AAAI-19
#2614

# Tensorial Change Analysis using Probabilistic Tensor Regression

Tsuyoshi Idé (“Ide-san”, 井手 剛, tide@us.ibm.com )
IBM Research, T. J. Watson Research Center

## Problem setting

**Learning regression function between tensor (multi-way array) and scalar**

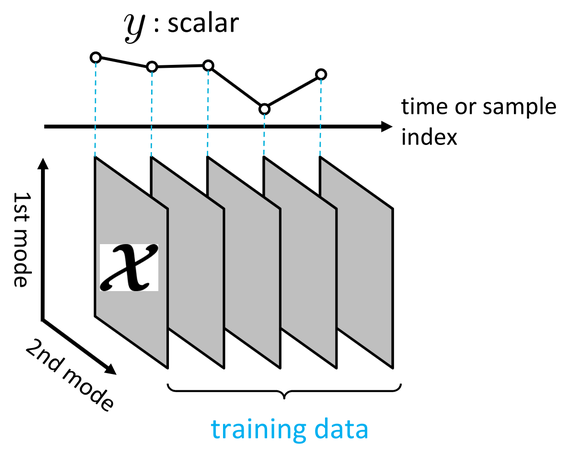
- Tensor regression
  - input: tensor
  - output: scalar

$$y = f(\boldsymbol{\mathcal{X}})$$

- Training data:

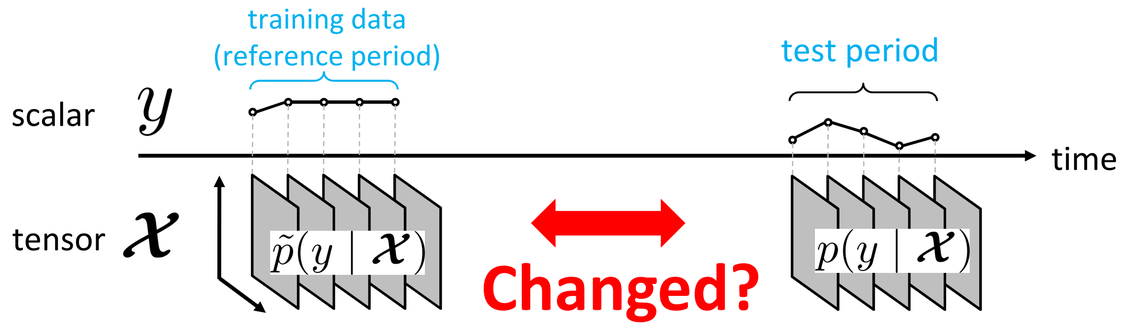
$$\{(y^{(1)}, \boldsymbol{\mathcal{X}}^{(1)}), \ldots, (y^{(N)}, \boldsymbol{\mathcal{X}}^{(N)})\}$$

1st sample  *N*-th sample




**The X-y relationship is used for change analysis (or diagnosis)**

**Goal: Learn probabilistic tensor regression model to compute change scores:**
**(1) overall change score**
**(2) mode- and dimension-wise change scores (responsibility scores)**




## Prior work

### Alternating least squares (ALS; [Zhou+ 13])
### → Non-probabilistic

- Linear model with CP (canonical polyadic) expansion

$$y \sim (\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}) \equiv \sum_{i_1,\ldots,i_M} \mathcal{X}_{i_1,\ldots,i_M} \mathcal{A}_{i_1,\ldots,i_M}$$

$$\boldsymbol{\mathcal{A}} = \sum_{r=1}^{R} \boldsymbol{a}^{1,r} \circ \boldsymbol{a}^{2,r} \circ \cdots \circ \boldsymbol{a}^{M,r}, \quad \text{or} \quad \mathcal{A}_{i_1,i_2,\ldots,i_M} = \sum_{r=1}^{R} a_{i_1}^{1,r} a_{i_2}^{2,r} \cdots a_{i_M}^{M,r}$$

- Factor out one component using tensor algebra

$$(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}) = \mathrm{Tr}\left[(\mathbf{A}^l)^\top \mathbf{X}_{(l)} (\mathbf{A}^M \odot \cdots \mathbf{A}^{l+1} \odot \mathbf{A}^{l-1} \odot \cdots \mathbf{A}^1)\right]$$

$$\mathbf{A}^l \equiv [\boldsymbol{a}^{l,1}, \ldots, \boldsymbol{a}^{l,R}], \ \odot = \text{Khatri-Rao product}$$

$$\mathbf{X}_{(l)} \equiv \text{mode-}l \text{ matricization of } \boldsymbol{\mathcal{X}},$$

- Find it, given the others. Repeat.

### Gaussian process regression [Zhao+ 14].
### → Much computational cost on testing

- Use standard GPR formula (**K**: kernel matrix)

$$(\text{predictive mean}) = \boldsymbol{k}(\boldsymbol{\mathcal{X}})^\top [\mathbf{K} + \sigma^2 \mathbf{I}_N]^{-1} \boldsymbol{y}_N$$

$$(\text{predictive variance}) = k_0 - \boldsymbol{k}(\boldsymbol{\mathcal{X}})^\top [\mathbf{K} + \sigma^2 \mathbf{I}_N]^{-1} \boldsymbol{k}(\boldsymbol{\mathcal{X}})$$

$$[\mathbf{K}]_{n,n'} = k(\boldsymbol{\mathcal{X}}^{(n)}, \boldsymbol{\mathcal{X}}^{(n')})$$

- Define proper kernel function for tensors
  - Note: You can **NOT** use the Frobenius norm as it reduces to the simple vectorization approach

$$\|\boldsymbol{\mathcal{X}}^{(n)} - \boldsymbol{\mathcal{X}}^{(n')}\|_{\mathrm{F}}^2 = \|\mathrm{vec}(\boldsymbol{\mathcal{X}}^{(n)}) - \mathrm{vec}(\boldsymbol{\mathcal{X}}^{(n')})\|_2^2$$

### Bayesian method [Guhaniyogi+ 17].
### → Needs complex Monte Carlo sampling

- Observation model with CP expanded coefficients
  - $p(y \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, \lambda) = \mathcal{N}(y \mid (\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}), \lambda^{-1})$
- Prior distributions allowing sparse model by shrinkage

$$p(\{\boldsymbol{a}^{l,r}\}) = \prod_{l=1}^{M} \prod_{r=1}^{R} \mathcal{N}(\boldsymbol{a}^{l,r} \mid \boldsymbol{0}, (\phi^r \tau)\mathrm{diag}(\boldsymbol{w}^{l,r})), \quad p(\lambda) = \mathcal{G}(\lambda \mid a_\lambda, b_\lambda)$$

$$p(\boldsymbol{w}^{l,r}) = \prod_{i=1}^{d_l} \mathrm{Exp}(w_i^{l,r} \mid \nu^{l,r}), \quad p(\sqrt{\nu^{l,r}}) = \mathcal{G}(\nu^{l,r} \mid a_\nu, b_\nu)$$

$$p(\boldsymbol{\phi}) = \mathrm{Dir}(\boldsymbol{\phi} \mid \alpha), \quad p(\tau) = \mathcal{G}(\tau \mid a_\tau, b_\tau)$$

- Inference is done by Markov chain Monte Carlo

## Probabilistic tensor regression

### Proposed Bayesian model

- Observation model: Gaussian + CP expansion

$$p(y \mid \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}, \lambda) = \mathcal{N}(y \mid (\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}), \lambda^{-1})$$ Gaussian distribution

$$\boldsymbol{\mathcal{A}} = \sum_{r=1}^{R} \boldsymbol{a}^{1,r} \circ \boldsymbol{a}^{2,r} \circ \cdots \circ \boldsymbol{a}^{M,r} \quad (\circ = \text{direct product})$$

- Transform to linear model by factoring out each component given the others

$$(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{A}}) = \sum_{r=1}^{R} (\boldsymbol{\mathcal{X}}, \boldsymbol{a}^{1,r} \circ \cdots \circ \boldsymbol{a}^{M,r}) = \sum_{r=1}^{R} (\boldsymbol{a}^{l,r})^\top \boldsymbol{\phi}^{l,r}(\boldsymbol{\mathcal{X}}),$$

$$\boldsymbol{\phi}^{l,r} = \mathbf{X}_{(l)} (\boldsymbol{a}^{M,r} \otimes \cdots \otimes \boldsymbol{a}^{l+1,r} \otimes \boldsymbol{a}^{l-1,r} \otimes \cdots \otimes \boldsymbol{a}^{1,r}),$$

“Mode-*l* matricization” $\otimes = (\text{Kronecker product})$

- Prior distribution: Gauss-Gamma

$$p(\boldsymbol{a}^{l,r} \mid b^{l,r}) = \mathcal{N}(\boldsymbol{a}^{l,r} \mid \boldsymbol{0}, (b^{l,r})^{-1} \mathbf{I}_{d_l}),$$

$$p(b^{l,r} \mid \alpha_0, \beta_0) = \frac{{\beta_0}^{\alpha_0}}{\Gamma(\alpha_0)} (b^{l,r})^{\alpha_0 - 1} \mathrm{e}^{-\beta_0 b^{l,r}}$$ gamma distribution

### Variational EM framework for inference

- Log complete likelihood

$$L(\boldsymbol{\mathcal{A}}, \lambda, \{b^{l,r}\}) \equiv \ln \underbrace{\prod_{n=1}^{N} p(y^{(n)} \mid \boldsymbol{\mathcal{X}}^{(n)}, \boldsymbol{\mathcal{A}}, \lambda)}_{\text{likelihood}} \times \underbrace{\prod_{l=1}^{M} \prod_{r=1}^{R} p(\boldsymbol{a}^{l,r} \mid b^{l,r}) p(b^{l,r} \mid \alpha_0, \beta_0)}_{\text{priors}}$$

- Assumed factorized form for posterior

$$Q(\{\boldsymbol{a}^{l,r}\}) = \prod_{r=1}^{R} q^{1,r}(\boldsymbol{a}^{1,r}) q^{2,r}(\boldsymbol{a}^{2,r}) \cdots q^{M,r}(\boldsymbol{a}^{M,r})$$

- Iterate between VB for $\{\boldsymbol{a}^{l,r}\}$ and $\{b^{l,r}\}$ and point-estimate of $\lambda$
- Results:
  - Posterior $q_1^{l,r}(\boldsymbol{a}^{l,r}) = \mathcal{N}(\boldsymbol{a}^{l,r} \mid \boldsymbol{\mu}^{l,r}, \boldsymbol{\Sigma}^{l,r}), \quad q_2^{l,r}(b^{l,r}) = \mathcal{G}(b^{l,r} \mid \alpha^{l,r}, \beta^{l,r})$
  - Precision $\lambda^{-1} = \frac{1}{N} \sum_{n=1}^{N} \left\{ y^{(n)} - \sum_{r=1}^{R} (\boldsymbol{\mathcal{X}}^{(n)}, \boldsymbol{\mu}^{1,r} \circ \cdots \circ \boldsymbol{\mu}^{M,r}) \right\}^2 + (\text{higher order term})$
- Predictive distribution $p(y \mid \boldsymbol{x}, \mathrm{D}) = \mathcal{N}(y \mid \mu(\boldsymbol{\mathcal{X}}), \sigma^2(\boldsymbol{\mathcal{X}}))$

$$\mu(\boldsymbol{\mathcal{X}}) = \eta + \sum_{r=1}^{R} (\boldsymbol{\mathcal{X}}, \boldsymbol{\mu}^{1,r} \circ \cdots \circ \boldsymbol{\mu}^{M,r}), \quad \sigma^2(\boldsymbol{\mathcal{X}}) = \lambda^{-1} + \sum_{r=1}^{R} \sum_{l=1}^{M} (\boldsymbol{\varphi}^{l,r})^\top \boldsymbol{\Sigma}^{l,r} \boldsymbol{\varphi}^{l,r}$$

$$\boldsymbol{\varphi}^{l,r} = \mathbf{X}_{(l)} (\boldsymbol{\mu}^{M,r} \otimes \cdots \otimes \boldsymbol{\mu}^{l+1,r} \otimes \boldsymbol{\mu}^{l-1,r} \otimes \cdots \otimes \boldsymbol{\mu}^{1,r}).$$

## Change scores

### Overall change: Use predictive distribution

- Outliner score for a new sample

$$c(y, \boldsymbol{\mathcal{X}}) \equiv -\ln p(y \mid \boldsymbol{\mathcal{X}}, \mathrm{D}) = \frac{\{y - \mu(\boldsymbol{\mathcal{X}})\}^2}{2\sigma^2(\boldsymbol{\mathcal{X}})} + \frac{1}{2} \ln\{2\pi\sigma^2(\boldsymbol{\mathcal{X}})\}$$

- Change-point score: smoothed version of it

### Mode- and dimension-wise change: Use posterior learned

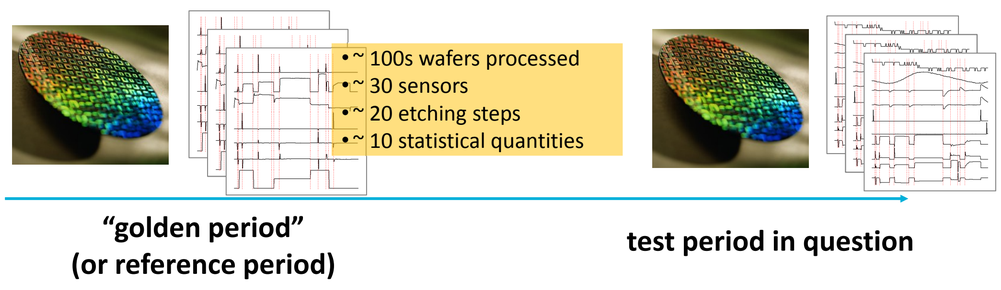
- Change analysis score of the *i*-th dimension of the *l*-th mode

$$c_i^l(\tilde{\mathrm{D}}, \mathrm{D}) \equiv \frac{1}{R} \sum_{r=1}^{R} \int \mathrm{d}\boldsymbol{a}^{l,r} q^{l,r}(\boldsymbol{a}^{l,r}) \ln \frac{q^{l,r}(a_i^{l,r} \mid \boldsymbol{a}_{-i}^{l,r})}{\tilde{q}^{l,r}(a_i^{l,r} \mid \boldsymbol{a}_{-i}^{l,r})}$$

- VB plays the key role in change analysis: It is the factorized assumption that makes possible mode- and dimension-wise scoring

## Application

### Change diagnosis for semiconductor etching tool




### Excursion event example detected

- Major change in ‘x12’ and ‘step 3’
- Found unusual fluctuation
  - Has been overlooked by existing monitoring tool