Generalization of Tensor Factorization and Applications

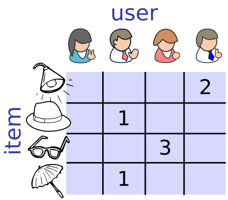
Kohei Hayashi

Collaborators:
T. Takenouchi, T. Shibata, Y. Kamiya, D. Kato, K. Kunieda, K. Yamada, K. Ikeda, R. Tomioka, H. Kashima

May 14, 2012

Relational data

is a collection of relationships among multiple objects.

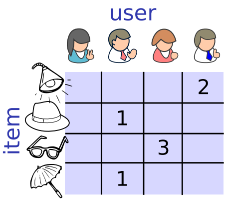


matrix

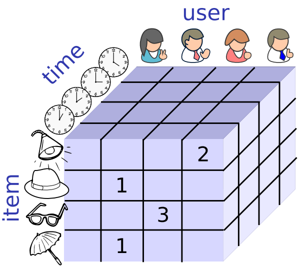
relationships of pair $\Leftrightarrow$ matrix

Relational data

is a collection of relationships among multiple objects.



matrix

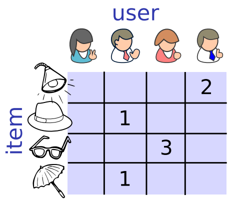


3 order tensor

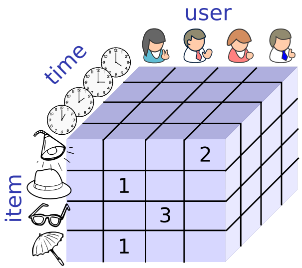
relationships of pair	$\Leftrightarrow$	matrix
relationships of 3-tuple	$\Leftrightarrow$	3 dim. array

Relational data

is a collection of relationships among multiple objects.



matrix



3 order tensor

$$\left.\begin{array}{ccc} \text{relationships of pair} & \Leftrightarrow & \text{matrix} \\ \text{relationships of 3-tuple} & \Leftrightarrow & \text{3 dim. array} \\ \vdots & & \vdots \end{array}\right\} \text{tensor}$$

can represent as a tensor with missing values.

Issue of tensor representation

Tensor representation is generally high-dimensional and large-scale

- e.g. $1,000$ users $\times$ $1,000$ items $\times$ $1,000$ times $=$ total $1,000,000,000$ relationships

Dimensional reduction techniques such as tensor factorization is used.

Tensor factorization

Tucker decomposition [Tucker 1966]: a tensor factorization method assuming

1. observation noise is Gaussian
2. underlying tensor is low-dimensional

These assumptions are general ... but are not always true.

Contributions

Generalize Tucker decomposition and propose two new models:

1. Exponential family tensor factorization (ETF)
   [Joint work with Takenouchi, Shibata, Kamiya, Kunieda, Yamada, and Ikeda]
   - Generalize the noise distribution.
   - Can handle a tensor containing mixed discrete and continuous values.
2. Full-rank tensor completion (FTC)
   [Joint work with Tomioka and Kashima]
   - Kernelize Tucker decomposition
   - Complete missing values without reducing the dimensionality.

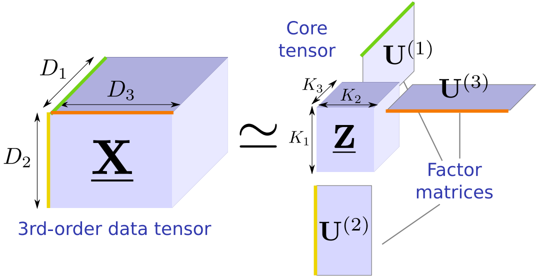
Outline

1. Tucker decomposition
2. Exponential family tensor factorization (ETF)
3. Full-rank tensor completion (FTC)
4. Conclusion

Outline

1. Tucker decomposition
2. Exponential family tensor factorization (ETF)
3. Full-rank tensor completion (FTC)
4. Conclusion

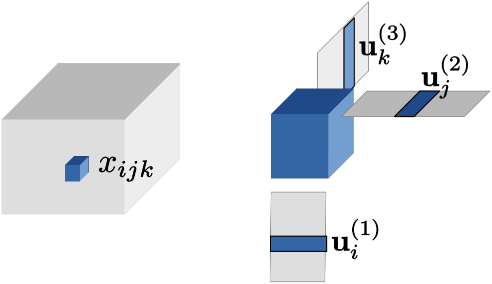
Tucker decomposition



$$x_{ijk} = \sum_{q=1}^{K_1} \sum_{r=1}^{K_2} \sum_{s=1}^{K_3} u_{iq}^{(1)} u_{jr}^{(2)} u_{ks}^{(3)} z_{qrs} + \varepsilon_{ijk} \tag{1}$$

- ε: i.i.d Gaussian noise

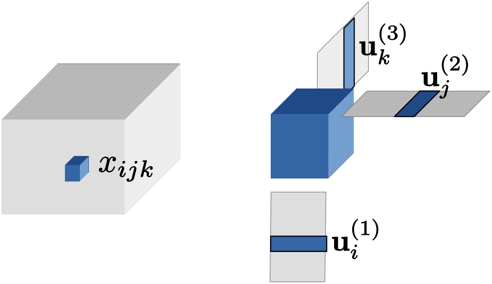
Tucker decomposition



$$x_{ijk} = \sum_{q=1}^{K_1} \sum_{r=1}^{K_2} \sum_{s=1}^{K_3} u_{iq}^{(1)} u_{jr}^{(2)} u_{ks}^{(3)} z_{qrs} + \varepsilon_{ijk} \tag{1}$$

- ε: i.i.d Gaussian noise

Tucker decomposition



$$x_{ijk} = \sum_{q=1}^{K_1} \sum_{r=1}^{K_2} \sum_{s=1}^{K_3} \underline{u_{iq}^{(1)} u_{jr}^{(2)} u_{ks}^{(3)}} \, \underline{z_{qrs}} + \varepsilon_{ijk} \tag{1}$$

- ε: i.i.d Gaussian noise

Vectorized form

Let $\vec{\mathbf{x}} \in \mathbb{R}^D$ and $\vec{\mathbf{z}} \in \mathbb{R}^K$ denote vectorized $\underline{\mathbf{X}}$ and $\underline{\mathbf{Z}}$, resp., then

$$\vec{\mathbf{x}} = \mathbf{W}\vec{\mathbf{z}} + \vec{\varepsilon} \quad \text{where} \quad \mathbf{W} \equiv \mathbf{U}^{(3)} \otimes \mathbf{U}^{(2)} \otimes \mathbf{U}^{(1)}.$$

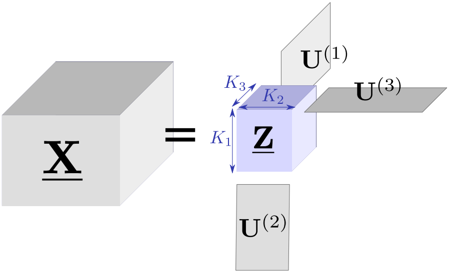
- $D \equiv D_1 D_2 D_3$, $K \equiv K_1 K_2 K_3$
- $\otimes$: the Kronecker product

Tucker decomposition = A linear Gaussian model

$$\vec{\mathbf{x}} \sim N(\vec{\mathbf{x}} \mid \mathbf{W}\vec{\mathbf{z}}, \sigma^2 \mathbf{I})$$

Rank of tensor

Call dimensionalities of the core tensor *rank of tensor*.



- Rank of $\underline{\mathbf{X}}$ is (K_1, K_2, K_3).

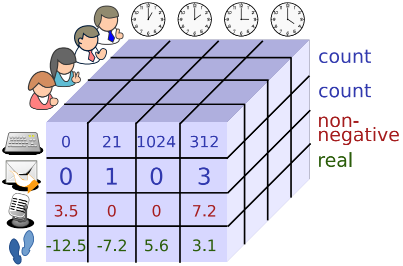
Rank of tensor represents its complexity

- low-rank tensor has less information

Outline

1. Tucker decomposition
2. Exponential family tensor factorization (ETF)
3. Full-rank tensor completion (FTC)
4. Conclusion

Motivation: heterogeneous tensor

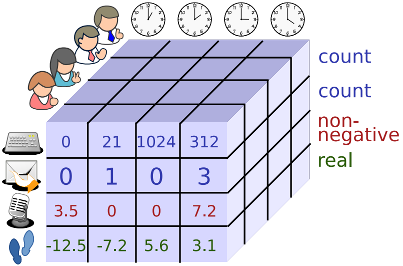


Multisensor measurements
- different data types for each sensor

Tucker decomposition assumes Gaussian noise

- not appropriate for such data

Motivation: heterogeneous tensor



Multisensor measurements
- different data types for each sensor

Tucker decomposition assumes Gaussian noise

- not appropriate for such data

Approach

generalize Tucker decomposition, assuming a different distribution for each element

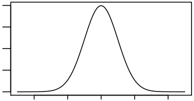
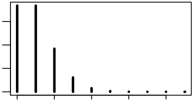
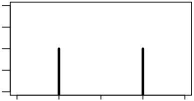
Exponential family tensor factorization

Likelihood

$$\vec{\mathbf{x}} \sim \prod_{d=1}^{D} \underbrace{\mathrm{Expon}_d(\vec{x}_d \mid \vec{\theta}_d)}_{\text{exp. family}}, \qquad \underbrace{\vec{\boldsymbol{\theta}} \equiv \mathbf{W}\vec{\mathbf{z}}}_{\text{Tucker decomp.}}$$

where $\quad \mathrm{Expon}(x \mid \theta) \equiv \exp\left[x\theta - \psi(\theta) + F(x)\right]$

exponential family: a class of distributions

	Gaussian	Poisson	Bernoulli
Density ($\theta = 0$)			
$\psi(\theta)$	$\theta^2/2$	$\exp[\theta]$	$\ln(1+\exp[\theta])$

Priors

- for $\vec{\mathbf{z}}$: a Gaussian prior $N(\mathbf{0}, \mathbf{I})$
- for $\mathbf{U}^{(m)}$: a Gaussian prior $N(\mathbf{0}, \alpha_m^{-1}\mathbf{I})$

Joint log-likelihood

$$\begin{aligned}\mathcal{L} = &\vec{\mathbf{x}}^\top \vec{\mathbf{W}} \vec{\mathbf{z}} - \sum_{d=1}^{D} \psi_{h_d}(\mathbf{w}_d^\top \vec{\mathbf{z}}) \\ &- \frac{1}{2} \left|\left|\vec{\mathbf{z}}\right|\right|^2 - \sum_{m=1}^{M} \frac{\alpha_m}{2} \left|\left|\mathbf{U}^{(m)}\right|\right|_{\mathrm{Fro}}^2 + \mathrm{const.}\end{aligned}$$

Estimate parameters by Bayesian inference.

Bayesian inference

Marginal-MAP estimator:

$$\underset{\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(M)}}{\operatorname{argmax}} \int \exp[\mathcal{L}(\vec{\mathbf{z}}, \mathbf{U}^{(1)}, \ldots, \mathbf{U}^{(M)})] \mathrm{d}\vec{\mathbf{z}}$$

- The integral is not analytical.

Develop efficient yet accurate approximation with *Laplace approximation* and *Gaussian process*.

- Computational cost is still higher than Tucker dcomp.
- For a long and thin tensor (e.g. time series), online algorithm is applicable (see thesis.)

Experiments

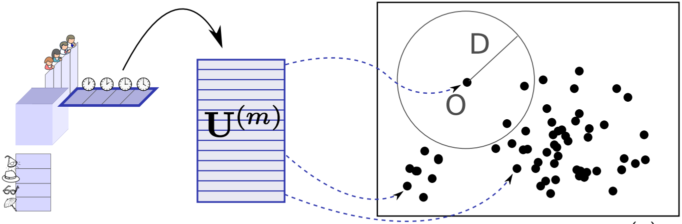
Anomaly detection by ETF

Purpose Find irregular parts of tensor data

Method Apply *distance-based outlier* $DB(p, D)$ [Knorr+ VLDB'00] to the estimated factor matrix $\mathbf{U}^{(m)}$

Definition of $DB(p, D)$

"An object O is a $DB(p, D)$ outlier if at least fraction p of the objects lies at a distance greater than D from O."



feature space of $\mathbf{U}^{(m)}$

Data set

A time series of multisensor measurements

- Each sensor recorded the human behavior (e.g. position) of researchers in NEC lab. for 8 month.
  - 6 (sensors) $\times$ 21 (persons) $\times$ 1927 (hours)

	Sensor	Type	Min	Max
$\mathbf{X}_1$	# of sent emails	Count	0.00	14.00
$\mathbf{X}_2$	# of received emails	Count	0.00	15.00
$\mathbf{X}_3$	# of typed keys	Count	0.00	50422.00
$\mathbf{X}_4$	X coordinate	Real	-550.17	3444.15
$\mathbf{X}_5$	Y coordinate	Real	128.71	2353.55
$\mathbf{X}_6$	Movement distance	Non-negative	0.00	203136.96

Evaluation

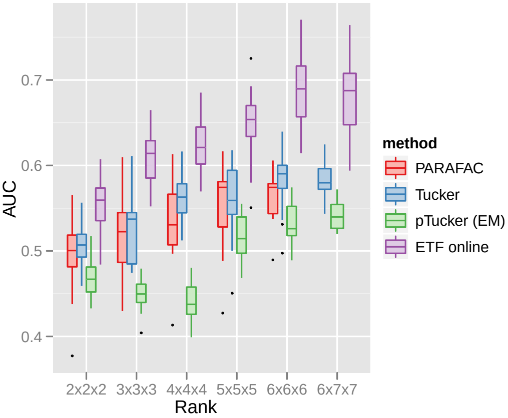
List of irregular events are provided.

Examples of irregular events

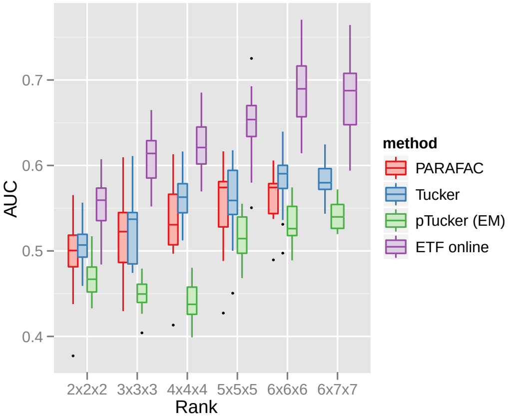
Date	Time	Description
Dec 21	All day	Private incident
Dec 22	15:00 16:00	Monthly seminar
Jun 15	13:00	Visiting tour
⋮	⋮	⋮

- Evaluate as a binary classification
- 50% are missing, 10 trials
- Apply ETF with online algorithm

Result: classification performance



Result: classification performance



✓ ETF well detect anomaly events

Skip the rest