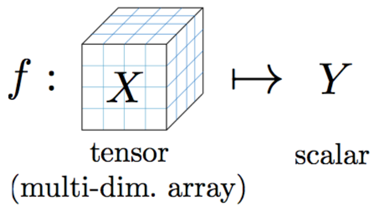
Doubly Decomposing Nonparametric Tensor Regression (ICML 2016)

M.Imaizumi (Univ. of Tokyo / JSPS DC)
K.Hayashi (AIST / JST ERATO)

2016/08/10

Outline

- **Topic**
  - Nonparametric Regression with Tensor input
    - Model
      $$Y = f(X) + \epsilon$$
    - Estimate (nonparametric) f

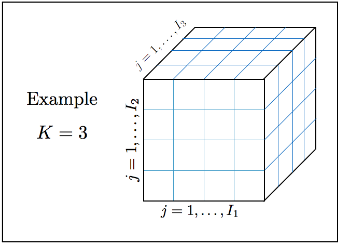


Outline

- **Method**
  - Propose a nonparametric model with a Bayes estimator
  - Improve its performance by controlling bias and variance trade-off

Tensor Regression Problem

- **Tensor data**
  - $X \in \mathbb{R}^{I_1 \times \ldots \times I_K}$
  - K : mode of tensor X
  - I_k : dim of k-th mode



- **Tensor Regression**
  - n observations $D_n = \{(X_i, Y_i)\}_{i=1}^n$
  - Input (tensor) : $X_i \in \mathbb{R}^{I_1 \times \ldots \times I_K}$ Output (scalar) : $Y_i \in \mathbb{R}$
  - D_n is generated with a function $f : \mathbb{R}^{I_1 \times \ldots \times I_K} \to \mathbb{R}$ as

$$Y_i = f(X_i) + \epsilon_i$$

    for $i = 1, \ldots, n$
  - ϵ_i is a Gaussian noise

Application of Tensor Regression Problem

- Predict health conditions from medical 3D images
  - X_i : medical 3D image of patient i, Y_i : health condition of i

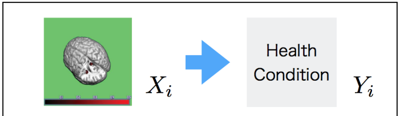


Image by Zhou+ (2013)

- Predict spread of epidemics on networks
  - X_i : adjacency matrix of network i, Y_i : $\#$ of infected nodes

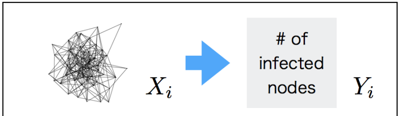


Image by Brenton+ (2011)

Related researches

- **Tensor linear regression**

$$Y = \langle W, X \rangle + \epsilon$$

  - $W \in \mathbb{R}^{I_1 \times \ldots \times I_K}$ is a parameter tensor
  - Dyrholm et al. (2007); Zhou et al. (2013); Suzuki (2015); Guhaniyogi et al. (2015),etc...

- **Nonparametric tensor regression**

$$Y = f(X) + \epsilon$$

  - $f : X \mapsto Y$ is possibly nonlinear function
  - Zhao et al. (2014); Hou et al. (2015),etc...

Motivation

- **Interest** : Convergence of an estimator $\hat{f}_n$

$$E\|\hat{f}_n - f^*\|^2 = O\left(n^{-?}\right)$$

  - n : # of observations
  - $\hat{f}_n$: estimator
  - f^* : target
- Take the nonparametric approach to reduce bias

Motivation

- **Starting point** : the (naive) nonparametric approach

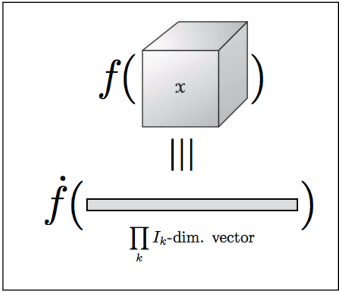
$$\min_{f \in \mathcal{F}} E_n \left[\ell(Y, f(X)) \right], \quad \ell : \text{loss function}$$

  - $\mathcal{F} := \left\{ f : \mathbb{R}^{I_1 \times \ldots \times I_K} \to \mathbb{R} | f \text{ is } \beta\text{-smooth} \right\}$
  - Let $\mathcal{F}$ be a hypothesis set

- **Problem** : the curse of dimensionality
  - An estimator by this approach has quite slow convergence

The curse of dimensionality

- Performance of the estimator of f gets worse with tensor input



Naive estimator $\tilde{f}_n$

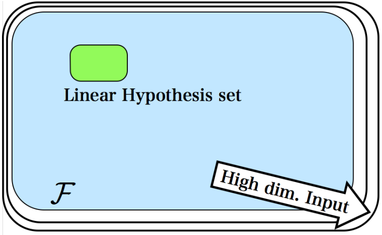
$$\|\tilde{f}_n - f^*\|^2 = O\left(n^{-2\beta/(2\beta+\prod_k I_k)}\right),$$

where β is smoothness of f.

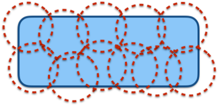
- $\prod_k I_k = \#$ of elements in X

Why the curse exists?

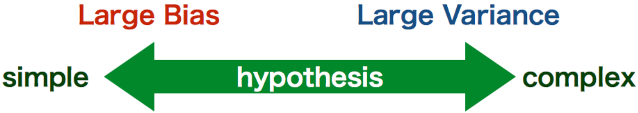
- The hypothesis set $\mathcal{F}$ is quite complex (large) due to high dimensionality of X



Metric entropy via ϵ-nets
(measure complexity)

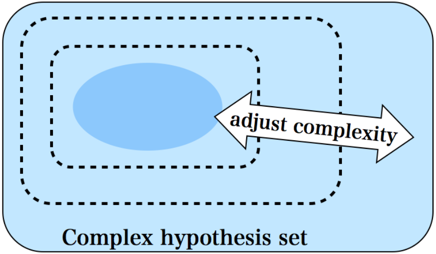


- Complex hypothesis sets make variance of estimators larger



Our idea

- **Reduce redundancy of hypotheses**
  - Data and models are often redundant
  - Represent X and f by less complex elements



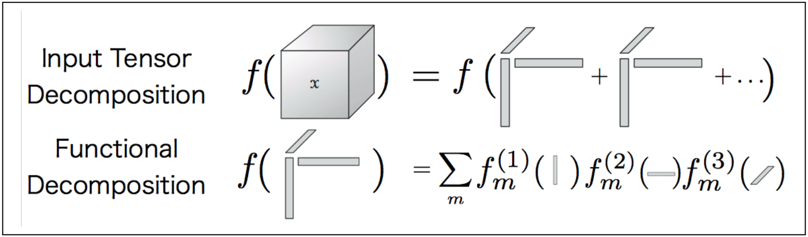
Example of reduction

1. Low-rank approx. of matrix
2. LASSO

...

Double Decomposition

- **Outline of Double Decomposition**
  1. Decompose input $X \in \mathbb{R}^{I_1 \times \ldots \times I_K}$
  2. Decompose function $f \in \mathcal{F}$



Double Decomposition 1

- **1.Tensor (CP) Decomposition**
  - Consider $X \in \mathbb{R}^{I_1 \times \ldots \times I_K}$
  - There exists a set of normalized vectors $\{x_r^{(k)} \in \mathbb{R}^{I_k}\}_{r,k=1,1}^{R^*,K}$ and scale term λ_r for all $r = 1, \ldots, R^*$, then

$$X = \sum_{r=1}^{R^*} \lambda_r x_r^{(1)} \otimes x_r^{(2)} \otimes \ldots \otimes x_r^{(K)}.$$

  - R^* is a tensor rank

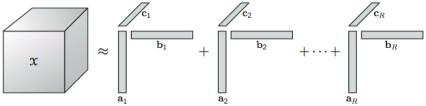


Image by Kolda+ (2009)

Double Decomposition 2

- **2.Functional Decomposition**
  - For each r, consider a function $f(x_r^{(1)}, x_r^{(2)}, \dots, x_r^{(K)})$
    - $x_r^{(k)}$ are I_k-dimensional vectors
  - There exist $M^* \in \mathbb{Z}_+ \cup \{\infty\}$ and a set of local functions $\{f_m^{(k)}\}_{k,m=1}^{K,M^*}$ satisfying

$$f(x_r^{(1)}, x_r^{(2)}, \dots, x_r^{(K)}) = \sum_{m=1}^{M^*} \prod_{k=1}^{K} f_m^{(k)}(x_r^{(k)}).$$

  - M^* is a model complexity

Proposed Framework

- Assumption
  - f is additive separable with respect to $r = 1, \dots, R^*$
- Consider doubly decomposed form of f

$$f(X) = \sum_{m=1}^{M^*} \sum_{r=1}^{R^*} \lambda_r \prod_{k=1}^{K} f_m^{(k)}(x_r^{(k)})$$

- **Additive-Multiplicative Nonparametric Regression (AMNR)**
  - Represent $f(X)$ by $f_m^{(k)}$ with a low-dimensional (I_k-dim.) vector as input
  - M^* (model complexity) and R^* (tensor rank) are tuning parameters

Proposed Framework

- **Our approach**

$$\min_{f \in \mathcal{G}} E_n \left[\ell(Y, f(X))\right]$$

  - $\mathcal{G} := \left\{ f : \mathbb{R}^{I_1 \times \ldots \times I_K} \to \mathbb{R} | f \text{ is AMNR, } f_m^{(k)} \text{ are } \beta\text{-smooth} \right\}$

- **Expected advantage**
  - $\mathcal{G}$ can be a less complex hypothesis set than $\mathcal{F}$ by tuning M^*
    - By calculating the metric entropy
  - $\mathcal{G}$ does not increase bias a few

Estimation Method

- The Bayes method with the Gaussian process prior.
- Prior

$$\pi(f) = \prod_m \prod_k \mathcal{GP}^{(k)}(f_m^{(k)}),$$

- Posterior

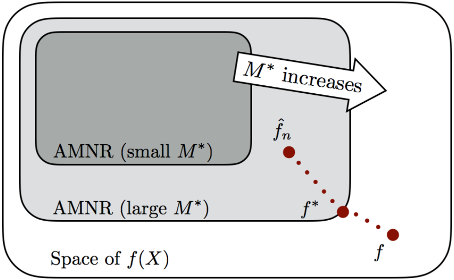
$$\pi(f|D_n) = \frac{\exp(-\sum_{i=1}^n (Y_i - G[f](X_i))^2)}{\int \exp(-\sum_{i=1}^n (Y_i - G[f'](X_i))^2)\pi(df')}\pi(f),$$

where $G[f](X_i) := \sum_{m=1}^{M^*} \sum_{r=1}^{R^*} \prod_{k=1}^{K} f_m^{(k)}\left(x_{r,i}^{(k)}\right).$

- Implementation
  - The estimation bases on Gibbs sampling

Analyze Convergence Theoretically

- In the form of AMNR, M^* controls the size of bias
  - As M^* increases, the bias decreases



- Focus on the distance between $\hat{f}_n$ and f^* with given M^*
  - We start with a case then finite M^* is sufficient to represent f
  - Then, we consider M^* is larger (infinite)

Finite M^* Case

- In the following, we assume that
  - i True $f_m^{(k)}$ belongs to Sobolev space with order β
  - ii Parameters of the prior estimation is appropriately selected

Theorem 1

Let $M^* < \infty$. Then, with some finite constant $C > 0$,

$$E\|\hat{f}_n - f^*\|_n^2 \leq C n^{-2\beta/(2\beta + \max_k I_k)}.$$

- Remind that the naive nonparametric estimator $\tilde{f}_n$ has a convergence rate $n^{-2\beta/(2\beta+\prod_k I_k)}$

Infinite M^* Case

- With infinite M^*, we estimate first M components with some assumption.

Theorem 2

Assume that with some constant $\gamma \geq 1$, $\left\| \sum_r \lambda_r \prod_k f_m^{(k)} \right\|_2 = o\left(m^{-\gamma-1}\right)$, as $m \to \infty$. Suppose we construct the estimator with a proximal complexity M such that

$$M \asymp \left(n^{2\beta/(2\beta+\max_k I_k)}\right)^{1/(1+\gamma)}.$$

Then, with some finite constant $C > 0$,

$$E\|\hat{f}_n - f^*\|_n^2 \leq C\left(n^{-2\beta/(2\beta+\max_k I_k)}\right)^{\gamma/(1+\gamma)}.$$

Convergence Rate

- Compare Nomparametric method for Tensor Regression
  - For example case, we set $K = 3, I_k = 100, \beta = \gamma = 2$

Method	Convergence Rate	Example
Naive	$n^{-2\beta/(2\beta+\prod_k I_k)}$	$n^{-1/2501}$
AMNR (Finite M^*)	$n^{-2\beta/(2\beta+\max_k I_k)}$	$n^{-1/26}$
AMNR (Infinite M^*)	$(n^{-2\beta/(2\beta+\max_k I_k)})^{\gamma/(1+\gamma)}$	$n^{-1/39}$

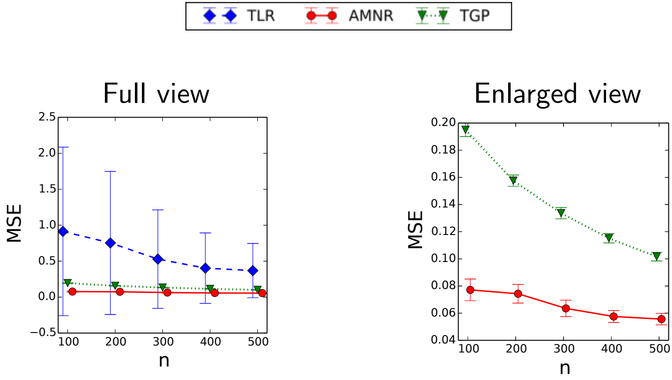
- AMNR achieves better convergence rate, by reducing the size of the model space by the double decomposition

Experiment Outline

- We introduce 3 experiments
  1. Prediction performance
  2. Convergence analysis
  3. Real data analysis
- Methods
  - AMNR (our method)
  - TGP (Tensor Gaussian Process)
    - Close to the naive nonparametric estimator
  - TLR (Tensor Linear Regression)
    - Not nonparametric method

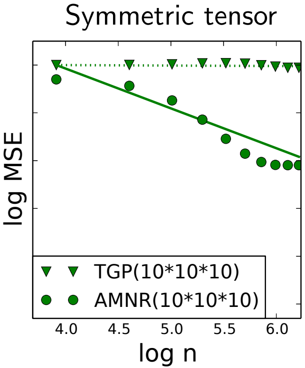
Prediction Performance

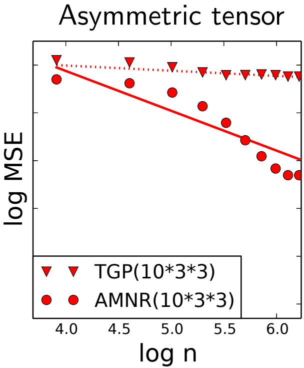
- Generate synthetic data with low rank tensor as
  $f(X) = \sum_{r=1}^{2} \lambda_r \prod_{k=1}^{K} (1 + \exp(-\gamma^T x_r^{(k)}))^{-1}$



Convergence analysis

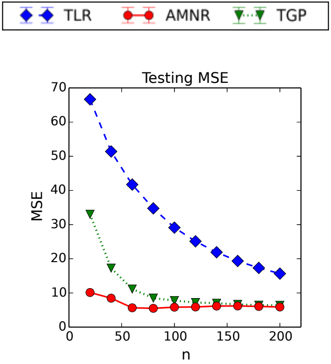
- Generate synthetic data with smoothness-controlled process $f(X) = \sum_{r=1}^{R} \prod_{k=1}^{K} \sum_{l} \mu_l \phi_l(\gamma^T x)$





Real Data Analysis

- Epidemic Spreading Data
  - X_i : Adjacency matrix of network i
  - Y_i : the number of total infected nodes of network i



Conclusion

- Summary
  - Proposed nonparametric regression model with tensor input
  - Doubly decomposition controls the hypothesis complexity
  - The control reduces the variance of the estimator
- Future work
  - Computational complexity / convergence
  - Tuning parameters (β, γ) selection
  - Measure bias size

Reference I

Dyrholm, M., Christoforou, C., and Parra, L. C. (2007). Bilinear discriminant component analysis. *The Journal of Machine Learning Research*, 8:1097–1111.

Guhaniyogi, R., Qamar, S., and Dunson, D. B. (2015). Bayesian tensor regression. *arXiv preprint arXiv:1509.06490*.

Hou, M., Wang, Y., and Chaib-draa, B. (2015). Online local gaussian process for tensor-variate regression: Application to fast reconstruction of limb movement from brain signal. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.

Reference II

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, 51(3):455–500.

Prettejohn, B. J., Berryman, M. J., and McDonnell, M. D. (2011). Methods for generating complex networks with selected structural properties for simulations: a review and tutorial for neuroscientists. *Frontiers in computational neuroscience*, 5:11.

Suzuki, T. (2015). Convergence rate of bayesian tensor estimator and its minimax optimality. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1273–1282.

Reference III

Zhao, Q., Zhou, G., Zhang, L., and Cichocki, A. (2014). Tensor-variate gaussian processes regression and its application to video surveillance. In *Acoustics, Speech and Signal Processing, 2014 IEEE International Conference*, pages 1265–1269. IEEE.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor regression with applications in neuroimaging data analysis. *Journal of the American Statistical Association*, 108(502):540–552.