# Gaussian process nonparametric tensor estimator and its minimax optimality

**Heishiro Kanagawa**[†] KANAGAWAH.AB@M.TITECH.AC.JP
**Taiji Suzuki**[†,‡] SUZUKI.T.CT@M.TITECH.AC.JP
[†] Tokyo Institute of Technology, Tokyo 152-8552, JAPAN
[‡] PRESTO, Japan Science and Technological Agency (JST), JAPAN

**Hayato Kobayashi**[⋆] HAKOBAYA@YAHOO-CORP.JP
**Nobuyuki Shimizu**[⋆] NOBUSHIM@YAHOO-CORP.JP
**Yukihiro Tagami**[⋆] YUTAGAMI@YAHOO-CORP.JP
[⋆] Yahoo Japan Corporation, Tokyo 107-6211, JAPAN

## Abstract

We investigate the statistical efficiency of a nonparametric Gaussian process method for a nonlinear tensor estimation problem. Low-rank tensor estimation has been used as a method to learn higher order relations among several data sources in a wide range of applications, such as multi-task learning, recommendation systems, and spatiotemporal analysis. We consider a general setting where a common linear tensor learning is extended to a nonlinear learning problem in reproducing kernel Hilbert space and propose a nonparametric Bayesian method based on the Gaussian process method. We prove its statistical convergence rate without assuming any strong convexity, such as restricted strong convexity. Remarkably, it is shown that our convergence rate achieves the minimax optimal rate. We apply our proposed method to multi-task learning and show that our method significantly outperforms existing methods through numerical experiments on real-world data sets.

## 1. Introduction

Tensor structure naturally arises in the analysis of complex interactions between several data sources. For example, in a movie recommendation system, user ratings of movies are described by a three-way tensor, which is referenced by (user, movie, context) (Karatzoglou et al., 2010). The noteworthy success of tensor data analysis is based on the notion of the low rank property of a tensor. The low rank decomposition of a tensor, which is analogous to that of a matrix, plays an important role in tensor learning, since it enables us to decompose a tensor into a few factors that can be analyzed, e.g., *CP-decomposition* (Hitchcock, 1927a;b) and *Tucker decomposition* (Tucker, 1966).

Parametric models based on low rank decomposition have been intensively investigated. A naive approach to computing tensor decomposition is to find a best approximate decomposition that minimizes the squared error (Kolda & Bader, 2009). However, this type of problem is not convex and therefore it is difficult to derive an optimal solution. In order to overcome the computational difficulty, convex relaxation methods have been proposed by some authors (Liu et al., 2009; Signoretto et al., 2010; Gandy et al., 2011; Tomioka et al., 2011; Tomioka & Suzuki, 2013). The main idea of convex relaxations is to unfold a tensor into some matrices and apply a well investigated trace norm regularization method to these matrices.

On the other hand, Bayesian methods have also been proposed (Chu & Ghahramani, 2009; Xiong et al., 2010; Xu et al., 2011; Rai et al., 2014). In the context of sparse estimation, it is known that a Bayesian method does not require strong conditions on the design to derive an optimal convergence rate (Dalalyan & Tsybakov, 2008; Alquier & Lounici, 2011). In fact, in low rank linear tensor estimation, a minimax optimal convergence rate for predictive error can be achieved without assuming strong convexity on the design (Suzuki, 2015).

In addition to the linear models as described above, nonparametric tensor learning has also been considered (Xu et al., 2011; Signoretto et al., 2013; Shen & Ghosal, 2016).

*Proceedings of the $33^{rd}$ International Conference on Machine Learning*, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

In particular, Signoretto et al. (2013) extended linear tensor learning to the non-parametric learning problem using a kernel method; they proposed applying a regularization method on the reproducing kernel Hilbert space (RKHS) of the tensor product. However, this approach is not guaranteed to produce a global optimal solution and the statistical optimality of that is not theoretically justified.

In this paper, we consider a Bayesian method that employs the Gaussian process prior as a distinct approach for learning functions on the RKHS. We theoretically investigate the Bayesian tensor estimator and derive a fast convergence rate without assuming any strong condition on the design through a PAC-Bayesian technique (Dalalyan & Tsybakov, 2008; Alquier & Lounici, 2011; Rigollet & Tsybakov, 2011). Our bound is considerably general in a sense that it gives the learning rate for estimation in a general RKHS, and it also covers a situation where the true function is not exactly included in the RKHS by utilizing the notion of *interpolation space* (Bennett & Sharpley, 1988). Furthermore, we derive the *minimax optimal* lower bound and it is shown that the convergence rate of the Gaussian process method achieves the minimax optimal rate.

We also apply our proposed method to a multi-task learning problem. We conduct extensive experiments, the results of which show that the nonlinear Gaussian process method significantly outperforms the existing linear tensor learning methods (Romera-Paredes et al., 2013; Wimalawarne et al., 2014; Suzuki, 2015).

## 2. Problem formulation

Suppose that we are given $n$ input-output samples $\{(x_i, y_i)\}_{i=1}^n$. The input $x_i$ is a concatenation of $K$ variables, i.e., $x_i = (x_i^{(1)}, \cdots, x_i^{(K)}) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_K = \mathcal{X}$, where each $x_i^{(k)}$ is an element of a set $\mathcal{X}_k$. We consider the regression problem where these samples are generated according to the non-parametric model (Signoretto et al., 2013):

$$y_i = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f_r^{*(k)}(x_i^{(k)}) + \epsilon_i, \tag{1}$$

where $\{\epsilon_i\}_{i=1}^n$ represents an i.i.d. zero-mean noise. In this regression problem, our objective is to estimate the true function $f^*(x^{(1)}, \dots, x^{(K)}) = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f_r^{*(k)}(x^{(k)})$.

This model captures the effect of non-linear higher order interactions among the input components $\{x^{(k)}\}_{k=1}^K$ to the output $y$, and thus, is useful for a regression problem where the output is determined by complex relations between the input components. This type of regression problem appears in several applications, such as multi-task learning, recommendation systems and spatiotemporal data analysis (Karatzoglou et al., 2010; Romera-Paredes et al., 2013; Bahadori et al., 2014) (see Sec. 6.1 for the multi-task learning formulation).

**Relation to the linear tensor model** To understand the model in Eq. (1), it is helpful to consider a linear case as a special case (Chu & Ghahramani, 2009; Tomioka et al., 2011). In general, the linear tensor model is formulated as

$$Y_i = \langle A^*, X_i \rangle + \epsilon_i. \tag{2}$$

Here, $X_i$, $A^*$ are tensors in $\mathbb{R}^{M_1 \times \cdots \times M_K}$ and the inner product $\langle \cdot, \cdot \rangle$ is defined by $\langle A, X \rangle = \sum_{i_1,\dots,i_K=1}^{M_1,\dots,M_K} A_{i_1 \dots i_K} X_{i_1 \dots i_K}$. $A^*$ is assumed to be low rank in the sense of CP-rank (Hitchcock, 1927a;b), i.e., $A^*$ is decomposed as $\sum_{r=1}^{d^*} u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}$, where the vector $u_r^{*(k)} \in \mathbb{R}^{M_k}$ and the symbol $\circ$ represents the vector outer product. If we also assume $X_i$ is rank-1, i.e., $X_i = x_i^{(1)} \circ \cdots \circ x_i^{(K)}$, then the inner product in Eq.(2) is written as: $\langle A^*, X_i \rangle = \langle \sum_{r=1}^{d} u_r^{*(1)} \circ \cdots \circ u_r^{*(K)}, x_i^{(1)} \circ \cdots \circ x_i^{(K)} \rangle = \sum_{r=1}^{d^*} \prod_{k=1}^{K} \langle u_r^{*(k)}, x_i^{(k)} \rangle$. This is equivalent to the case where we limit $f_r^{*(k)}$ in Eq. (1) to the linear function $\langle u_r^{*(k)}, x^{(k)} \rangle$. Hence, the linear model based on CP-decomposition can be understood as a special case of our proposed model.

## 3. Estimation of nonlinear tensor model

Our approach analyzed in this paper is a Bayesian method in which a Gaussian process prior is employed for each $f_r^{(k)}$. We can compute the posterior distribution by using MCMC technique and give the statistical convergence rate of the posterior mean estimator.

### 3.1. Gaussian process prior and corresponding reproducing kernel Hilbert space

We place a zero-mean Gaussian process prior $\mathrm{GP}_{r,k}$ with a kernel $k_{(r,k)}$ to estimate the function $f_r^{*(k)}$ on $\mathcal{X}_k$. A zero-mean Gaussian process $f = (f(x) : x \in \mathcal{X})$ on some input space $\mathcal{X}$ is a set of random variables $(f(x))_{x \in \mathcal{X}}$ indexed by $\mathcal{X}$ such that each finite subset $(f(x_1), \dots, f(x_j))$ $(j = 1, 2, \dots)$ obeys a zero-mean multivariate normal distribution, where $(x_1, \dots, x_j) \subseteq \mathcal{X}$ is an arbitrary finite subset of $\mathcal{X}$. The kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ corresponding to the Gaussian process is the covariance function defined by $k(x, x') = E[f(x)f(x')]$. Since the kernel function is symmetric and positive definite, we can define its corresponding RKHS in the usual manner (Aronszajn, 1950).

We denote by $\mathcal{H}_{(r,k)}$ the RKHS corresponding to the kernel $k_{(r,k)}$. It is known that the RKHS is usually much smaller than the support of the Gaussian process in an infinite dimensional setting. In fact, typically the prior has probability mass 0 on the infinite dimensional RKHS (van der

Vaart & van Zanten, 2011). This leads to the fact that, under the assumption $f_r^{*(k)} \in \mathcal{H}_{(r,k)}$, estimating the function $f_r^{*(k)}$ through the standard Bayesian procedure with a Gaussian process prior never achieves the optimal rate in some important examples (van der Vaart & van Zanten, 2011). To overcome this issue, we scale the process by the factor of $\lambda_{(r,k)}$ and make the estimator close to the small space $\mathcal{H}_{(r,k)}$.

### 3.2. The posterior distribution and the corresponding estimator

Given a rank $d$, let $\mathcal{F} = (f_r^{(k)})_{r=1,\dots,d,\ k=1,\dots,K}$ be a concatenation of functions $\{f_r^{(k)}\}_{r=1,\dots,d,\ k=1,\dots,K}$. Let the Gaussian process prior $\mathrm{GP}_{r,k}(\cdot|\lambda_{(r,k)})$ with a parameter $\lambda_{(r,k)} > 0$ be the process associated with a "scaled" kernel function $k_{(r,k)}/\lambda_{(r,k)}$. We consider the following prior distribution on the product space $\mathrm{d}\mathcal{F} = (\mathrm{d}f_r^{(k)})_{r=1,\dots,d,\ k=1,\dots,K}$:

$$\Pi(\mathrm{d}\mathcal{F}|d) = \prod_{r=1}^{d}\prod_{k=1}^{K}\int_{\lambda_{(r,k)}>0}\mathrm{GP}_{r,k}(\mathrm{d}f_r^{(k)}|\lambda_{(r,k)})\mathcal{G}(\mathrm{d}\lambda_{(r,k)}),$$

where $\mathcal{G}$ denotes the exponential distribution, $\mathcal{G}(\mathrm{d}\lambda_{(r,k)}) = \exp(-\lambda_{(r,k)})\mathrm{d}\lambda_{(r,k)}$, which is a conjugate prior for the scale of the Gaussian process priors. It will be shown that, by involving the scaling parameter $\lambda_{(r,k)}$, the estimator is able to achieve the optimal convergence rate while it can not without scaling as described above. Putting a prior distribution on $\lambda_{(r,k)}$ rather than fixing it to some optimally chosen value is rather for theoretical purpose, but by doing so, the estimator possesses an adaptivity against a property of $f^*$. Finally, we place a prior distribution on the rank $1 \le d \le d_{\max}$ as

$$\pi(d) = \frac{\xi^d}{\sum_{d'=1}^{d_{\max}}\xi^{d'}}, \tag{3}$$

where $0 < \xi < 1$ is some positive real number and $d_{\max}$ is a sufficiently larger number than the supposed true rank $d^*$.

We now provide the posterior distribution and the corresponding Bayesian estimator. For some $\beta > 0$, the posterior measure is constructed as

$$\Pi(\mathrm{d}\mathcal{F}|D_n) = \frac{\sum_{d=1}^{d_{\max}}\Pi(D_n|\mathcal{F})}{\sum_{d=1}^{d_{\max}}\int\Pi(D_n|\tilde{\mathcal{F}})\Pi(\mathrm{d}\tilde{\mathcal{F}}|d)\pi(d)}\Pi(\mathrm{d}\mathcal{F}|d)\pi(d),$$

where $\Pi(D_n|\mathcal{F})$ is a *quasi likelihood* defined by

$$\Pi(D_n|\mathcal{F}) = \exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n}\left(y_i - \sum_{r=1}^{d}\prod_{k=1}^{K}f_r^{(k)}(x_i^{(k)})\right)^2\right\}$$

with a temperature parameter $\beta > 0$. Although the noise $\epsilon_i$ is not necessarily Gaussian, we suggest using the Gaussian likelihood as above. It will be shown that even with this quasi likelihood, we obtain a nice convergence property. Corresponding to the posterior, we have the postierior mean estimator $\hat{f}$: $\hat{f} = \int f\Pi(\mathrm{d}\mathcal{F}|D_n)$.

### 3.3. Posterior sampling

Here, we describe the generation of samples from the posterior distribution. In practice, we may estimate the scale parameter $\{\lambda_{(r,k)}\}_{r,k}$ and the rank $d$ by minimizing the validation error. In the following, we fix these parameters for simplicity. Let $\mathbf{f}_r^{(k)} = (f_r^{(k)}(x_1^{(k)}), \dots, f_r^{(k)}(x_n^{(k)})) \in \mathbb{R}^n$ and $\mathbf{f}_{-r}^{(-k)} = \{\mathbf{f}_{r'}^{(k')} \mid (r',k') \neq (r,k)\}$. Since the prior and the likelihood constitute a Gaussian distribution, the conditional posterior distribution of $\mathbf{f}_r^{(k)}$ given the other components $\mathbf{f}_{-r}^{(-k)}$ is also a Gaussian distribution. Thus, we may apply *Gibbs sampling* for the computation of the mean estimator. Let $\boldsymbol{y} = (y_i)_{i=1}^n$ and the Gram matrix be $\boldsymbol{K}_r^{(k)} = (k_{(r,k)}(x_i^{(k)}, x_j^{(k)}))_{i,j} \in \mathbb{R}^{n\times n}$. Then, a simple calculation gives the conditional distribution of $\mathbf{f}_r^{(k)}$ as

$$\begin{aligned}
&\pi(\mathbf{f}_r^{(k)} \mid \mathbf{f}_{-r}^{(-k)}, D_n) = N(\boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_r^{(k)}), \qquad (4)\\
&\boldsymbol{\mu}_r^{(k)} = \boldsymbol{\Sigma}_r^{(k)}\big(\boldsymbol{a} * (\boldsymbol{y} - \boldsymbol{b})\big)/\beta,\\
&\boldsymbol{\Sigma}_r^{(k)} = \boldsymbol{K}_r^{(k)} - \boldsymbol{K}_r^{(k)}\big(\boldsymbol{K}_r^{(k)} + \mathrm{diag}(a_1^2, \dots, a_n^2)/\beta\big)^{-1}\boldsymbol{K}_r^{(k)},
\end{aligned}$$

where the symbol $*$ is the Hadamard product and $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}$,

$$\boldsymbol{a} = \Big(\prod_{\tilde{k}\neq k} f_r^{(\tilde{k})}(x_i^{(\tilde{k})})\Big)_{i=1}^n, \ \boldsymbol{b} = \Big(\sum_{r'\neq r}\prod_{\tilde{k}=1}^{K} f_{r'}^{(\tilde{k})}(x_i^{(\tilde{k})})\Big)_{i=1}^n. \tag{5}$$

To perform the Gibbs sampling, we iteratively sample $\mathbf{f}_r^{(k)} \in \mathbb{R}^n$ from the conditional distribution given above by shifting $(r,k)$ cyclically.

Suppose the $i$-th posterior sample is denoted by $(\mathbf{f}_{r,[i]}^{(k)})_{r,k}$. Then, to predict the function value $f^*(x)$ at a new input $x = (x^{(1)}, \dots, x^{(K)})$, we compute the conditional posterior mean of $f_r^{(k)}(x^{(k)})$ conditioned by $\mathbf{f}_{r,[i]}^{(k)}$ according to

$$\hat{f}_{r,[i]}^{(k)} = \mathbf{k}_{(r,k)}^\top\big(\boldsymbol{K}_r^{(k)}\big)^{-1}\mathbf{f}_{r,[i]}^{(k)}. \tag{6}$$

where $\mathbf{k}_{(r,k)} := (k_{(r,k)}(x_1^{(k)}, x^{(k)}), \dots, k_{(r,k)}(x_n^{(k)}, x^{(k)}))^\top$. Then we take an average $\frac{1}{N}\sum_{i=1}^{N}\left(\sum_{r=1}^{d}\prod_{k=1}^{K}\hat{f}_{r,[i]}^{(k)}\right)$, (where $N$ is the number of sampling iterations), over the entire iteration as the predicted function value at $x$.

## 4. Convergence rate analysis

In this section, we provide the statistical convergence rate of our Gaussian process tensor estimator, and show that the derived convergence rate is actually minimax optimal (up to constants).

### 4.1. Upper bound for correctly specified setting

First, we assume a condition on the noise $\epsilon_i$ as follows.

**Assumption 1** $\mathrm{E}[\epsilon_1^2] < \infty$ *and* $\mathrm{E}[\epsilon_1] = 0$. *Let* $m_\epsilon(z) := \int_z^\infty y \mathrm{d}F_\epsilon(y)$ *where* $F_\epsilon(z) = P(\epsilon_1 \leq z)$ *is the cumulative distribution function of the noise* $\epsilon_i$. *The measure* $m_\epsilon(z)\mathrm{d}z$ *is absolutely continuous with respect to the distribution function* $F_\epsilon(z)$ *with a bounded Radon-Nikodym derivative, i.e., there exists a bounded function* $g_\epsilon : \mathbb{R} \to \mathbb{R}_+$ *such that*

$$\int_a^b m_\epsilon(z)\mathrm{d}z = \int_a^b g_\epsilon(z)\mathrm{d}F_\epsilon(z), \quad \forall a, b \in \mathbb{R}.$$

Roughly speaking, this assumption indicates the noise has a light tail probability. In fact, the Gaussian noise $N(0,1)$ satisfies this assumption with $g_\epsilon(z) = \sigma^2$. See Dalalyan & Tsybakov (2008) for more details.

Next, we introduce a quantity that measures the complexity of the RKHSs. More specifically, we assume that the RKHSs defined by the kernels have a polynomial order complexity of the *metric entropy* of their unit balls. Let $N(B, \epsilon, d)$ denote the $\epsilon$-covering number of the space $B$ with respect to the metric $d$ (van der Vaart & Wellner, 1996), that is, the smallest number of $\epsilon$-balls that are required to cover $B$, where the radius $\epsilon$ of the $\epsilon$-balls is measured by the metric $d$. The metric entropy is the logarithm of the covering number. Let $\mathcal{B}_{\mathcal{H}_{(r,k)}}$ be the unit ball of the RKHS $\mathcal{H}_{(r,k)}$.

**Assumption 2** *There exists a real value* $0 < s_{(r,k)} < 1$ *and* $C_0 > 0$ *such that*

$$\log N(\mathcal{B}_{\mathcal{H}_{(r,k)}}, \epsilon, \|\cdot\|_n) \leq C_0 \epsilon^{-2 s_{(r,k)}} \quad (\epsilon > 0). \tag{7}$$

*Moreover, the kernel function is bounded as* $\sup_x k_{r,k}(x,x) \leq 1$.

An interesting fact is that the metric entropy condition in Eq. (7) controls the *small ball probability* of the corresponding Gaussian process as $-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_n \leq \epsilon\})) = O\left(\epsilon^{-2s_{(r,k)}/(1-s_{(r,k)})}\right)$ (Kuelbs & Li, 1993; Li & Shao, 2001). This assumption is usually satisfied by practically used kernels. For example, the Gaussian kernel satisfies this condition with an arbitrary $s_{(r,k)}$ with a different constant $C_0$ with high probability.

Next, we assume that the prior has a sufficient mass on bounded functions. This is a technical assumption and practically used kernels usually satisfy this assumption.

**Assumption 3** *There exists* $c_1 > 0$ *such that*

$$-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq 1\})) \leq c_1 \quad (\forall r, k).$$

Moreover, we assume the following condition on the true function $f^*$.

**Assumption 4** $f_r^{*(k)}$ *is included in* $\mathcal{H}_{(r,k)}$ *for all* $1 \leq r \leq d_{\max}$ *and* $1 \leq k \leq K$. *There exists* $R$ *such that* $\max_{(r,k)} \|f_r^{*(k)}\|_{\mathcal{H}_{(r,k)}} \leq R$. *The true tensor is low rank, that is, there exists* $d^*$ *such that* $f_r^{*(k)} = 0$ *for all* $r > d^*$ *and* $1 \leq k \leq K$.

Under these assumptions, we have the following estimation error bound.

**Theorem 1** *Suppose that Assumptions 1, 2, 3, and 4 are satisfied, and* $\beta \geq 4\|g_\epsilon\|_\infty$. *Then, there exists a constant* $C > 0$ *depending on* $\beta$, $C_0$, $c_1$ *and* $s_{(r,k)}$ *such that*

$$\begin{aligned}&\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^*\|_n^2\right] \\ \leq & C\left\{(3R \vee 1)^{2(K-1)} \sum_{r=1}^{d^*}\sum_{k=1}^{K} n^{-\frac{1}{1+s_{(r,k)}}} + \frac{d^*}{n}\log\left(\frac{1}{\kappa}\right)\right\},\end{aligned}$$

*where* $\mathrm{E}_{Y_{1:n}|x_{1:n}}$ *indicates the expectation with respect to the outputs* $Y_1, \ldots, Y_n$ *conditioned by the inputs* $x_1, \ldots, x_n$, *and* $\kappa = \xi(1-\xi)$.

The proof is given in the supplementary material. Basically, the proof is obtained by using the PAC-Bayes bound (McAllester, 1998; 1999; Catoni, 2004) (the version we used was developed by Dalalyan & Tsybakov (2008)), and applying the small ball probability theorems of Gaussian processes (Kuelbs & Li, 1993; Li & Shao, 2001).

If $K = 1$, the convergence rate coincides with the usual one of the ordinary kernel ridge regression (Steinwart & Christmann, 2008). Note that we do not assume any (restricted) strong convexity on the design. Remarkably, the convergence rate is determined by the true rank $d^*$ (not $d_{\max}$). This implies that the posterior of the rank based on the prior in Eq. (3) properly concentrates around the true rank. The second term $\frac{d^*}{n}\log\left(\frac{1}{\kappa}\right)$ represents the complexity of the model selection. This term is almost negligible as $n \to \infty$. Moreover, if we know $d^*$ beforehand and fix $d = d_{\max}$, then this term disappears. It will be shown that this convergence rate is actually minimax optimal (see Theorem 4 below).

We analyze the convergence rate more closely. To do so, let us consider a special case of the *Matérn prior*, and assume the domain of the input is a hypercube: $x^{(k)} \in [0,1]^{p_k}$. The Matérn prior is a Gaussian process prior corresponding to a kernel function that has a spectral density given as $\psi(s) = \frac{1}{(1+\|s\|^2)^{\alpha+p/2}}$, where $\alpha$ is a smoothness parameter and $p$ is the dimension of the input. It is known that the corresponding RKHS is included in a Sobolev space $W^{\alpha+p/2}[0,1]^p$ with the smoothness $\alpha + p/2$ (van der Vaart & van Zanten, 2011), and thus, the metric entropy exponent can be evaluated as $s \leq p/(2\alpha + p)$ (with high probability). We consider a simple situation where the Gaussian process prior on $f_r^{(k)}$ is the Matérn prior with the same smoothness parameter $\alpha$ for all $r, k$. Then, according to Theorem 1, we obtain the following convergence rate in this situation:

$$\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^*\|_n^2\right] \leq C\left\{\sum_{r=1}^{d^*}\sum_{k=1}^{K} n^{-\frac{1}{1+\frac{p_k}{2\alpha+p_k}}}\right\}.$$

This could be much smaller than the optimal convergence rate $O(n^{-\frac{1}{1+p/(2\alpha+p)}})$ for the nive estimation of $f^* \in W^{\alpha+p/2}[0,1]^p$ on the whole space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$ because the full dimension $p = \sum_{k=1}^K p_k$ is larger than individual dimension $p_k$. However, an estimation fully utilizing the nonlinear tensor product model in Eq. (1) can alleviate the curse of dimensionality.

### 4.2. Upper bound for misspecified setting

Here, we give a convergence rate in a situation where the true function is not necessarily included in the RKHS. In a practical situation, it is slightly demanding to assume that the true function is included *exactly* in our specified RKHS. For example, the RKHS corresponding to the Gaussian kernel is dense in $L_2(\mathcal{X})$, but the function space itself is much smaller than $L_2(\mathcal{X})$. Thus, we develop an extended analysis where such a *misspecified* situation is allowed. To handle a function $f : \mathcal{X} \to \mathbb{R}$, which might be outside an RKHS $\mathcal{H}$, we consider the following norm:

$$\|f\|_{\theta,\infty,\mathcal{H}} := \sup_{t>0} t^{-\theta}\left[\inf_{h\in\mathcal{H}}\{\|f-h\|_\infty + t\|h\|_{\mathcal{H}}\}\right],$$

where $0 < \theta \leq 1$ is a parameter. This norm defines a *real interpolation space*

$$[L_\infty, \mathcal{H}]_{\theta,\infty} := \{f : \mathcal{X} \to \mathbb{R} \mid \|f\|_{\theta,\infty,\mathcal{H}} < \infty\}.$$

We can check that $\mathcal{H} \hookrightarrow [L_\infty, \mathcal{H}]_{\theta,\infty} \hookrightarrow L_\infty$ if the corresponding kernel is bounded (where $\hookrightarrow$ represents continuous embedding) (Bennett & Sharpley, 1988). The parameter $\theta$ controls the size of the interpolation space as compared with the RKHS $\mathcal{H}$. In particular, $\theta = 1$ indicates $[L_\infty, \mathcal{H}]_{\theta,\infty} = \mathcal{H}$.

Accordingly, we relax Assumption 4 as follows.

**Assumption 5** *There exists $0 < \theta \leq 1$ such that $f_r^{*(k)} \in [L_\infty, \mathcal{H}_{(r,k)}]_{\theta,\infty}$ for all $r$ and $k$. There exists $R$ such that $\max_{(r,k)} \|f_r^{*(k)}\|_{\theta,\infty,\mathcal{H}_{(r,k)}} \leq R$. The true tensor is low rank, that is, there exists $d^*$ such that $f_r^{*(k)} = 0$ for all $r > d^*$ and $1 \leq k \leq K$.*

To control the infinity norm of the estimator, we replace Assumption 3 to the following one.

**Assumption 6** *If $1 - \theta - s_{(r,k)} \geq 0$, there exists $0 < \tilde{s}_{(r,k)} < 0$ and $C_1' \geq 0$ such that*

$$\log N(\mathcal{B}_{\mathcal{H}_{(r,k)}}, L, \|\cdot\|_\infty) \leq c_1' L^{-2\tilde{s}_{(r,k)}} \quad (L > 0). \tag{8}$$

*Otherwise, Assumption 3 is satisfied.*

Note that Eq. (8) implies $-\log(\mathrm{GP}_{r,k}(\{f : \|f\|_\infty \leq L\})) \leq C_1' L^{-\frac{2\tilde{s}_{(r,k)}}{1-\tilde{s}_{(r,k)}}}$ for some $C_1' > 0$ by the relation between a small ball probability and a metric entropy (Kuelbs & Li, 1993; Li & Shao, 2001). Under these assumptions, we obtain the following convergence rate.

**Theorem 2** *Suppose that Assumptions 1, 2, 5, and 6 are satisfied, and $\beta \geq 4\|g_\epsilon\|_\infty$. Then, there exists a constant $C > 0$ depending on $\beta$, $C_0$, $C_1'$ and $s_{(r,k)}$ such that*

$$\begin{aligned}&\mathrm{E}_{Y_{1:n}|x_{1:n}}\left[\|\hat{f} - f^*\|_n^2\right]\\ \leq& C'\left\{\left[\sum_{r=1}^{d^*}\left(\sum\nolimits_{k=1}^{K} n^{-\frac{1}{1+s_{(r,k)}/\theta}}\right)^{1/2}\right]^2 + \frac{d^*}{n}\log\left(\frac{1}{\kappa}\right)\right\}\end{aligned}$$

*where* $C' = CK\left[R + 2(R\vee 1)^{\frac{\max_{r,k}\{1-s_{(r,k)}\}}{\theta}}\right]^{2(K-1)\vee\max_{r,k}\{2s_{(r,k)}/\theta\}}$.

The proof is given in the supplementary material. Now, we again consider the Matérn prior on $\mathcal{X}_k = [0,1]^{p_k}$. It is shown that the interpolation space $[L_\infty(\mathcal{X}_k), W^\alpha(\mathcal{X}_k)]$ with respect to a Sobolev space $W^\alpha(\mathcal{X}_k)$ is included in a *Besov space* and satisfies the metric entropy condition (7), where $s_{(r,k)}$ is replaced with $s'_{(r,k)} = \frac{p_k}{2\alpha\theta} = \frac{s_{(r,k)}}{\theta}$ (Theorem 2 of Edmunds & Triebel (1996) and Section A.5.6 of Steinwart & Christmann (2008))). Thus, the convergence rate to estimate one function $f_r^{*(k)}$ by the ordinary kernel ridge regression is given by $n^{-\frac{1}{1+s_{(r,k)}/\theta}}$. In that sense, the convergence rate in Theorem 1 yields a more general result in a tensor estimation situation. Moreover, we do not need to know the parameter $\theta$ beforehand, but the Gaussian process estimator possesses adaptivity against the unknown parameter $\theta$.

**Remark 3** *A more general theorem that includes both Theorems 1 and 2 is given in the supplementary material (Theorem B.1), by using which it is possible to derive a* mixture *of both theorems; that is, some components $f_r^{*(k)}$ are included in the RKHS and the others are not. Moreover, a convergence rate for a linear model as given in Suzuki (2015) is also derived from the generalized theory.*

### 4.3. Minimax lower bound

Here, we give the minimax lower bound. To simplify the problem, we specify the structure of the problem. We assume that each component $x^{(k)} \in \mathcal{X}_k$ of the input $x = (x^{(1)}, \ldots, x^{(K)}) \in \mathcal{X}$ can be further decomposed as

$$x^{(k)} = (x_{(1,k)}, \ldots, x_{(d^*,k)}) \in \mathcal{X}_{(1,k)} \times \cdots \times \mathcal{X}_{(d^*,k)} = \mathcal{X}_k.$$

Then, each RKHS $\mathcal{H}_{(r,k)}$ takes $x_{(r,k)} \in \mathcal{X}_{(r,k)}$ as an input; that is, for any $f_r^{(k)} \in \mathcal{H}_{(r,k)}$, there is a function

$\tilde{f}_{(r,k)} : \mathcal{X}_{(r,k)} \to \mathbb{R}$ such that $f_r^{(k)}(x_k) = \tilde{f}_{(r,k)}(x_{(r,k)})$. We assume that the distribution of the input $x_k \in \mathcal{X}_k$ is a product measure $P_{\mathcal{X}_k} = P_{\mathcal{X}_{(1,k)}} \times \cdots \times P_{\mathcal{X}_{(d^*,k)}}$ and the distribution of the whole input $x = (x^{(1)}, \dots, x^{(K)}) \in \mathcal{X}$ is also a product of $P_{\mathcal{X}_k}$: $P_{\mathcal{X}} = P_{\mathcal{X}_1} \times \cdots \times P_{\mathcal{X}_K}$. We may assume that all functions $f_r^{(k)} \in \mathcal{H}_{(r,k)}$ have zero mean without loss of generality: $\mathrm{E}_{X \sim P_{\mathcal{X}_k}}[f_r^{(k)}(X)] = 0$. Then, by the set up of $P_{\mathcal{X}}$, we have that

$$\|f\|_{L_2(P_{\mathcal{X}})}^2 = \mathrm{E}_{X \sim P(X)}[f^2(X)] = \sum_{r=1}^{d^*} \prod_{k=1}^{K} \|f_r^{(k)}\|_{L_2(P_{\mathcal{X}_k})}^2$$

for $f = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f_r^{(k)}$ where $f_r^{(k)} \in \mathcal{H}_{(r,k)}$. Moreover, we assume that the noise is distributed from a normal distribution: $\epsilon_i \sim N(0, \sigma^2)$ (i.i.d.).

To simplify the analysis, we assume that the complexities of all RKHSs $\mathcal{H}_{(r,k)}$ are the same and have the following lower bound of the metric entropy.

**Assumption 7** *There exists a real value $0 < s < 1$ such that*

$$\log N(\mathcal{B}_{\mathcal{H}_{(r,k)}}, \epsilon, L_2(P_{\mathcal{X}_{(r,k)}})) \sim \epsilon^{-2s}. \tag{9}$$

*Moreover, the kernel function is bounded as $\sup_x k_{r,k}(x,x) \leq 1$, and there exists $c_1 > 0$ such that $\exists \widehat{f}_r^{(k)} \in \mathcal{B}_{\mathcal{H}_{(r,k)}}$ satisfying $\|\widehat{f}_r^{(k)}\|_{L_2(P_{\mathcal{X}_k})} \geq c_1$ for all $r, k$.*

Let $\mathcal{H}_{(r,k)}(R) := \{f \in \mathcal{H}_{(r,k)} \mid \|f\|_{\mathcal{H}_{(r,k)}} \leq R\}$ be the ball with radius $R$ in $\mathcal{H}_{(r,k)}$. Then, we define a set of tensors as

$$\mathscr{H}_{(d^*,K)}(R) := \left\{ f = \sum_{r=1}^{d^*} \prod_{k=1}^{K} f_r^{(k)} \;\middle|\; f_r^{(k)} \in \mathcal{H}_{(r,k)}(R) \right\}.$$

Under these settings, we have the following minimax optimal lower bound of the estimation error.

**Theorem 4** *If every $\mathcal{X}_k$ is a compact metric space, every $k_{(r,k)}$ is continuous, and the radius $R$ of the tensor set $\mathscr{H}_{(d^*,K)}(R)$ satisfies $R \geq \frac{1+c_1}{c_1}$, then there is a constant $C_1 > 0$ independent of $d^*, K, n$ such that*

$$\inf_{\hat{f}} \sup_{f^* \in \mathscr{H}_{(d^*,K)}(R)} \mathrm{E}[\|f - \hat{f}\|_{L_2(P_{\mathcal{X}})}^2] \geq C_1 d^* K n^{-\frac{1}{1+s}},$$

*where the inf is taken over all estimators $\hat{f}$.*

The proof is given in the supplementary material. In the proof, we utilize the information theoretic technique developed by Yang & Barron (1999). This theorem states that the learning rate of our Gaussian process tensor estimator (Theorem 1) is actually minimax-optimal up to constants.

## 5. Related work

Here, we discuss the relations and differences between our work and related work.

Our model is based on Signoretto et al. (2013). They proposed using an RKHS to model the nonlinear relations between several data sources $x^{(1)}, \dots, x^{(K)}$. Their proposed learning method is an alternating regularized least squares method. That is, $f_r^{(k)}$ is updated by minimizing the regularized empirical risk, while other components are fixed. However, there is no statistical guarantee of the alternating minimization in the context of nonparametric estimation.

Recently, independently of our study, Imaizumi & Hayashi (2016) developed a novel theory of a Bayes estimator for a product of functions on a tensor space. The point at which their study differs most widely is that they assume that the input $x = (x^{(1)}, \dots, x^{(K)})$ is given by the CP-decomposition of a certain tensor $\mathscr{X} \in \mathbb{R}^{p_1 \times \cdots \times p_K}$. Therefore, the input is restricted to an Euclidean vector and cannot be applied to more general input such as graphs and "tasks" as in multi-task learning. Moreover, CP-decomposition is not necessarily uniquely determined (at least there is a freedom of signs) (see Kolda & Bader (2009) for more details). The minimax lower bound was not given. In contrast, our method can be applied to any space where kernel functions can be defined, and our analysis shows the minimax optimality in the general setting.

Another relevant study is that of Suzuki (2015) in which a statistical convergence rate of a Bayes estimator for the *linear tensor model* Eq. (2) was given and it was shown that the Bayes estimator achieves the minimax optimal rate. However, his analysis is limited to the linear model. In contrast, our analysis is for a considerably more general nonlinear model in RKHS.

## 6. Numerical experiments

Here we numerically evaluate the practical performance of our Gaussian process tensor estimator. We applied our method to multi-task learning and conducted several experiments on three real datasets.

### 6.1. Multi-task learning

We executed our experiments on a nonlinear multi-task learning problem that is a nonlinear extension of multilinear multi-task learning (MLMTL) (Romera-Paredes et al., 2013; Wimalawarne et al., 2014). In MTMTL, several regression tasks are referenced by multiple indices. For example, if our objective is to predict the users' ratings of several aspects of movies, such as the story, cast, and overall quality, then a natural indexing of the tasks would be a two dimensional array represented by (movie, aspect). Here, as in this example, we considered multiple tasks that are aligned on a 2-dimensional space.

The original MLMTL was proposed as a special case of the linear tensor regression problem (2) (Romera-Paredes et al., 2013; Wimalawarne et al., 2014). We extend the model to a nonlinear regression problem which has a form of Eq. (1). Suppose that we observe an input-output pair $(\boldsymbol{w}_i, y_i)$ for some task $(p,q) \in \{p_1, \dots, p_{M_1}\} \times \{q_1, \dots, q_{M_2}\}$ as the $i$-th sample, where $\boldsymbol{w}_i$ is the input feature and $y_i$ is the corresponding label. Then we construct a concatenation $x_i = (x_i^{(1)}, x_i^{(2)}, x_i^{(3)}) = (p, q, \boldsymbol{w}_i)$ as an input for the nonlinear model (1), and the output $y_i$ is given as

$$
\begin{aligned}
y_i &= \sum_{r=1}^{d} \underbrace{f_r^{(1)}(x_i^{(1)})}_{f_r^{(1)}(p)=:\alpha_{r,p}} \times \underbrace{f_r^{(2)}(x_i^{(2)})}_{f_r^{(2)}(q)=:\beta_{r,q}} \times \underbrace{f_r^{(3)}(x_i^{(3)})}_{f_r^{(3)}(\boldsymbol{w}_i)=:f_r(\boldsymbol{w}_i)} + \epsilon_i \\
&= \textstyle\sum_{r=1}^{d} \alpha_{r,p}\beta_{r,q} f_r(\boldsymbol{w}_i) \; + \; \epsilon_i.
\end{aligned}
$$

If the function $f_r(\boldsymbol{w})$ is a linear function, then this model is reduced to the original MLMTL. An intuition behind this model is that the regression function is represented by a linear combination of a latent factor $f_r$ and its linear coefficient is given by the degree of relevance to each task. The multi-task learning is reduced to the estimation of $\alpha_{r,p}$, $\beta_{r,q}$, and $f_r$. In the following, we consider the input $\boldsymbol{w}_i$ as a feature vector in $\mathbb{R}^{M_3}$.

### 6.2. Real benchmark data sets

Here, we tested our proposed method with two real-world benchmark data sets, namely the Restaurant data set (Blanca et al., 2011) and the School data set (Goldstein, 1991). We compared our method with the following methods:

- Scaled latent convex regularization (scaled latent) (Wimalawarne et al., 2014): A state-of-the-art convex regularization method for MLMTL. The regularization parameter was chosen so that the test MSE is minimized.
- Alternating regularized least squares (ALS) (Signoretto et al., 2013): The method that alternatively updates $f_r^{(k)}$. We used the GRBF (Gaussian radial basis function) kernel for $f_r$. We chose the regularization parameter 50 (ALS(50)) and the best performance parameter (ALS(Best)).

The tensor rank for ALS and GP was fixed $d = 3$ in both data sets. For our Gaussian process method, we employed the linear kernel (linear) and the GRBF kernel (GRBF) as the kernel function for the GP on $f_r$. We also tested a mixture of them, i.e., some of three kernels for $f_r$ were the linear and the rest were the GRBF. We did this with the number of the linear kernels 1 and 2 (indicated by GRBF(2)+lin(1) and GRBF(1)+lin(2) respectively).

**Restaurant data set.** The Restaurant & Consumer Dataset (Blanca et al., 2011) is a dataset for a recommendation system used to predict consumer ratings given to different restaurants. Each of the $M_1 = 138$ consumers gave scores to restaurants from $M_2 = 3$ different aspects, i.e., food quality, service quality, and overall quality. Following the approach of (Romera-Paredes et al., 2013), we obtained $M_3 = 44$ features from descriptive attributes of the restaurants and modeled this as a multi-task learning problem where the objective was to predict a consumer's response to a restaurant given the features of that restaurant. The total number of instances for all the tasks was 3483. The kernel width $\sigma$ for the GRBF kernel was set at 100, and the delta kernel was chosen for the kernel functions for Task 1 (restaurant) and Task 2 (aspect). Figure 1a illustrates the MSEs against the sample size.

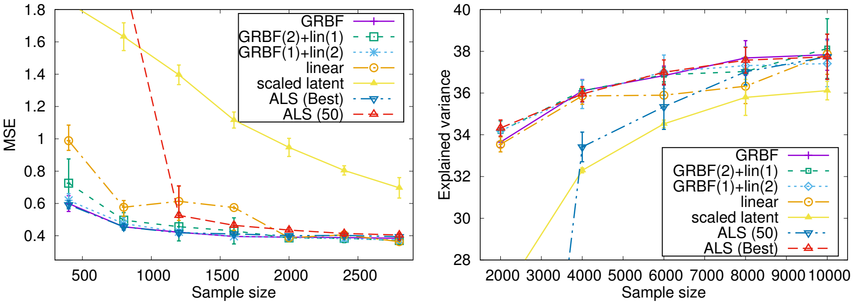
The scaled latent method performed much worse than the ALS and our method. The linear kernel of our method outperformed the scaled latent, even though both of them are linear model and it was numerically unstable when the proportion of observed samples was small. ALS (50) was not as good as the other nonlinear methods. This is because the regularization parameter was not optimally chosen. Overall, the nonlinear methods (ALS and GRBF) outperformed the linear models (linear and scaled latent) in small sample size when the parameters were well tuned. The GRBF of our method performed slightly better than the ALS method. Moreover, when the number of samples was above 2000, the mixture of the linear kernel and the GRBF kernel outperformed the results of the GRBF kernel.

**School data set.** The school data set (Goldstein, 1991) was taken from the Inner London Education Authority (ILEA) and consists of the examination records of 15362 students at 139 secondary schools in the years 1985, 1986, and 1987. The objective is to predict examination scores for students at schools in certain years based on student- and school-dependent inputs. Following Bakker & Heskes (2003), we obtained 44 features. Therefore, in this setting, $M_1 = 139$ (school), $M_2 = 3$ (year), and $M_3 = 44$. The kernel width $\sigma$ for the GRBF kernel was set at 15. Following Wimalawarne et al. (2014), we used the percentage of explained variance, $100 \cdot (\text{test MSE})/(\text{variance of } y)$, as the evaluation metric. Figure 1b shows the expected variances against the sample size.

Here again, the performance of the scaled latent was worst. We can see that the best performance was achieved by methods involving the GRBF kernel such as GRBF, linear-and-GRBF-mixiture, and ALS (Best).

### 6.3. Prediction of online shopping sales

Next, we apply our method to an online shopping dataset which is a collection of consumer purchase histories on



(a) Restaurant data

(b) School data

*Figure 1.* Predictive accuracy comparisons between the scaled latent method (scaled latent), the alternating regularized lease squares (ALS), and the Gaussian process method with different kernels in the restaurant and the school datasets.

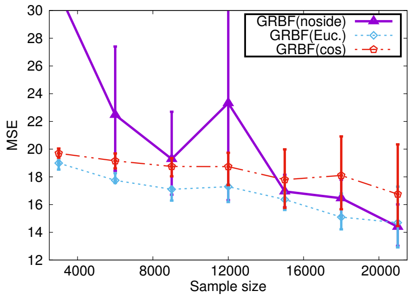


*Figure 2.* Predictive accuracy (MSE) of the Gaussian process method with GRBF on shops for online shopping prediction. Different kernels on shops are tested.

Yahoo! Japan shopping. The data set also contains the identities of its registered users: age, gender, industry type of their occupation, and so on. The website employs a shopping mall type system where customers can purchase goods offered by different shops in one place. We selected 100 different products and 570 shops to set up a multi-task learning problem in which the objective was to predict the quantity of a product which a consumer would buy at a particular shop given the features of that consumer. 65 features were obtained by user identities and the purchase time.

We investigated the effect of kernels on tasks by comparing three kernels on shops: the delta kernel (no side information about shops; (noside)) and two kernels with side information ((Euc.) and (cos) respectively). To exploit side information, we constructed a weighted and undirected graph $G$ with its vertices $V$ representing shops. Each node had a real-valued vector whose elements expressed demographic characteristics, such as the percentage of female customers, which were calculated using purchase histories including records of other products. The Laplacian matrix $L$ of $G$ is defined as the matrix with $L_{i,j} = \left(\sum_{j\in V} w_{i,j}\right)\delta_{i,j} - w_{i,j}$ where $w_{i,j}$ is the weight between two shops $i$ and $j$. Then we constructed the commute-time kernel $K = L^\dagger$ (Fouss et al., 2007) for the kernel on shops, where † denotes the psuedoinverse matrix. The weight $w_{i,j}$ was calculated in two manners: the "similarity" weight which is the cosine similarity (cos) and the "anti-similarity" weight which is the Euclidean distance (Euc.) between two shops. In contrast to the "cos" method, the "Euc." method puts a large weight between *dissimilar* shops. It was observed that in the data that some customers stay at their own favorite shops among a similar shops. By using the "anti-similarity" regularization, we can exclude the effect from the *similar but not the same* shops but includes information from shops in a different category. Finally, we chose the linear kernel and the GRBF kernel for the GP on $f_r$, and the delta kernel for products.

Figure 2 shows the MSE on the test data against the sample size. The MSE of the linear kernel was above 100, which was much worse than that of the GRBF kernel, and thus it was omitted. We can see that the kernels with the side-information performs better than the kernel without side-information in the small sample size region. In particular, the anti-similarity kernel always outperforms the others. This results indicates the effectiveness of using non-linear kernels on a tensor learning problem.

## 7. Conclusion and discussion

In this paper, we analyzed a nonparametric tensor learning method based on the Gaussian process technique. The statistical convergence rate was derived for both correctly specified and misspecified settings. It was shown that the derived bound is minimax optimal up to constants. Moreover, we applied our method to multi-task learning and showed that our method outperformed the existing convex optimization method based on the linear model.

The alternating least squares method (ALS) performed almost comparably to the Gaussian process method. This is not so surprising because ALS can be interpreted as a "non-stochastic" method of the Gaussian process approach. It would be an important future work to analyze the statistical properties of ALS.

## Acknowledgment

This work was partially supported by MEXT kakenhi (25730013, 25120012, 26280009, 15H01678 and 15H05707), JST-PRESTO and JST-CREST.

## References

P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

N. Aronszajn. Theory of reproducing kernels. *Transac-*

*tions of the American Mathematical Society*, 68:337–404, 1950.

M. T. Bahadori, Q. R. Yu, and Y. Liu. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *Advances in Neural Information Processing Systems 27*, pp. 3491–3499. 2014.

B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

C. Bennett and R. Sharpley. *Interpolation of Operators*. Academic Press, Boston, 1988.

V.-G. Blanca, G.-S. Gabriel, and P.-M. Rafael. Effects of relevant contextual features in the performance of a restaurant recommender system. In *Proceedings of 3rd Workshop on Context-Aware Recommender Systems*, 2011.

O. Catoni. *Statistical Learning Theory and Stochastic Optimization*. Lecture Notes in Mathematics. Springer, 2004. Saint-Flour Summer School on Probability Theory 2001.

W. Chu and Z. Ghahramani. Probabilistic models for incomplete multi-dimensional arrays. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5 of *JMLR Workshop and Conference Proceedings*, 2009.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, 72:39–61, 2008.

D. E. Edmunds and H. Triebel. *Function Spaces, Entropy Numbers, Differential Operators*. Cambridge University Press, Cambridge, 1996.

F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.

S. Gandy, B. Recht, and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

H. Goldstein. Multilevel modelling of survey data. *The Statistician*, pp. 235–244, 1991.

F. L. Hitchcock. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics*, 6:164–189, 1927a.

F. L. Hitchcock. Multilple invariants and generalized rank of a p-way matrix or tensor. *Journal of Mathematics and Physics*, 7:39–79, 1927b.

M. Imaizumi and K. Hayashi. Nonparametric tensor regression with low-rank decomposition. In *the 33rd International Conference on Machine Learning (ICML)*, 2016. to appear.

A. Karatzoglou, X. Amatriain, L. Baltrunas, and N. Oliver. Multiverse recommendation: N-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys)*, pp. 79–86, 2010.

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

J. Kuelbs and W. V. Li. Metric entropy and the small ball problem for Gaussian measures. *Journal of Functional Analysis*, 116(1):133–157, 1993.

W. V. Li and Q.-M. Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Stochastic Processes: Theory and Methods*, 19:533–597, 2001.

J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In *Proceedings of the 12th International Conference on Computer Vision (ICCV)*, pp. 2114–2121, 2009.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pp. 230–234, 1998.

D. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Anual Conference on Computational Learning Theory (COLT)*, pp. 164–170, 1999.

P. Rai, Y. Wang, S. Guo, G. Chen, D. Dunson, and L. Carin. Scalable Bayesian low-rank decomposition of incomplete multiway tensors. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 1800–1808, 2014.

P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 1444–1452, 2013.

W. Shen and S. Ghosal. Adaptive Bayesian density regression for high-dimensional data. *Bernoulli*, 22(1):396–420, 02 2016.

M. Signoretto, L. D. Lathauwer, and J. Suykens. Nuclear norms for tensors and their use for convex multilinear estimation. Technical Report 10-186, ESAT-SISTA, K.U.Leuven, 2010.

M. Signoretto, L. D. Lathauwer, and J. A. K. Suykens. Learning tensors in reproducing kernel Hilbert spaces with multilinear spectral penalties. *CoRR*, abs/1310.4977, 2013.

I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, 2008.

T. Suzuki. Convergence rate of Bayesian tensor estimator and its minimax optimality. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1273–1282, 2015.

R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *Advances in Neural Information Processing Systems 26*, pp. 1331–1339, 2013. NIPS2013.

R. Tomioka, T. Suzuki, K. Hayashi, and H. Kashima. Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems 24*, pp. 972–980, 2011. NIPS2011.

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, 1996.

A. W. van der Vaart and J. H. van Zanten. Information rates of nonparametric Gaussian process methods. *Journal of Machine Learning Research*, 12:2095–2119, 2011.

K. Wimalawarne, M. Sugiyama, and R. Tomioka. Multitask learning meets tensor factorization: task imputation via convex optimization. In *Advances in Neural Information Processing Systems 27*, pp. 2825–2833. 2014. NIPS2011.

L. Xiong, X. Chen, T.-K. Huang, J. Schneider, and J. G. Carbonell. Temporal collaborative filtering with bayesian probabilistic tensor factorization. In *Proceedings of the 2010 International Conference on SIAM Data Mining*, pp. 211–222, 2010.

Z. Xu, F. Yan, and Y. A. Qi. Inftucker: t-process based infinite tensor decomposition. *CoRR*, abs/1108.6296, 2011.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *The Annals of Statistics*, 27(5):1564–1599, 1999.