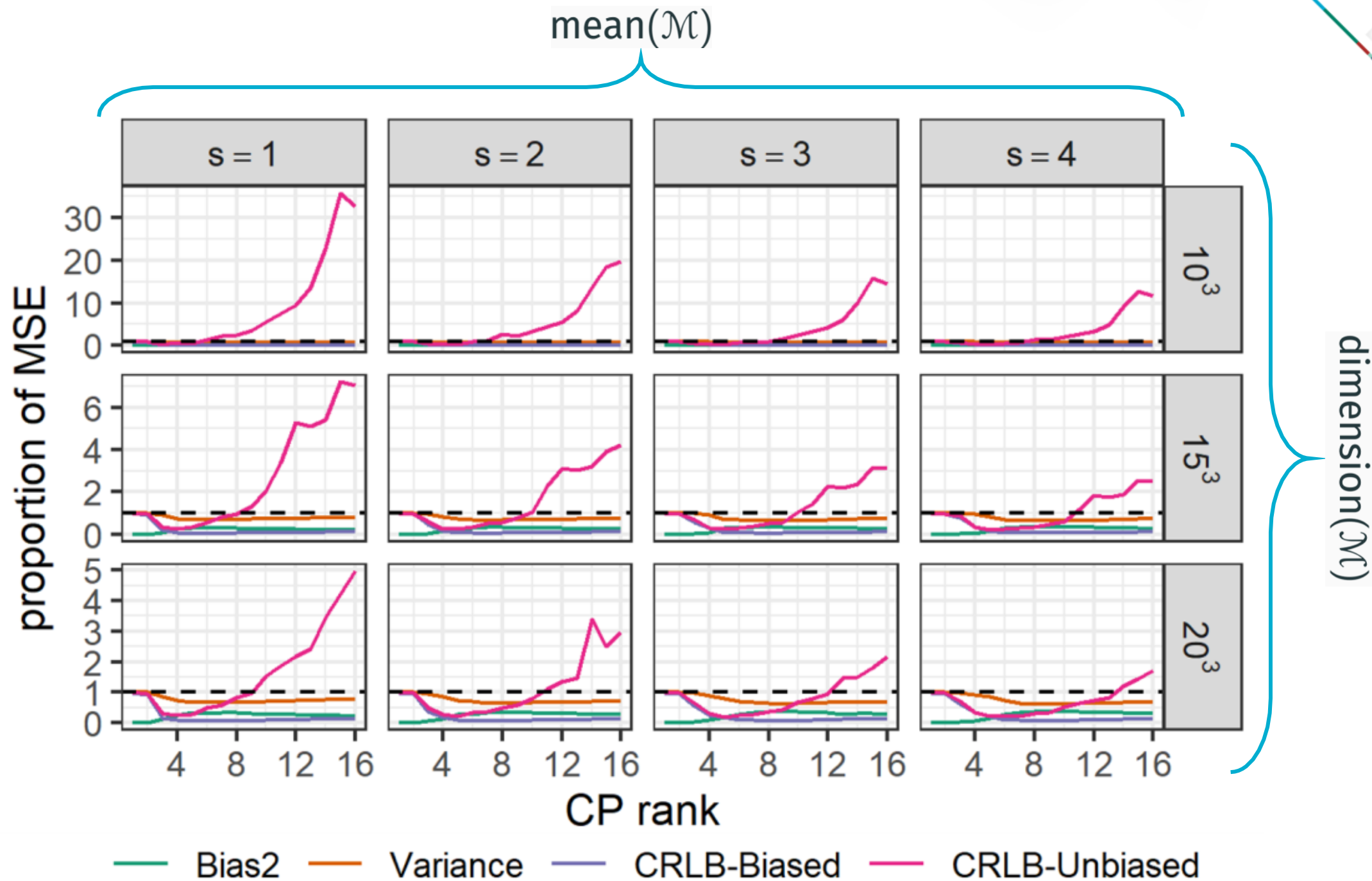
CONCLUSIONS AND PATH FORWARD

- PCP is a very popular method, we demonstrate it's a latent variable model
  - Applications in topic modeling, document clustering and classification, poll analysis, etc.
  - We allow for parameter inference through complete loglikelihood

- We rediscover popular estimating algorithms as instances of EM algorithms
  - Shed light on the properties of existing algorithms
  - Help bridge two fields of machine learning research

- Derived novel Fisher information matrix, using the missing information principle
  - Can be used to propose new Fisher scoring algorithms, Cramer Rao inequalities
  - Allows us to gauge the conditions for a well-posed parameter inference problem

- Variance-trade off simulation study
  - Bias-variance trade-off of PCP
  - Comparison against CRLB

THANK YOU!

The Poisson Canonical Polyadic Tensor Model as a Latent-Variable Model

Carlos Llosa cjllosa@sandia.gov