Doubly Decomposing Nonparametric Tensor Regression

Masaaki Imaizumi INSOU11@HOTMAIL.COM
University of Tokyo

Kohei Hayashi HAYASHI.KOHEI@GMAIL.COM
National Institute of Advanced Industrial Science and Technology,
JST, ERATO, Kawarabayashi Large Graph Project

Abstract

Nonparametric extension of tensor regression is proposed. Nonlinearity in a high-dimensional tensor space is broken into simple local functions by incorporating low-rank tensor decomposition. Compared to naive nonparametric approaches, our formulation considerably improves the convergence rate of estimation while maintaining consistency with the same function class under specific conditions. To estimate local functions, we develop a Bayesian estimator with the Gaussian process prior. Experimental results show its theoretical properties and high performance in terms of predicting a summary statistic of a real complex network.

1. Introduction

Tensor regression deals with matrices or tensors (i.e., multi-dimensional arrays) as covariates (inputs) to predict scalar responses (outputs) (Wang et al., 2014; Hung & Wang, 2013; Zhao et al., 2014; Zhou et al., 2013; Tomioka et al., 2007; Suzuki, 2015; Guhaniyogi et al., 2015). Suppose we have a set of n observations $D_n = \{(Y_i, X_i)\}_{i=1}^n$; $Y_i \in \mathcal{Y}$ is a respondent variable in the space $\mathcal{Y} \subset \mathbb{R}$ and $X_i \in \mathcal{X}$ is a covariate with Kth-order tensor form in the space $\mathcal{X} \subset \mathbb{R}^{I_1 \times \ldots \times I_K}$, where I_k is the dimensionality of order k. With the above setting, we consider the regression problem of learning a function $f : \mathcal{X} \to \mathcal{Y}$ as

$$Y_i = f(X_i) + u_i, \tag{1}$$

where u_i is zero-mean Gaussian noise with variance σ^2. Such problems can be found in several applications. For example, studies on brain-computer interfaces attempt to predict human intentions (e.g., determining whether a subject imagines finger tapping) from brain activities. Electroencephalography (EEG) measures brain activities as electric signals at several points (channels) on the scalp, giving *channel* $\times$ *time* matrices as covariates. Functional magnetic resonance imaging captures blood flow in the brain as three-dimensional voxels, giving *X-axis* $\times$ *Y-axis* $\times$ *Z-axis* $\times$ *time* tensors.

There are primarily two approaches to the tensor regression problem. One is assuming linearity to f as

$$f(X_i) = \langle B, X_i \rangle, \tag{2}$$

where $B \in \mathbb{R}^{I_1 \times \ldots \times I_K}$ is a weight parameter with the same dimensionalities as X and $\langle B, X \rangle = \sum_{j_1,\ldots,j_K=1}^{I_1,\ldots,I_K} B_{j_1 \ldots j_K} X_{j_1 \ldots j_K}$ denotes the inner product. Since B is very high-dimensional in general, several authors have incorporated a low-rank structure to B (Dyrholm et al., 2007; Zhou et al., 2013; Hung & Wang, 2013; Wang et al., 2014; Suzuki, 2015; Guhaniyogi et al., 2015). We collectively refer to the linear models (2) with low-rank B as *tensor linear regression (TLR)*. As an alternative, a nonparametric approach has been proposed (Zhao et al., 2013; Hou et al., 2015). When $f(X)$ belongs to a proper functional space, with an appropriately choosing kernel function, the nonparametric method can estimate f perfectly even if f is nonlinear.

In terms of both theoretical and practical aspects, the *bias-variance tradeoff* is a central issue. In TLR, the function class that the model can represent is critically restricted due to its linearity and the low-rank constraint, implying that the variance error is low but the bias error is high if the true function is either nonlinear or full rank. In contrast, the nonparametric method can represent a wide range of functions and the bias error can be close to zero. However, at the expense of the flexibility, the variance error will be high due to the high dimensionality, the notorious nature of tensors. Generally, the optimal convergence rate of nonparametric models is given by

$$O(n^{-\beta/(2\beta+d)}), \tag{3}$$

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

which is dominated by the input dimensionality d and the smoothness of the true function β (Tsybakov, 2008). For tensor regression, d is the total number of X's elements, i.e., $\prod_k I_k$. When each dimensionality is roughly the same as $I_1 \simeq \cdots \simeq I_K$, $d = O(I_1^K)$, which significantly worsens the rate, and hinders application to even moderate-sized problems.

In this paper, to overcome the curse of dimensionality, we propose *additive-multiplicative nonparametric regression (AMNR)*, a new class of nonparametric tensor regression. Intuitively, AMNR constructs f as the sum of local functions taking the component of a rank-one tensor as inputs. In this approach, functional space and the input space are concurrently decomposed. This "double decomposition" simultaneously reduces model complexity and the effect of noise. For estimation, we propose a Bayes estimator with the Gaussian Process (GP) prior. The following theoretical results highlight the desirable properties of AMNR. Under some conditions,

- AMNR represents the same function class as the general nonparametric model, while
- the convergence rate (3) is improved as $d = I_{k'}$ ($k' = \mathrm{argmax}_k\, I_k$), which is $\prod_{k \neq k'} I_k$ times better.

We verify the theoretical convergence rate by simulation and demonstrate the empirical performance for real application in network science.

2. AMNR: Additive-Multiplicative Nonparametric Regression

First, we introduce the basic notion of tensor decomposition. With a finite positive integer R^*, the *CANDECOMP/PARAFAC (CP) decomposition* (Harshman, 1970; Carroll & Chang, 1970) of $X \in \mathcal{X}$ is defined as

$$X = \sum_{r=1}^{R^*} \lambda_r x_r^{(1)} \otimes x_r^{(2)} \otimes \ldots \otimes x_r^{(K)}, \tag{4}$$

where $\otimes$ denotes the tensor product, $x_r^{(k)} \in \mathcal{X}^{(k)}$ is a unit vector in a set $\mathcal{X}^{(k)} := \{v | v \in \mathbb{R}^{I_k}, \|v\| = 1\}$, and λ_r is the scale of $\{x_r^{(1)}, \ldots, x_r^{(K)}\}$ satisfying $\lambda_r \geq \lambda_{r'}$ for all $r > r'$. In this paper, R^* is the rank of X.

A similar relation holds for functions. Here, $\mathcal{W}^\beta(\mathcal{X})$ denotes a Sobolev space, which is β times differentiable functions with support $\mathcal{X}$. Let $g \in \mathcal{W}^\beta(S)$ be such a function. If S is given by the direct product of multiple supports as $S = S_1 \times \cdots \times S_J$, there exists a (possibly infinite) set of local functions $\{g_m^{(j)} \in \mathcal{W}^\beta(S_j)\}_m$ satisfying

$$g = \sum_m \prod_j g_m^{(j)} \tag{5}$$

for any g (Hackbusch, 2012, Example 4.40). This relation can be seen as an extension of tensor decomposition with infinite dimensionalities.

2.1. The Model

For brevity, we start with the case wherein X is rank one. Let $X = \bigotimes_k x_k := x_1 \otimes \cdots \otimes x_K$ with vectors $\{x_k \in \mathcal{X}^{(k)}\}_{k=1}^K$ and $f \in \mathcal{W}^\beta(\bigotimes_k \mathcal{X}^{(k)})$ be a function on a rank one tensor. For any f, we can construct $\tilde{f}(x_1, \ldots, x_K) \in \mathcal{W}^\beta(\mathcal{X}^{(1)} \times \ldots \times \mathcal{X}^{(k)})$ such that $\tilde{f}(x_1, \ldots, x_K) = f(X)$ using function composition as $\tilde{f} = f \circ h$ with $h : (x_1, \ldots, x_K) \mapsto \bigotimes_k x_K$. Then, using (5), f is decomposed into a set of local functions $\{f_m^{(k)} \in \mathcal{W}^\beta(\mathcal{X}^{(k)})\}_m$ as:

$$f(X) = \tilde{f}(x_1, \ldots, x_K) = \sum_{m=1}^{M^*} \prod_{k=1}^{K} f_m^{(k)}(x^{(k)}), \tag{6}$$

where M^* represents the complexity of f (i.e., the "rank" of the model).

With CP decomposition, (6) is amenable to extend for $X \in \mathcal{X}$ having a higher rank. For $R^* \geq 1$, we define AMNR as follows:

$$f^{\mathrm{AMNR}}(X) := \sum_{m=1}^{M^*} \sum_{r=1}^{R^*} \lambda_r \prod_{k=1}^{K} f_m^{(k)}(x_r^{(k)}). \tag{7}$$

Aside from the summation with respect to m, AMNR (7) is very similar to CP decomposition (4) in terms of that it takes summation over ranks and multiplication over orders. In addition, as λ_r indicates the importance of component r in CP decomposition, it controls how component r contributes to the final output in AMNR. Note that, for $R^* > 1$, equality between f^{AMNR} and $f \in \mathcal{W}^\beta(\mathcal{X})$ does not hold in general; see Section 4.

3. Truncated GP Estimator

3.1. Truncation of M^* and R^*

To construct AMNR (7), we must know M^*. However, this is unrealistic because we do not know the true function. More crucially, M^* can be infinite, and in such a case the exact estimation is computationally infeasible. We avoid these problems using predefined $M < \infty$ rather than M^* and ignore the contribution from $\{f_m^{(k)} : m > M\}$. This may increase the model bias; however, it decreases the variance of estimation. We discuss how to determine M in Section 4.2.

For R^*, we adopt the same strategy as M^*, i.e., we prepare some $R < R^*$ and approximate X as a rank- R tensor. Because this approximation reduces some information in X, the prediction performance may degrade. However, if R is

not too small, this preprocessing is justifiable for the following reasons. First, this approximation possibly removes the noise in X. In real data such as EEG data, X often includes observation noise that hinders the prediction performance. However, if the power of the noise is sufficiently small, the low-rank approximation discards the noise as the residual and enhances the robustness of the model. In addition, even if the approximation discards some intrinsic information of X, its negative effects could be limited because λs of the discarded components are also small.

3.2. Estimation Method and Algorithm

For each local function $f_m^{(k)}$, consider the GP prior $GP(f_m^{(k)})$, which is represented as multivariate Gaussian distribution $\mathcal{N}(0_{Rn}, K_m^{(k)})$ where 0_{Rn} is the zero element vector of size Rn and $K_m^{(k)}$ is a kernel Gram matrix of size $Rn \times Rn$. The prior distribution of the local functions $\mathfrak{F} := \{f_m^{(k)}\}_{m,k}$ is then given by:

$$\pi(\mathfrak{F}) = \prod_{m=1}^{M} \prod_{k=1}^{K} GP(f_m^{(k)}).$$

From the prior $\pi(\mathfrak{F})$ and the likelihood $\prod_i N(Y_i | f(X_i), \sigma^2)$, Bayes' rule yields the posterior distribution:

$$\pi(\mathfrak{F}|D_n) = \frac{\exp(-\sum_{i=1}^{n}(Y_i - G[\mathfrak{F}](X_i))^2/\sigma)}{\int \exp(-\sum_{i=1}^{n}(Y_i - G[\tilde{\mathfrak{F}}](X_i))^2/\sigma)\pi(\tilde{\mathfrak{F}})d\tilde{\mathfrak{F}}}\pi(\mathfrak{F}), \tag{8}$$

where $G[\mathfrak{F}](X_i) = \sum_{m=1}^{M}\sum_{r=1}^{R}\lambda_{r,i}\prod_{k=1}^{K} f_m^{(k)}(x_{r,i}^{(k)})$. $\tilde{\mathfrak{F}} = \{\tilde{f}_m^{(k)}\}_{m,k}$ are dummy variables for the integral. We use the posterior mean as the Bayesian estimator of AMNR:

$$\hat{f}_n = \int \sum_{m=1}^{M}\sum_{r=1}^{R}\lambda_{r,i}\prod_{k=1}^{K} f_m^{(k)} d\pi(\mathfrak{F}|D_n)d\mathfrak{F}. \tag{9}$$

To obtain predictions with new inputs, we derive the mean of the predictive distribution in a similar manner.

Since the integrals in the above derivations have no analytical solution, we compute them numerically by sampling. The details of the entire procedure are summarized as follows. Note that Q denotes the number of random samples.

- **Step 1: CP decomposition of input tensors**
  With the dataset D_n, apply rank-R CP decomposition to X_i and obtain $\{\lambda_{r,i}\}$ and $\{x_{r,i}^{(k)}\}$ for $i = 1, \ldots, n$.
- **Step 2: Construction of the GP prior distribution** $\pi(\mathfrak{F})$
  Construct a kernel Gram matrix $K_m^{(k)}$ from $\{x_r^{(k)}\}$ for each m and k, and obtain random samples of the multivariate Gaussian distribution $\mathcal{N}(0_{Rn}, K_m^{(k)})$. For each sampling $q = 1, \ldots, Q$, obtain a value $f_m^{(k)}(x_{r,i}^{(k)})$ for each r, m, k, and $i = 1, \ldots, n$.
- **Step 3: Computation of likelihood**
  To obtain the likelihood, calculate $\sum_m \sum_r \lambda_r \prod_k f_m^{(k)}(x_{r,i}^{(k)})$ for each sampling q and obtain the distribution by (8). Obtain the Bayesian estimator $\hat{f}$ and select the hyperparameters (optional).
- **Step 4: Prediction with the predictive distribution**
  Given a new input X', compute CP decomposition and obtain λ'_r and $\{x'^{(k)}_r\}_{r,k}$. Then, sample $f_m^{(k)}(x'^{(k)}_r)$ from the prior for each q. By multiplying the likelihood calculated in Step 3, derive the predictive distribution of $\sum_m \sum_r \lambda_r \prod_k f_m^{(k)}(x'^{(k)}_r)$ and obtain its expectation with respect to q.

4. Theoretical Analysis

Our main interest here is the asymptotic behavior of distance between the true function that generates data and an estimator (9). Preliminarily, let $f^0 \in \mathcal{W}^\beta(\mathcal{X})$ be the true function and $\hat{f}_n$ be the estimator of f^0. To analyze the distance in more depth, we introduce the notion of *rank additivity*[1] for functions, which is assumed implicitly when we extend (6) to (7).

Definition 1 (Rank Additivity)**.** *A function $f : \mathcal{X} \to \mathcal{Y}$ is rank additive if*

$$f\left(\sum_{r=1}^{R^*} \bar{x}_r\right) = \sum_{r=1}^{R^*} f(\bar{x}_r),$$

where $\bar{x}_r := \lambda_r x_r^{(1)} \otimes \ldots \otimes x_r^{(K)}$.

Letting f^* be a projection of f^0 onto the Sobolev space $f \in \mathcal{W}^\beta$ satisfying rank additivity, the distance is bounded above as

$$\|f^0 - \hat{f}_n\| \le \|f^0 - f^*\| + \|f^* - \hat{f}_n\|. \tag{10}$$

Unfortunately, the first term $\|f^0 - f^*\|$ is difficult to evaluate, aside from a few exceptions; if $R^* > 0$ or f^0 is rank additive, $\|f^0 - f^*\| = 0$.

Therefore, we focus on the rest term $\|f^* - \hat{f}_n\|$. By definition, f^* is rank additive and the functional tensor decomposition (6) guarantees that f^* is decomposed as the

[1]This type of additivity is often assumed in multivariate and additive model analysis (Hastie & Tibshirani, 1990; Ravikumar et al., 2009).

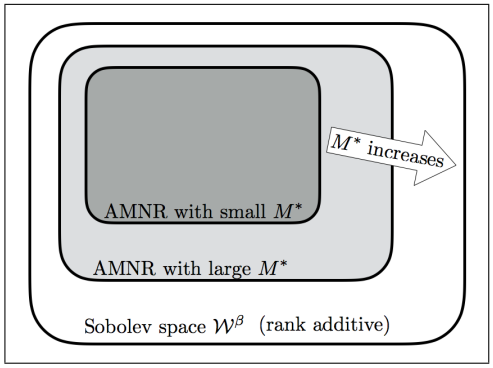


Figure 1. Functional space and the effect of M^*.

AMNR form (7) with some M^*. Here, the behavior of the distance strongly depends on M^*. We consider the following two cases: (i) M^* is finite and (ii) M^* is infinite. In case (i), the consistency of $\hat{f}_n$ to f^* is shown with an explicit convergence rate (Theorem 1). More surprisingly, the consistency also holds in case (ii) with a mild assumption (Theorem 2).

Figure 1 illustrates the relations of these functions and the functional space. The rectangular areas are the classes of functions represented by AMNR with small M^*, AMNR with large M^*, and Sobolev space $\mathcal{W}^\beta$ with rank additivity.

Note that the formal assumptions and proofs of this section are shown in supplementary material.

4.1. Estimation with Finite M^*

The consistency of Bayesian nonparametric estimators is evaluated in terms of posterior consistency (Ghosal et al., 2000; Ghosal & van der Vaart, 2007; van der Vaart & van Zanten, 2008). Here, we follow the same strategy. Let $\|f\|_n^2 := \frac{1}{n}\sum_{i=1}^n f(x_i)^2$ be the empirical norm. We define $\epsilon_n^{(k)}$ as the contraction rate of the estimator of local function $f^{(k)}$, which evaluates the probability mass of the GP around the true function. Note that the order of $\epsilon_n^{(k)}$ depends on the covariance kernel function of the GP prior, in which the optimal rate of $\epsilon_n^{(k)}$ is given by (3) with $d = I_k$. For brevity, we suppose that the variance of the noise u_i is known and the kernel in the GP prior is selected to be optimal.[2] Then, we obtain the following result.

Theorem 1 (Convergence analysis)**.** *Let $M = M^* < \infty$. Then, with Assumption 1 and some finite constant $C > 0$,*

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

Theorem 1 claims the validity of the estimator (9). Its convergence rate corresponds to the minimax optimal rate of estimating a function in $\mathcal{W}^\beta$ on compact support in $\mathbb{R}^{I_k}$, showing that the convergence rate of AMNR depends only on the largest dimensionality of X.

[2]We assume that the Matérn kernel is selected and the weight of the kernel is equal to β. Under these conditions, the optimal rate is achieved (Tsybakov, 2008).

4.2. Estimation with Infinite M^*

When M^* is infinite, we cannot use the same strategy used in Section 4.1. Instead, we truncate M^* by finite M and evaluate the bias error caused by the truncation. To evaluate the bias, we assume that the local functions are in descending order of their volumes $\bar{f}_m := \sum_r \prod_k f_m^{(k)}$, i.e., $\{\bar{f}_m\}$ are ordered as satisfying $\|\bar{f}_{m'}\|_2 \geq \|\bar{f}_m\|_2$ for all $m' > m$. We then introduce the assumption that $\{\|\bar{f}_m\|_2\}_{m=1}^\infty$ decays to zero polynomially with respect to m.

Assumption 1. *With some constant $\gamma \geq 1$,*

$$\|\bar{f}_m\|_2 = o\left(m^{-\gamma-1}\right),$$

as $m \to \infty$.

Then we claim the main result in this section.

Theorem 2. *Suppose we construct the estimator* (9) *with*

$$M \asymp (n^{-2\beta/(2\beta+\max_k I_k)})^{\gamma/(1+\gamma)},$$

where $\asymp$ denotes equality up to a constant. Then, with some finite constant $C > 0$,

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

The above theorem states that, even if we truncate M^* by finite M, the convergence rate is nearly the same as the case of finite M^* (Theorem 1), which is slightly worsened by the factor $\gamma/(1+\gamma)$.

Theorem 2 also suggests how to determine M. For example, if $\gamma = 2, \beta = 1$, and $\max_k I_k = 100$, $M \asymp n^{1/70}$ is recommended, which is much smaller than the sample size. Our experimental results (Section 6.3) also support this. Practically, very small M is sufficient, such as 1 or 2, even if n is greater than 300.

Here, we show the conditional consistency of AMNR, which is directly derived from Theorem 2.

Corollary 1. *For all function $f^* \in \mathcal{W}^\beta$ with finite $M = M^*$ or Assumption 1, the estimator* (9) *satisfies*

$$E\|\hat{f}_n - f^0\|_n^2 \to 0,$$

as $n \to \infty$.

5. Related Work

5.1. Nonparametric Tensor Regression

The *tensor GP (TGP)* (Zhao et al., 2014; Hou et al., 2015) in a method that estimates the function in $\mathcal{W}^\beta(\mathcal{S})$ directly.

TGP is essentially a GP regression model that flattens a tensor into a high-dimensional vector and takes the vector as an input. Zhao et al. (2014) proposed its estimator and applied it to image recognition from a monitoring camera. Hou et al. (2015) applied the method to analyze brain signals. Although both studies demonstrated the high performance of TGP, its theoretical aspects such as convergence have not been discussed.

Signoretto et al. (2013) proposed a regression model with tensor product reproducing kernel Hilbert spaces (TP-RKHSs). Given a set of vectors $\{x_k\}_{k=1}^K$, their model is written as

$$\sum_j \alpha_j \prod_k f_j^{(k)}(x_k), \tag{11}$$

where α_j is a weight. The key difference between TP-RKHSs (11) and AMNR is in the input. TP-RKHSs take only a single vector for each order, meaning that the input is implicitly assumed as rank one. On the other hand, AMNR takes rank-R tensors where R can be greater than one. This difference allows AMNR to be used for more general purposes, because the tensor rank observed in the real world is mostly greater than one. Furthermore, the properties of the estimator, such as convergence, have not been investigated. Kanagawa et al. (2016) proposed a similar model and investigated its theoretical properties more deeply such as minimax optimality.

5.2. TLR

For the matrix case ($K = 2$), Dyrholm et al. (2007) proposed a classification model as (2), where B is assumed to be low rank. Hung & Wang (2013) proposed a logistic regression where the expectation is given by (2) and B is a rank-one matrix. Zhou et al. (2013) extended these concepts for tensor inputs. Suzuki (2015) and Guhaniyogi et al. (2015) proposed a Bayes estimator of TLR and investigated its convergence rate.

Interestingly, AMNR is interpretable as a piecewise nonparametrization of TLR. Suppose B and X have rank-M and rank-R CP decompositions, respectively. The inner product *in the tensor space* in (2) is then rewritten as the product of the inner product *in the low-dimensional vector space*, i.e.,

$$\langle B, X_i \rangle = \sum_{m=1}^{M} \sum_{r=1}^{R} \lambda_{r,i} \prod_{k=1}^{K} \langle b_m^{(k)}, x_{r,i}^{(k)} \rangle, \tag{12}$$

where $b_m^{(k)}$ is the order-K decomposed vector of B. The AMNR formation is obtained by replacing the inner product $\langle b_m^{(k)}, x_r^{(k)} \rangle$ with local function $f_m^{(k)}$.

From this perspective, we see that AMNR incorporates the advantages of TLR and TGP. AMNR captures nonlinear relations between Y and X through $f_m^{(k)}$, which is impossible for TLR due to its linearity. Nevertheless, in contrast to TGP, an input of the function constructed in a nonparametric way is given by an I_k-dimensional vector rather than an $(I_1, \ldots, I_K)$-dimensional tensor. This reduces the dimension of the function's support and significantly improves the convergence rate (Section 4).

5.3. Other Studies

Koltchinskii & Yuan (2010) and Suzuki & Sugiyama (2013) investigated *Multiple Kernel Learning (MKL)* considering a nonparametric p-variate regression model with an additive structure: $\sum_{j=1}^p f_j(x_j)$. To handle high dimensional inputs, MKL reduces the input dimensionality by the additive structure for f_j and x_j. Note that both studies deal with a vector input, and they do not fit to tensor regression analysis.

Table 1. Comparison of related methods.

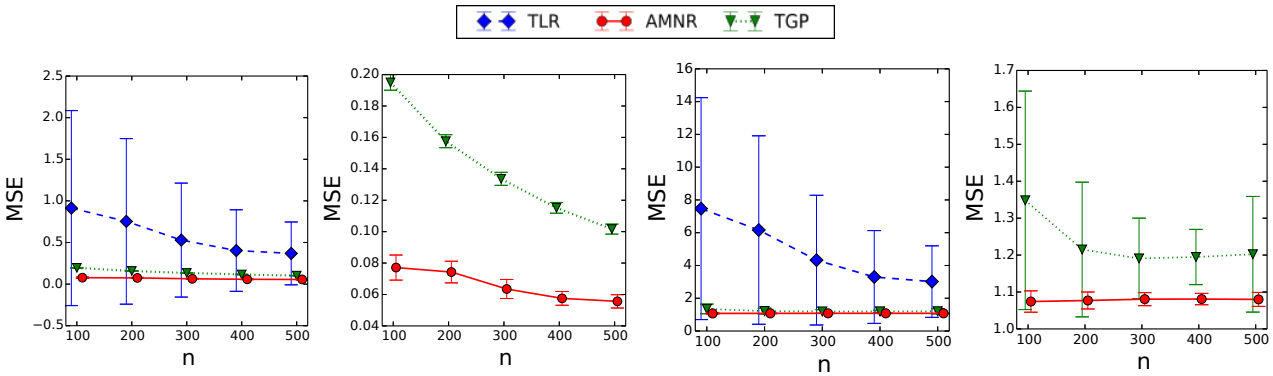
METHOD	TENSOR INPUT	NON-LINEARITY	CONVERGENCE RATE WITH (3)
TLR	$\surd$		$d = 0$
TGP	$\surd$	$\surd$	$d = \prod_k I_k$
TP-RKHSS	RANK-1	$\surd$	N/A
MKL		$\surd$	N/A
AMNR	$\surd$	$\surd$	$d = \max I_k$

Table 1 summarizes the methods introduced in this section. As shown, MKL and TP-RKHSs are not applicable for general tensor input. In contrast, TLR, TGP, and AMNR can take multi-rank tensor data as inputs, and their applicability is much wider. Among the three methods, AMNR is only the one that manages nonlinearity and avoids the curse of dimensionality on tensors.

6. Experiments

6.1. Synthetic Data

We compare the prediction performance of three models: TLR, TGP, and AMNR. In all experiments, we generate datasets by the data generating process (dgp) as $Y = f^*(X) + u$ and fix the noise variance as $\sigma^2 = 1$. We set the size of $X \in \mathbb{R}^{20 \times 20}$, i.e., $K = 2$ and $I_1 = I_2 = 20$. By varying the sample size as $n \in \{100, 200, 300, 400, 500\}$, we evaluate the empirical risks by the mean-squared-error (MSE) for the testing data, for which we use one-half of the samples. For each experiment, we derive the mean and variance of the MSEs in 100 trials. For TGP and AMNR, we optimize the parameter of the kernel function by grid search in the training phase.

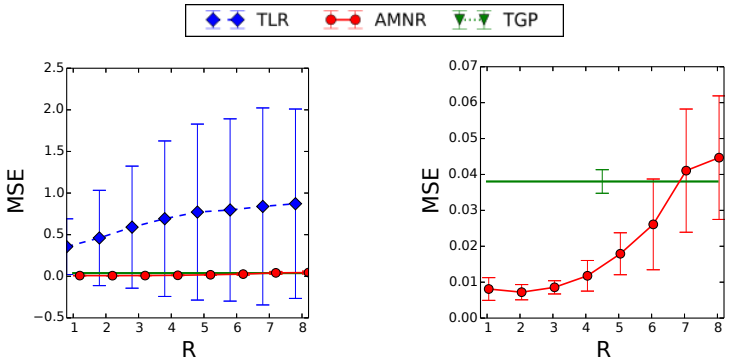


(a) Full view (b) Enlarged view

Figure 2. Synthetic data experiment: Low-rank data.

(a) Full view (b) Enlarged view

Figure 3. Synthetic data experiment: Full-rank data.



(a) Full view (b) Enlarged view

Figure 4. Synthetic data experiment: Sensitivity of R.

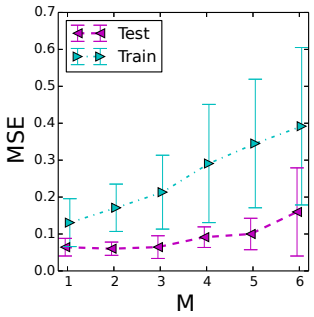


Figure 5. Synthetic data experiment: Sensitivity of M.

6.1.1. Low-Rank Data

First, we consider the case that X and the dgp are exactly low rank. We set R^*, the true rank of X, as $R^* = 4$ and

$$f^*(X) = \sum_{r=1}^{R^*} \lambda_r \prod_{k=1}^{K} (1 + \exp(\gamma^T x_r^{(k)}))^{-1}$$

where $[\gamma]_j = 0.1j$. The results (Figure 2) show that AMNR and TGP clearly outperform TLR, implying that they successfully capture the nonlinearity of the true function. To closely examine the difference between AMNR and TGP, we enlarge the corresponding part (Figure 2(b)), which shows that AMNR consistently outperforms TGP. Note that the performance of TGP improves gradually as n increases, implying that the sample size is insufficient for TGP due to its slow convergence rate.

6.1.2. Full-Rank Data

Next, we consider the case that X is full rank and the dgp has no low-rank structure, i.e., model misspecification will occur in TLR and AMNR. We generate X as $X_{j_1 j_2} \sim \mathcal{N}(0, 1)$ with

$$f^*(X) = \prod_{k=1}^{K} (1 + \exp(-\|X\|_2 / \prod_{k} I_k))^{-1}.$$

The results (Figure 3) show that, as in the previous experiment, AMNR and TGP outperform TLR. Although the difference between AMNR and TGP (Figure 3(b)) is much smaller, AMNR still outperforms TGP. This implies that, while the effect of AMNR's model misspecification is not negligible, TGP's slow convergence rate is more problematic.

6.2. Sensitivity of Hyperparameters

Here, we investigate how the truncation of R^* and M^* affect prediction performance. In the following experiments, we fix the sample size as $n = 300$.

First, we investigate the sensitivity of R. We use the same low-rank dgp used in Section 6.1.1 (i.e., $R^* = 4$.) The results (Figure 4) show that AMNR and TGP clearly outperform TLR. Although their performance is close, AMNR

beats TGP when R is not too large, implying that the negative effect of truncating R^* is limited.

Next, we investigate the sensitivity of M. We use the same full-rank dgp used in Section 6.1.2. Figure 5 compares the training and testing MSEs of AMNR, showing that both errors increase as M increases. These results imply that the model bias decreases quickly and estimation error is more dominant. Indeed, the lowest testing MSE is achieved at $M = 2$. This agrees satisfactory with the analysis in Section 4.2, which recommends small M.

6.3. Convergence Rate

Here, we confirm how the empirical convergence rates of AMNR and TGP meet the theoretical convergence rates. To clarify the relation, we generate synthetic data from dgp with $\beta = 1$ such that the difference between TGP and AMNR is maximized. Let parameters of the kernel function fit the known β. To do so, we design the dgp function as $f^*(X) = \sum_{r=1}^{R} \prod_{k=1}^{K} f^{(k)}(x_r^{(k)})$ and

$$f^{(k)} = \sum_l \mu_l \phi_l(\gamma^T x),$$

where $\phi_l(z) = \sqrt{2}\cos((l-1/2)\pi z)$ is an orthonormal basis function of the functional space and $\mu_l = l^{-3/2}\sin(l)$.[3]

SETTING NO.	K	R^*	I_1	I_2	I_3	d IN (3) TGP	d IN (3) AMNR
(I)	3	2	10	10	10	1000	10
(II)	3	2	3	3	3	27	3
(III)	3	2	10	3	3	90	10

Table 2. Synthetic data experiment: Settings for convergence rate.

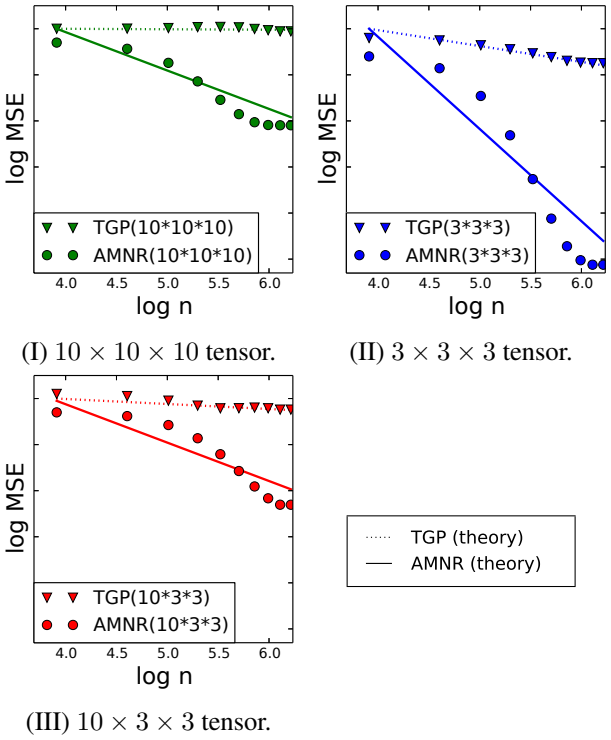
For X we consider three variations: $3 \times 3 \times 3$, $10\times, 3 \times 3$, and $10 \times 10 \times 10$ (Table 2). Figure 6 shows testing MSEs averaged over 100 trials. The theoretical convergence rates are also depicted by the dashed line (TGP) and the solid line (AMNR). To align the theoretical and empirical rates, we adjust them at $n = 50$. The result demonstrates the theoretical rates agree with the practical performance.

6.4. Prediction of Epidemic Spreading

Here, we deal with the *epidemic spreading* problem in complex networks (Anderson & May, 1991; Vespignani, 2012) as a matrix regression problem. More precisely, given an adjacency matrix network X_i, we simulate the spreading process of a disease by the susceptible-infected-recovered (SIR) model as follows.

1. We select 10 nodes as the initially infected nodes.

[3]This dgp is derived from a theory of Sobolev ellipsoid; see (Tsybakov, 2008).



(I) $10 \times 10 \times 10$ tensor. (II) $3 \times 3 \times 3$ tensor.

(III) $10 \times 3 \times 3$ tensor.

Figure 6. Comparison of convergence rate.

2. The nodes adjacent to the infected nodes become infected with probability 0.01.

3. Repeat Step 2. After 10 epochs, the infected nodes recover and are no longer infected (one iteration = one epoch).

After the convergence of the above process, we count the total number of infected nodes as Y_i. Note that the number of infected nodes depends strongly on the network structure and its prediction is not trivial. Conducting the simulation is of course a reliable approach; however, it is time-consuming, especially for large-scale networks. In contrast, once we obtain a trained model, regression methods make prediction very quick.

As a real network, we use the Enron email dataset (Klimt & Yang, 2004), which is a collection of emails. We consider these emails as undirected links between senders and recipients (i.e., this is a problem of estimating the number of email addresses infected by an email virus). First, to reduce the network size, we select the top $1,000$ email addresses based on frequency and delete emails sent to and received from other addresses. After sorting the remaining emails by timestamp, we sequentially construct an adjacency matrix from every $2,000$ emails, and we finally obtain 220 input matrices.

For the analysis, we set $R = 2$ for the AMNR and TLR

methods.[4] Although $R = 2$ seems small, we can still use the top-two eigenvalues and eigenvectors, which contain a large amount of information about the original tensor. In addition, the top eigenvectors are closely related to the threshold of outbreaks in infection networks (Wang et al., 2003). From these perspectives, the good performance demonstrated by AMNR with $R = 2$ is reasonable. The parameter of the kernel function is optimized by grid search in the training phase.

Figure 7 shows the training and testing MSEs. Firstly, there is a huge performance gap between TLR and the nonparametric models in the testing error. This indicates that the relation between epidemic spreading and a graph structure is nonlinear and the linear model is deficient for this problem. Secondly, AMNR outperforms TGP for every n in both training and testing errors. In addition, the performance of AMNR is constantly good and almost unaffected by n. This suggests that the problem has some extrinsic information that inflates the dimensionality, and the efficiency of TGP is diminished. On the other hand, it seems AMNR successfully captures the intrinsic information in a low-dimensional space by its "double decomposition" so that AMNR achieves the low-bias and low-variance estimation.

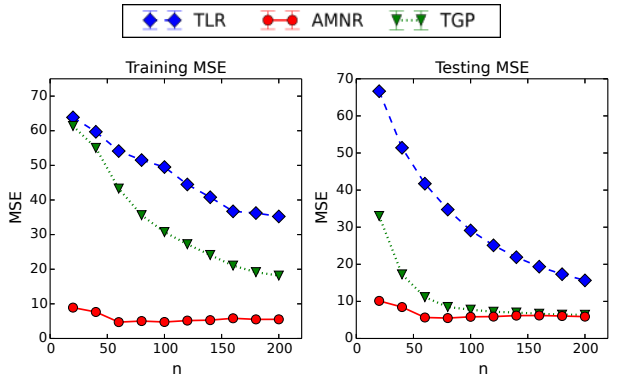


Figure 7. Epidemic spreading experiment: Prediction performance.

7. Conclusion and Discussion

We have proposed AMNR, a new nonparametric model for the tensor regression problem. The key of AMNR is that two different decompositions (4) and (5) are involved, which make the estimation problem easier and improve the convergence rate (Theorem 1). Although they produce an estimation bias, we empirically confirmed the bias was not critical (Sections 6.1.2 and 6.2).

The most important limitation of AMNR is the computational complexity, which is better than TGP but worse than TLR. The time complexity of AMNR with the GP estimator is $O(nR\prod_k I_k + M(nR)^3\sum_k I_k)$, where CP decomposition requires $O(R\prod_k I_k)$ (Phan et al., 2013) for n inputs and the GP prior requires $O(I_k(nR)^3)$ for the MK local functions. On the contrary, TGP requires $O(n^3\prod_k I_k)$ computation because it must evaluate all the elements of X to construct the kernel Gram matrix. When $R \ll n^2$ and $MR^3\sum_k I_k \ll \prod_k I_k$, which are satisfied in many practical situations, the proposed method is more efficient than TGP. Approximation methods for GP regression can be used to reduce the computational burden of AMNR. For example, Williams & Seeger (2001) proposed the Nyström method, which approximates the kernel Gram matrix by a low-rank matrix. If we apply rank-L approximation, the computational cost of AMNR can be reduced to $O((L^3 + nL^2)\prod_k I_k)$.

Note that the vector set $\{x_r^{(k)}\}_{r,k}$ obtained by CP decomposition (4) is not unique in general (Kolda & Bader, 2009), and this non-uniqueness violates the i.i.d. condition of $\{x_r^{(k)}\}_{r,k}$, which is a necessary condition for the convergence of the estimator (Corollary 1). However, in some cases, there exists an operation that makes the i.i.d. condition satisfied. For example, suppose X satisfies Kruskal's condition (Kruskal, 1977), which is

$$\sum_{k=1}^{K} r_k \geq 2R + K - 1, \tag{13}$$

where r_k is the largest number s such that every subset of s elements of $\{x_r^{(k)}\}_r$ are linearly independent. If (13) is satisfied, $\{x_r^{(k)}\}_{r,k}$ is unique, except for sign inversion. For instance, a tensor X with $R^* = 1$ and $K = 3$ has two equivalent decompositions: (A) $x_1 \otimes x_2 \otimes x_3$ and (B) $x_1 \otimes (-x_2) \otimes (-x_3)$. Nevertheless, by flipping the sign of the vectors randomly, i.e., of x_1, x_2, and x_3 at random while maintaining the original sign of X, we can regard the flipped vectors as i.i.d. When Kruskal's condition is not satisfied, it is unknown that whether such the operation exists. This is an important issue for future research.

Acknowledgement We would like to thank Taiji Suzuki and Taro Takaguchi for valuable discussions and comments. MI was supported by MEXT KAKENHI 15J10206. KH was supported by MEXT KAKENHI 15K16055.

References

Anderson, Roy M and May, Robert McCredie. *Infectious diseases of humans*, volume 1. Oxford university press, 1991.

Carroll, J Douglas and Chang, Jih-Jie. Analysis of individual differences in multidimensional scaling via an n-

[4] We also tested $R = 1, 2, 4$, and 8; however, the results were nearly the same.

way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319, 1970.

Dyrholm, Mads, Christoforou, Christoforos, and Parra, Lucas C. Bilinear discriminant component analysis. *The Journal of Machine Learning Research*, 8:1097–1111, 2007.

Ghosal, Subhashis and van der Vaart, Aad. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Ghosal, Subhashis, Ghosh, Jayanta K, and van der Vaart, Aad W. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

Guhaniyogi, Rajarshi, Qamar, Shaan, and Dunson, David B. Bayesian tensor regression. *arXiv preprint arXiv:1509.06490*, 2015.

Hackbusch, Wolfgang. *Tensor spaces and numerical tensor calculus*, volume 42. Springer Science & Business Media, 2012.

Harshman, Richard. Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16, 1970.

Hastie, Trevor J and Tibshirani, Robert J. *Generalized additive models*, volume 43. CRC Press, 1990.

Hou, Ming, Wang, Yali, and Chaib-draa, Brahim. 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*, 2015.

Hung, Hung and Wang, Chen-Chien. Matrix variate logistic regression model with application to EEG data. *Biostatistics*, 14(1):189–202, 2013.

Kanagawa, Heishiro, Suzuki, Taiji, Kobayashi, Hayato, and Yukihiro, Tagami. Gaussian process nonparametric tensor estimator and its minimax optimality. international conference on machine learning. In *Proceedings of The 33nd International Conference on Machine Learning*, 2016.

Klimt, Bryan and Yang, Yiming. The enron corpus: A new dataset for email classification research. In *Proceedings of 15th European Conference on Machine Learning*, volume 15, pp. 217. Springer Science & Business Media, 2004.

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

Koltchinskii, Vladimir and Yuan, Ming. Sparsity in multiple kernel learning. *The Annals of Statistics*, 38(6): 3660–3695, 2010.

Kruskal, Joseph B. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Phan, Anh-Huy, Tichavsky, Petr., and Cichocki, Andrzej. Fast alternating ls algorithms for high order candecomp/parafac tensor factorizations. *IEEE Transactions on Signal Processing*, 61(19):4834–4846, 2013.

Ravikumar, Pradeep, Lafferty, John, Liu, Han, and Wasserman, Larry. Sparse additive models. *Journal of the Royal Statistical Society: Series B*, 71(5):1009–1030, 2009.

Signoretto, Marco, De Lathauwer, Lieven, and Suykens, Johan AK. Learning tensors in reproducing kernel hilbert spaces with multilinear spectral penalties. *arXiv preprint arXiv:1310.4977*, 2013.

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

Suzuki, Taiji and Sugiyama, Masashi. Fast learning rate of multiple kernel learning: Trade-off between sparsity and smoothness. *The Annals of Statistics*, 41(3):1381–1405, 2013.

Tomioka, Ryota, Aihara, Kazuyuki, and Müller, Klaus-Robert. Logistic regression for single trial eeg classification. In *Advances in Neural Information Processing Systems 19*, pp. 1377–1384, 2007.

Tsybakov, Alexandre B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

van der Vaart, AW and van Zanten, H. Rates of contraction of posterior distributions based on gaussian process priors. *The Annals of Statistics*, 36(3):1435–1463, 2008.

Vespignani, Alessandro. Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8: 32–39, 2012.

Wang, Fei, Zhang, Ping, Qian, Buyue, Wang, Xiang, and Davidson, Ian. Clinical risk prediction with multilinear sparse logistic regression. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

Wang, Yang, Chakrabarti, Deepayan, Wang, Chenxi, and Faloutsos, Christos. Epidemic spreading in real networks: An eigenvalue viewpoint. In *Proceedings. 22nd International Symposium on Reliable Distributed Systems, 2003.*, pp. 25–34. IEEE, 2003.

Williams, Christopher KI and Seeger, Matthias. Using the nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems 13*, pp. 682–688, 2001.

Zhao, Qibin, Zhang, Liqing, and Cichocki, Andrzej. A tensor-variate gaussian process for classification of multidimensional structured data. In *AAAI Conference on Artificial Intelligence*, 2013.

Zhao, Qibin, Zhou, Guoxu, Zhang, Liqing, and Cichocki, Andrzej. Tensor-variate gaussian processes regression and its application to video surveillance. In *Acoustics, Speech and Signal Processing, 2014 IEEE International Conference*, pp. 1265–1269. IEEE, 2014.

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