Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Multi-Dimensional Domain Generalization with Low-Rank Structures

Sai Li & Linjun Zhang

To cite this article: Sai Li & Linjun Zhang (11 Apr 2025): Multi-Dimensional Domain Generalization with Low-Rank Structures, Journal of the American Statistical Association, DOI: 10.1080/01621459.2025.2471055

To link to this article: https://doi.org/10.1080/01621459.2025.2471055

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2025, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2025.2471055

Taylor & Francis
Taylor & Francis Group

Multi-Dimensional Domain Generalization with Low-Rank Structures

Sai Li[a] and Linjun Zhang[b]

[a]Institute of Statistics and Big Data, Renmin University of China, Beijing, China; [b]Department of Statistics, Rutgers University, New Brunswick, NJ

ABSTRACT
In conventional statistical and machine learning methods, it is typically assumed that the test data are identically distributed with the training data. However, this assumption does not always hold, especially in applications where the target population are not well-represented in the training data. This is a notable issue in health-related studies, where specific ethnic populations may be underrepresented, posing a significant challenge for researchers aiming to make statistical inferences about these minority groups. In this work, we present a novel approach to addressing this challenge in linear regression models. We organize the model parameters for all the sub-populations into a tensor. By studying a structured tensor completion problem, we can achieve robust domain generalization, that is, learning about sub-populations with limited or no available data. Our method novelly leverages the structure of group labels and it can produce more reliable and interpretable generalization results. We establish rigorous theoretical guarantees for the proposed method and demonstrate its minimax optimality. To validate the effectiveness of our approach, we conduct extensive numerical experiments and a real data study focused on diabetes prediction for multiple subgroups, comparing our results with those obtained using other existing methods. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received September 2023
Accepted February 2025

KEYWORDS
Heterogeneous data; Tensor completion; Transfer learning; Zero-shot learning

1. Introduction

Conventional machine learning methods typically assume that the test data, sampled from a target distribution, are well-represented in the training domains. However, in many practical scenarios, data from the target domains can be scarce or completely unseen during the training phase. A prominent example of this occurs in biomedical research: different clinical centers may employ varied medical devices and serve diverse patient demographics, leading to significant discrepancies between the training and test distributions (Zhang et al. 2023). Such scenarios introduce multiple levels of heterogeneity between the test and training domains, causing the test data distributed differently from the training data. To address this challenge, the field of *domain generalization* has emerged. Also referred to as out-of-distribution generalization or zero-shot domain adaptation, domain generalization aims at developing models that can effectively generalize to new and even unseen populations that are not represented in the training domains (Wang et al. 2022; Zhou et al. 2022).

Domain generalization has attracted significant attention from various disciplines, including computer vision, healthcare, and biological studies (Hendrycks et al. 2021; Lotfollahi et al. 2021; Sharifi-Noghabi et al. 2021). Consider a typical example where the goal is to train a classification model to distinguish between cats and dogs based on images. If the training data consists of images in cartoon or painting styles, while the test data comprises real photos, there is a clear divergence between the style features in the training and test domains. In such cases, a classification rule trained on the artistic images may not perform well when applied to real photos. In health-related studies, this issue becomes even more critical. Populations with different demographic features, such as gender and race, may exhibit distinct biological mechanisms underlying diseases (Woodward et al. 2022). For instance, the polygenic risk score (PRS) for schizophrenia showed significant variation among ancestral groups with substantially higher scores observed in African subjects compared to European subjects from HapMap (Curtis 2018). Moreover, certain sub-populations may be underrepresented in medical databases, exacerbating the challenge of ensuring that predictive models, trained on well-represented groups, are generalizable and beneficial to those underrepresented populations. This underscores the urgent need for reliable domain generalization methods.

1.1. Main Results

In this work, we consider a multi-task linear regression framework. Suppose that we have observations $(\boldsymbol{x}_i^\top, y_i, \boldsymbol{g}_i^\top)$, for $i = 1, \ldots, N$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ represents the vector of covariates, $y_i \in \mathbb{R}$ is the response, and $\boldsymbol{g}_i \in \mathcal{G}$ is a q-dimensional group, task, or environment index. Without loss of generality, the whole set of group indices can be expressed as $\mathcal{G} = [p_1] \circ \cdots \circ [p_q]$, where "$\circ$" denotes the Cartesian product. For example, a two-dimensional ($q = 2$) group index might comprise gender and

CONTACT Sai Li saili@ruc.edu.cn Institute of Statistics and Big Data, Renmin University of China, Beijing, 100872, China.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

race indicators. If $\boldsymbol{g}_i = \boldsymbol{g}_{i'}$, then the ith and i'th individuals belong to the same group.

For each group, we consider the following linear model:

$$\mathbb{E}[y_i|\boldsymbol{x}_i, \boldsymbol{g}_i = \boldsymbol{g}] = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(g)}, \quad \forall g \in \mathcal{G}, \tag{1}$$

where $\boldsymbol{\beta}^{(g)} \in \mathbb{R}^p$ denotes the coefficient vector for group g. However, data is available only for a subset of groups. Denote the indices of observed groups by $\mathcal{O}$ for $\mathcal{O} \subseteq \mathcal{G}$. Let $n^{(g)}$ represent the sample size for group g. By definition, $n^{(g)} > 0$ for each $g \in \mathcal{O}$, and $n^{(g)} = 0$ for each $g \notin \mathcal{O}$. If g is the gender and race indicator, then $\mathcal{O}$ may contain well-represented groups and $\mathcal{G} \setminus \mathcal{O}$ main contain underrepresented groups. Our primary objective is to establish prediction rules for some unseen domains $g \notin \mathcal{O}$.

We propose to organize the coefficient vectors from multiple groups as a high-order tensor and develop new tensor completion methods tailored for domain generalization. Notably, our coefficient tensor presents structured missing patterns, which is different from commonly studied random missing tensor completion scenarios. To this end, we present a novel algorithm named "TensorDG", which stands for **Tensor** completion-based algorithm for **D**omain **G**eneralization. We further establish the convergence rates of our proposal and show that it is minimax optimal for estimating $\{\boldsymbol{\beta}^{(g)}\}_{g\in\mathcal{G}}$ in tensor Frobenius norm under mild conditions. Additionally, we introduce 'TensorTL', which extends our core methodology for transfer learning when the target domain possesses limited samples.

We highlight two key features of our proposal. First, it leverages the group structures rather than simply labeling the observed groups as $1, \dots, |\mathcal{O}|$ as in many existing works on multi-task learning and domain generalization. This structural information is helpful for understanding the similarity of different domains and further sheds light on devising more explainable and reliable domain generalization methods. Second, our proposal has solid theoretical guarantees and enjoys minimax optimality under mild conditions. Moreover, by employing the rank determination techniques (Han, Chen, and Zhang 2022), practitioners can ascertain the degree to which the model may be misspecified, adding another layer of reliability to the method.

1.2. Related Literature

Multi-task learning. Recently, multi-task regression and transfer learning have been widely studied in various models (Du et al. 2020; Lee et al. 2021; Li, Cai, and Li 2022). Many existing works follow the idea to first estimate some shared representations based on all the training samples and then calibrate the model using the data from the target domain. Our framework in (1) reduces to the multi-task learning problem when $\mathcal{G} = \mathcal{O}$, that is, all the domains have some training data available. We focus on the more challenging scenario where $\mathcal{O} \subsetneq \mathcal{G}$, that is, some target domains are not observed at all. Therefore, the multi-task learning methods cannot be directly used for domain generalization. An example and further comparisons are given in (4).

Domain generalization. Existing literature has studied identifying causal features for domain generalization, that is, the features that are responsible for the outcome but are independent of the unmeasured confounders in each domain. Identification of causal features has been connected with the estimation of invariant representations (Bühlmann 2020). Rojas-Carulla et al. (2018) propose methods to find invariant representations for domain generalization without theoretical guarantees. Chen and Bühlmann (2020) propose new estimands with theoretical guarantees under linear structural equation models. Pfister et al. (2021) investigate so-called stable blankets for domain generalization but the proposed algorithm does not have theoretical guarantees.

Beyond the causal framework, other popular domain generalization methods include Maximin estimator, self-training, and invariant risk minimization. To name a few, distributionally robust optimization (Volpi et al. 2018; Sagawa et al. 2019) or Maximin estimator (Meinshausen and Bühlmann 2015; Guo 2023) minimizes the max prediction errors among the training groups. Kumar, Ma, and Liang (2020) study the theoretical properties of self-training with gradual shifts. Baktashmotlagh et al. (2013) propose a domain invariant projection approach by extracting the information that is invariant across the source and target domains but it lacks theoretical guarantees. Wimalawarne, Sugiyama, and Tomioka (2014), Li et al. (2017), and Feng, Han, and Du (2021) use the low-rank matrix or tensor for domain generalization in deep neural networks. However, these methods are either purely empirical or computationally demanding, lacking of statistical optimality guarantee with efficient algorithms. In the realm of invariant predictors, Arjovsky et al. (2019) introduce invariant risk minimization, with extended discussions available in Rosenfeld, Ravikumar, and Risteski (2020), Zhou et al. (2022), and Fan et al. (2023). In contrast to our work, the works mentioned above do not consider the structural information of group labels and the generalizability is simply based on model assumptions. Our model leverages the structure of group indices which better explains why and how the model generalizes. It is also possible to diagnose whether our model assumptions fail or not.

Tensor completion. Our work is also closely related to tensor estimation and completion, which has significantly advanced in recent years (Bi et al. 2021). Montanari and Sun (2018) study tensor completion with random missing patterns in the noiseless setting. When having noisy entries, Zhang (2019) study tensor completion under low-rank assumptions with structural missing. Xia, Yuan, and Zhang (2021) study noisy tensor completion with random missing patterns under low-rank assumptions. The tensor completion problem can also be rewritten as a tensor regression model whose design consists of indicator functions. Chen, Raskutti, and Yuan (2019) study projected gradient descent for tensor regression and Zhang et al. (2020) propose a minimax optimal method for low-rank tensor regression with independent Gaussian designs. Mu et al. (2014) and Raskutti, Yuan, and Chen (2019) study tensor recovery with convex regularizers without and with noises, respectively. From the application perspective, tensor models have been widely used in recommender systems and modeling biomedical image data (Adomavicius and Tuzhilin 2010; Zhou, Li, and Zhu 2013).

1.3. Organization and Notation

In the rest of this article, we introduce the low-rank tensor model for multi-task regression in Section 2. In Section 3, we present the rationale of the proposed method and introduce the formal algorithm. We provide theoretical guarantees for the proposed method in Section 4 and discuss extensions of the main proposal in Section 5. In Section 6, we demonstrate the numerical performance of our proposals. In Section 7, we apply the proposed method to predict diabetes for people in different age and education groups. We conclude this article with discussions in Section 8. The proofs are all given in the supplementary materials.

For a generic matrix $T \in \mathbb{R}^{p_1\times p_2}$, let $\|T\|_2$ denote its spectral norm, $\|T\|_{2,\infty}$ denote $\max_{j\le p_2}\|T_{.,j}\|_2$, and $\|T\|_{\infty,2}$ denote $\max_{j\le p_1}\|T_{j,.}\|_2$. For a generic semi-positive definite matrix $\Sigma \in \mathbb{R}^{m\times m}$, let $\Lambda_k(\Sigma)$, $\Lambda_{\max}(\Sigma)$, $\Lambda_{\min}(\Sigma)$, and $\mathrm{Tr}(\Sigma)$ denote the kth, largest, smallest singular values, and trace of Σ, respectively. For a generic tensor $M \in \mathbb{R}^{p_1\times\cdots\times p_q}$, let $\|M\|_{\ell_2}$ denote the vectorized ℓ_2-norm of M, which is also known as tensor Frobenius norm. For a generic set A, let $|A|$ denote the cardinality of A. Let $a\vee b$ denote $\max\{a,b\}$ and $a\wedge b$ denote $\min\{a,b\}$. We use $c, c_0, c_1,\dots$ to denote generic constants which can be different in different statements. Let $a_n = O(b_n)$ and $a_n \lesssim b_n$ denote $|a_n/b_n| \le c$ for some constant c when n is large enough.

2. Set-Up and Data Generation Model

In this section, we outline the basic concepts related to the low-rank tensor model and establish its connection with the domain generalization problem.

2.1. Low-Rank Tensor Model

Invoking Section 1.1, the observed data can be reshaped as $(X^{(g)}, \boldsymbol{y}^{(g)}) \in \mathbb{R}^{n^{(g)}\times(p+1)}$, where each row corresponds to a sample $(\boldsymbol{x}_i^\top, y_i)$ with group label $\boldsymbol{g}_i = g$ for each $g \in \mathcal{O}$. For each $g \notin \mathcal{O}$, let $\boldsymbol{x}^{(g)}$ and $y^{(g)}$ denote the design and response variables generated from the oracle model for domain g.

Condition 1 (Data generating process). Suppose that

$$y_i^{(g)} = (\boldsymbol{x}_i^{(g)})^\top\boldsymbol{\beta}^{(g)} + \epsilon_i^{(g)},\ i \in [n^{(g)}],\ \forall g \in \mathcal{O},$$
$$\mathbb{E}[y^{(g)}|\boldsymbol{x}^{(g)}] = (\boldsymbol{x}^{(g)})^\top\boldsymbol{\beta}^{(g)},\ \forall g \in \mathcal{G}\setminus\mathcal{O}, \tag{2}$$

where $\epsilon_i^{(g)}$ are the independent random noises such that $\mathbb{E}[\epsilon_i^{(g)}|\boldsymbol{x}_i^{(g)}] = 0$ for each $g \in \mathcal{O}$. We assume that $\epsilon_i^{(g)}$ is independent of $\epsilon_{i'}^{(g')}$ for any $g \neq g' \in \mathcal{O}$, $i \in [n^{(g)}]$, $i' \in [n^{(g')}]$.

Each group in $\mathcal{G}$ has distinct model parameters $\boldsymbol{\beta}^{(g)}$. The noises in different domains are independent. For the unseen groups $g \in \mathcal{G}\setminus\mathcal{O}$, we only make assumptions on the conditional mean models as there are no samples.

We arrange the regression coefficients $\{\boldsymbol{\beta}^{(g)}\}_{g\in\mathcal{G}}$ into a tensor $\boldsymbol{\beta}(\mathcal{G}) \in \mathbb{R}^{p\times p_1\times p_2\times\cdots\times p_q}$ such that

$$\{\boldsymbol{\beta}(\mathcal{G})\}_{j,i_1,\dots,i_q} = \beta_j^{(\mathcal{G}_{i_1,\dots,i_q})} = \beta_j^{(i_1,\dots,i_q)}.$$

That is, the first mode of $\boldsymbol{\beta}(\mathcal{G})$ represents the regression coefficients and the remaining q modes represent group indices. We refer to the first mode of $\boldsymbol{\beta}(\mathcal{G})$ as the "coefficient mode" and the last q modes as the "group modes".

Denote the Tucker rank of tensor $\boldsymbol{\beta}(\mathcal{G})$ as $(r_0, r_1, \cdots, r_q)$, where r_t can be unknown a priori. The Tucker rank is defined based on matrix unfolding. Specifically, the unfolding $\mathcal{M}_t[\mathcal{X}]$ maps a tensor $\mathcal{X} \in \mathbb{R}^{h_0\times\cdots\times h_q}$ into a matrix $\mathcal{M}_t[\mathcal{X}] \in \mathbb{R}^{h_t\times(\prod_{0\le s\neq t\le q} h_s)}$ such that

$$(\mathcal{M}_t[\mathcal{X}])_{i_t,j} = X_{i_0,i_1,\dots,i_q},\ \text{for } j = 1 + \sum_{0\le l\le q, l\neq t}(i_l - 1)J_l \text{ and}$$
$$J_l = \prod_{0\le m\le l-1, m\neq t} h_m. \tag{3}$$

The Tucker rank $(r_0, r_1, \dots, r_q)$ is defined such that $\mathrm{rank}(\mathcal{M}_t[\boldsymbol{\beta}(\mathcal{G})]) = r_t$. We illustrate the implications of the low-rankness in the following example.

If $q = 2$, the order-3 tensor $\boldsymbol{\beta}(\mathcal{G})$ can be unfolded as

$$\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{G})] = \begin{pmatrix}\boldsymbol{\beta}^{(1,1)} & \dots & \boldsymbol{\beta}^{(p_1,p_2)}\end{pmatrix} \in \mathbb{R}^{p\times(p_1p_2)},$$
$$\mathcal{M}_1[\boldsymbol{\beta}(\mathcal{G})] = \begin{pmatrix}(\boldsymbol{\beta}^{(1,1)})^\top & \dots & (\boldsymbol{\beta}^{(1,p_2)})^\top \\ & \dots & \\ (\boldsymbol{\beta}^{(p_1,1)})^\top & \dots & (\boldsymbol{\beta}^{(p_1,p_2)})^\top\end{pmatrix},$$
$$\mathcal{M}_2[\boldsymbol{\beta}(\mathcal{G})] = \begin{pmatrix}(\boldsymbol{\beta}^{(1,1)})^\top & \dots & (\boldsymbol{\beta}^{(p_1,1)})^\top \\ & \dots & \\ (\boldsymbol{\beta}^{(1,p_2)})^\top & \dots & (\boldsymbol{\beta}^{(p_1,p_2)})^\top\end{pmatrix}. \tag{4}$$

The low-rank nature of $\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{G})]$ implies that $\boldsymbol{\beta}^{(g)} = B^*\boldsymbol{\alpha}^{(g)}$ for some basis matrix $B^* \in \mathbb{R}^{p\times r_0}$ and a latent score vector $\boldsymbol{\alpha}^{(g)} \in \mathbb{R}^{r_0}$. Similar low-rank assumptions for linear coefficients have been adopted by Du et al. (2020) and Tripuraneni, Jin, and Jordan (2021) in the context of transfer learning for linear models. As $\boldsymbol{\alpha}^{(g)}$ is task-specific, it needs to be learned based on a certain amount of target data. In domain generalization, however, $\boldsymbol{\alpha}^{(g)}$ cannot be directly estimated when there are no samples available from the target domain g. Fortunately, the low Tucker rank structure suggests that the matrices $\mathcal{M}_1[\boldsymbol{\beta}(\mathcal{G})]$ and $\mathcal{M}_2[\boldsymbol{\beta}(\mathcal{G})]$ also possess a low-rank nature. Such a correlation structure among different groups enables the possibility of domain generalization or zero-shot learning.

While we focus on the Tucker rank, another common metric of tensor rank is the canonical polyadic (CP) rank (Hitchcock 1927). Let us denote the CP rank of $\boldsymbol{\beta}(\mathcal{G})$ as $R(\boldsymbol{\beta}(\mathcal{G}))$. It holds that $\max_{0\le t\le q} r_t \le R(\boldsymbol{\beta}(\mathcal{G})) \le \prod_{t=0}^q r_t/(\max_{0\le t\le q} r_t)$. As a result, a tensor with low CP rank will imply a low Tucker rank structure. Hence, we focus on the Tucker rank characterization in this work.

Additionally, we define the mode product as follows. Let $\mathcal{X} \in \mathbb{R}^{h_0\times\cdots\times h_q}$ denote a generic $(q+1)$-order tensor. For $E_t \in \mathbb{R}^{h_t\times m_t}$ and $t = 0,\dots,q$, the tth mode product $\mathcal{X}\times_t E_t \in \mathbb{R}^{h_0\times\cdots\times h_{t-1}\times m_t\times h_{t+1}\times\cdots\times h_q}$ is defined as

$$\{\mathcal{X}\times_t E_t\}_{i_0,i_1,\dots,i_q} = \sum_{s=1}^{h_t}\{\mathcal{X}\}_{i_0,\dots,i_{t-1},s,i_{t+1},\dots,i_q}\{E_t\}_{s,i_t}$$

for $i_0 \in [h_0], \dots, i_q \in [h_q]$ and $i_t \in [m_t]$.

2.2. Observed Group Structures

To recover the tensor $\boldsymbol{\beta}(\mathcal{G})$, the observed groups $\mathcal{O}$ need to contain some crucial elements. For $t = 1, \ldots, q$, the *arm* set for mode t is defined as

$$\mathcal{A}_t = \underset{1\leq k\neq t\leq q}{\circ} S_k \text{ such that } S_k \subseteq [p_k] \text{ and } \\ S_1 \circ \cdots \circ S_{t-1} \circ [p_t] \circ S_{t+1} \circ \cdots \circ S_q \subseteq \mathcal{O}, \qquad (5)$$

where $\mathcal{A}_t$ is a $(q-1)$-dimensional set. For example, if $\boldsymbol{a} \in \mathcal{A}_1$, then $(1, \boldsymbol{a}^\top), \ldots, (p_1, \boldsymbol{a}^\top)$ are all elements of $\mathcal{O}$, that is, $(1, \boldsymbol{a}^\top), \ldots, (p_1, \boldsymbol{a}^\top)$ are all observed groups. To ease our notation, for $\mathcal{A}_t$ defined in (5), we denote the product $S_1 \circ \cdots \circ S_{t-1} \circ [p_t] \circ S_{t+1} \circ \cdots \circ S_q$ as $\mathcal{A}_t \circ_t [p_t]$. Further, we define the *body* set as

$$\Omega = \underset{1\leq t\leq q}{\circ} \Omega_t \text{ such that } \Omega_t \subseteq [p_t] \text{ and } \Omega \subseteq \mathcal{O}. \qquad (6)$$

As a consequence, it holds that

$$\mathcal{O} \supseteq \Omega \cup (\mathcal{A}_1 \circ_1 [p_1]) \cup \cdots \cup (\mathcal{A}_q \circ_q [p_q]). \qquad (7)$$

That is, the observed groups should consist of a body set and q arm sets (Figure 1). Without loss of generality, we assume $\mathcal{A}_t \subseteq \Omega_{-t}$ for $\Omega_{-t} = \underset{1\leq s\neq t\leq q}{\circ} \Omega_s$. Otherwise, if $\mathcal{A}_t \supset \Omega_{-t}$, then we can treat $\mathcal{A}_t \circ_t \Omega_t$ as the body set which is larger than Ω and Condition 2 sequel is easier to be satisfied.

To summarize, the missing data pattern described in (7) is structural, distinguishing it from missing at random. Xia, Yuan, and Zhang (2021) studies minimax optimal tensor completion methods with missing completely at random (MCAR) among many others. The MCAR assumption cannot be directly applied in our problem, given that $\boldsymbol{\beta}(\mathcal{G})$ has no missingness in the 0th mode (Figure 1).

3. Method

We first outline the rationale of our proposal based on Tucker decomposition in Section 3.1. The formal algorithm is provided in Section 3.2.

3.1. Rationale from Tucker Decomposition

For $\mathcal{S} = S_1 \circ \cdots \circ S_q$ with $S_t \subseteq [p_t]$, $t = 1, \ldots, q$, let $\boldsymbol{\beta}(\mathcal{S}) = \{\boldsymbol{\beta}(\mathcal{G})\}_{S_1,\ldots,S_q,\cdot} \in \mathbb{R}^{|S_1|\times\cdots\times|S_q|\times p}$. We make the following assumption on the whole tensor.

Condition 2 (Overall low-rank structure). The tensor $\boldsymbol{\beta}(\mathcal{G})$ has Tucker rank $(r_0, r_1, \ldots, r_q)$ and (7) holds. Moreover, $\mathcal{M}_t[\boldsymbol{\beta}(\Omega)]$ and $\mathcal{M}_t[\boldsymbol{\beta}(\mathcal{A}_t \circ_t \Omega_t)]$ both have rank r_t for $t = 1, \ldots, q$ with finite q, and $\mathcal{M}_0[\boldsymbol{\beta}(\Omega)]$ has rank r_0.

Under Condition 2, our target tensor $\boldsymbol{\beta}(\mathcal{G})$ has a Tucker decomposition

$$\boldsymbol{\beta}(\mathcal{G}) = \boldsymbol{\beta}(\Omega) \times_0 R_0 \times_1 R_1 \times_2 \cdots \times_q R_q \in \mathbb{R}^{p\times p_1\times\cdots\times p_q}, \quad (8)$$

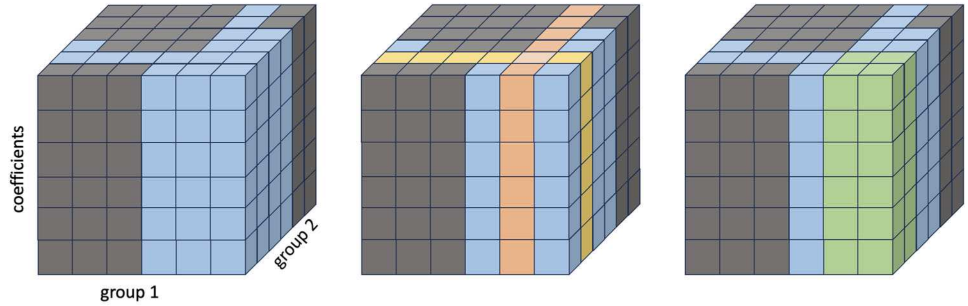
where $R_0 \in \mathbb{R}^{p\times p}$ and $R_t \in \mathbb{R}^{|\Omega_t|\times p_t}$, $t = 1, \ldots, q$ are computed based on $\boldsymbol{\beta}(\mathcal{A}_t \circ_t [p_t])$ as in the forthcoming (9). Hence, the full tensor can be recovered by using a subset of groups $\mathcal{O}$ satisfying (7), which only involves a small number of groups. As Condition 2 can be violated in practice, we will discuss the model diagnostics at the end of Section 4. In the supplement (Section A.1), we provide a sufficient condition which guarantees Condition 2 when the group indices are randomly revealed.

We now provide more details about R_t in (8). For $t = 1, \ldots, q$, let $\omega_t = |\Omega_t|$ and $a_t = |\mathcal{A}_t|$. Define $B_t^{(jo)} = \mathcal{M}_t^\top[\boldsymbol{\beta}(\mathcal{A}_t \circ_t \Omega_t)] \in \mathbb{R}^{(a_t p)\times\omega_t}$ as the joint measurements for mode t and $B_t^{(ar)} = \mathcal{M}_t^\top[\boldsymbol{\beta}(\mathcal{A}_t \circ_t [p_t])] \in \mathbb{R}^{(a_t p)\times p_t}$ as the arm measurements for mode t. For $t = 0$, especially, we define $B_0^{(jo)} = B_0^{(ar)} = (\boldsymbol{\beta}^{(g_1)} \ \ldots \ \boldsymbol{\beta}^{(g_{|\mathcal{O}|})})^\top \in \mathbb{R}^{|\mathcal{O}|\times p}$ such that $\{g_1, \ldots g_{|\mathcal{O}|}\} = \mathcal{O}$. In words, $B_0^{(jo)}$ is a matrix consisting of the coefficient vectors of all the observed groups. Then (8) holds with

$$R_t = (B_t^{(jo)})^\dagger B_t^{(ar)} \in \mathbb{R}^{\omega_t\times p_t} \text{ and} \\ R_0 = (B_0^{(jo)})^\dagger B_0^{(ar)} \in \mathbb{R}^{p\times p}, \qquad (9)$$

where the notation $A^\dagger$ denotes the pseudo-inverse of a matrix A. Condition 2 guarantees that $\{R_t\}_{t=0}^q$ are well-defined and estimable. Indeed, we use $\boldsymbol{\beta}(\mathcal{A}_t \circ_t \Omega_t)$ instead of $\boldsymbol{\beta}(\Omega)$ to realize R_t because the definition of $\mathcal{A}_t$ guarantees that $\mathcal{A}_t \circ_t [p_t] \subseteq \mathcal{O}$ but $\Omega_{-t} \circ_t [p_t] \not\subseteq \mathcal{O}$ in general. In fact, R_t also lives in a low-dimensional subspace. Consider the SVD of $B_t^{(jo)} = U_t \Lambda_t V_t^\top$. By (9), we have the following relationships

$$V_t V_t^\top R_t = V_t \Lambda_t^{-1} U_t^\top B_t^{(ar)} = R_t, \ t = 0, \ldots, q. \qquad (11)$$



(a). $\{\boldsymbol{\beta}^{(g)}\}_{g\in\mathcal{O}}$ (b). $\boldsymbol{\beta}(\mathcal{A}_1 \circ_1 [p_1])$ and $\boldsymbol{\beta}(\mathcal{A}_2 \circ_2 [p_2])$ (c). $\boldsymbol{\beta}(\Omega)$

Figure 1. A graphical illustration of the coefficient tensor $\boldsymbol{\beta}(\mathcal{G})$. In the upper plane, the blue squares correspond to the group indices in $\mathcal{O}$, the orange and yellow squares correspond to the group indices in $\mathcal{A}_t \circ_t [p_t]$, $t = 1, 2$, and the green squares correspond to the group indices in Ω.

Algorithm 1: TensorDG: Domain Generalization via Tensor Completion

```
Input: ${X^{(g)}, \boldsymbol{y}^{(g)}}_{g\in\mathcal{O}}$, body set $\Omega$, arm sets $\mathcal{A}_t$, $t = 1, \dots, q$.
Output: $\widehat{\boldsymbol{\beta}}(\mathcal{G}) \in \mathbb{R}^{p\times p_1\times\cdots\times p_q}$.
Step 0: Sample splitting. For each $g \in \mathcal{O}$, split the samples into two disjoint folds such that $I_1^{(g)} \cup I_2^{(g)} = [n^{(g)}]$ and
  $|I_1^{(g)}| \approx |I_2^{(g)}|$. Let $(\widetilde{X}^{(g)}, \tilde{\boldsymbol{y}}^{(g)}) \in \mathbb{R}^{|I_1^{(g)}|\times(p+1)}$ and $(\mathring{X}^{(g)}, \mathring{\boldsymbol{y}}^{(g)}) \in \mathbb{R}^{|I_2^{(g)}|\times(p+1)}$ denote the observations within first and
  second folds, respectively.
Step 1: Estimation of the rank and basis for each mode. For $t = 0, \dots, q$, estimate $\tilde{r}_t$ and $\widetilde{V}_t$ via Algorithm 2 with input
  ${\widetilde{X}^{(g)}, \tilde{\boldsymbol{y}}^{(g)}}_{g\in\mathcal{O}}$.
Step 2: Estimate $\Gamma_t$. For each $g \in \mathcal{O}$, compute $\tilde{\boldsymbol{\beta}}^{(g)} = \{(\widetilde{X}^{(g)})^\top \widetilde{X}^{(g)}\}^{-1}(\widetilde{X}^{(g)})^\top \tilde{\boldsymbol{y}}^{(g)}$ and $\mathring{\boldsymbol{\beta}}^{(g)} = \{(\mathring{X}^{(g)})^\top \mathring{X}^{(g)}\}^{-1}(\mathring{X}^{(g)})^\top \mathring{\boldsymbol{y}}^{(g)}$.
for mode $t = 1, \dots, q$ do
    Unfold $\widetilde{B}_t^{(jo)} = \mathcal{M}_t^\top[\tilde{\boldsymbol{\beta}}(\mathcal{A}_t \circ_t \Omega_t)] \in R^{(|\mathcal{A}_t|p)\times\omega_t}$, $\mathring{B}_t^{(jo)} = \mathcal{M}_t^\top[\mathring{\boldsymbol{\beta}}(\mathcal{A}_t \circ_t \Omega_t)] \in \mathbb{R}^{(|\mathcal{A}_t|p)\times\omega_t}$, and $\mathring{B}_t^{(ar)} = \mathcal{M}_t^\top[\mathring{\boldsymbol{\beta}}(\mathcal{A}_t \circ_t [p_t])] \in \mathbb{R}^{(|\mathcal{A}_t|p)\times p_t}$.
    Compute
        $\widehat{\Gamma}_t = (\widetilde{V}_t^\top (\widetilde{B}_t^{(jo)})^\top \mathring{B}_t^{(jo)} \widetilde{V}_t)^{-1} (\widetilde{B}_t^{(jo)} \widetilde{V}_t)^\top \mathring{B}_t^{(ar)} \in \mathbb{R}^{\tilde{r}_t\times p_t}$.    (10)
end
For $t = 0$, unfold $\widetilde{B}_0 = \mathcal{M}_0^\top[\tilde{\boldsymbol{\beta}}(\mathcal{O})] \in \mathbb{R}^{|\mathcal{O}|\times p}$, $\mathring{B}_0 = \mathcal{M}_0^\top[\mathring{\boldsymbol{\beta}}(\mathcal{O})] \in \mathbb{R}^{|\mathcal{O}|\times p}$ and compute
  $\widehat{\Gamma}_0 = (\widetilde{V}_0^\top \widetilde{B}_0^\top \mathring{B}_0 \widetilde{V}_0)^{-1} (\widetilde{B}_0 \widetilde{V}_0)^\top \mathring{B}_0 \in \mathbb{R}^{\tilde{r}_0\times p_0}$.
Step 3: Tensor completion. Compute
        $\widehat{\boldsymbol{\beta}}(\mathcal{G}) = (\mathring{\boldsymbol{\beta}}(\Omega) \times_{t=0}^q \widetilde{V}_t) \times_{t=0}^q \widehat{\Gamma}_t$.
```

That is, R_t lies in the linear subspace spanned by V_t. Together with (8), we have

$$\begin{aligned}\boldsymbol{\beta}(\mathcal{G}) &= \boldsymbol{\beta}(\Omega) \times_0 V_0 V_0^\top \times_{t=1}^q V_t V_t^\top R_t \\ &= (\boldsymbol{\beta}(\Omega) \times_{t=0}^q V_t) \times_{t=0}^q \Gamma_t, \qquad (12)\end{aligned}$$

where $\Gamma_0 = V_0^\top \in \mathbb{R}^{r_0\times p}$ and $\Gamma_t = V_t^\top R_t \in \mathbb{R}^{r_t\times p_t}$, $t = 1, \dots, q$. We see from (12) that to recover the whole tensor, it suffices to estimate $\boldsymbol{\beta}(\Omega) \times_{t=0}^q V_t \in \mathbb{R}^{r_0\times r_1\cdots\times r_q}$ and Γ_t for $t = 0, \dots, q$, which has degree of freedom $\prod_{t=0}^q r_t + \sum_{t=0}^q (p_t - r_t) r_t$, where we use the convention $p_0 = p$. Indeed, $\boldsymbol{\beta}(\Omega) \times_{t=0}^q V_t$ is always referred to as the core tensor (Zhang and Xia 2018) as it is the smallest possible tensor containing the coefficients of each mode and Γ_t can be viewed as multiplying coefficients for the tth direction and it spans the subspace of tth mode. We will propose an algorithm to estimate the core tensor and $\{\Gamma_t\}_{t=0}^q$ in the next subsection.

Remark 1. Model (8) is also related to the tensor factor models (Han et al. 2020; Chen, Yang, and Zhang 2022), which have been studied in high-dimensional tensor time series. Using the terminology in factor models, Γ_t are the loading matrices and $\boldsymbol{\beta}(\Omega) \times_{t=0}^q V_t$ is the tensor of factors. In the time series applications, the observed data always form a complete tensor which is different from our setting.

3.2. Proposed Algorithm

We now devise an algorithm to estimate $\boldsymbol{\beta}(\mathcal{G})$ based on (12). Our proposal has three main steps. The first step is to estimate the low-dimensional subspace V_t for $t = 0, \dots, q$. Then we estimate Γ_t based on (9) and (11). Finally, we assemble the estimated tensor based on (12). The proposal, termed as TensorDG, is presented in Algorithm 1. Algorithm 1 starts with a sample-splitting step, which is mainly for technical convenience. Similar approaches have also been used in other works to derive theoretical guarantees for tensor estimation and completion (Zhang et al. 2020).

In Step 1, Algorithm 1 estimates V_t by $\widetilde{V}_t$ using half of the samples via Algorithm 2. Algorithm 2 is motivated by the fact that V_t is the column space of $\Theta_t = \{\mathcal{M}_t[\boldsymbol{\beta}(\Omega)]\}^\top \mathcal{M}_t[\boldsymbol{\beta}(\Omega)]/|\Omega_{-t}|$. The proposed estimate $\widetilde{\Theta}_t$ of Θ_t is based on the least square estimates of $\boldsymbol{\beta}^{(g)}$, $g \in \Omega$ and the last term of its expression further corrects the bias caused by the product of OLS estimates. The rank $\tilde{r}_t$ is estimated as the number of significantly nonzero eigenvalues of $\widetilde{\Theta}_t$. The threshold level λ_t is chosen based on the concentration properties of $\widetilde{\Theta}_t$. Determination of the tensor rank can also be based on information criteria (Han, Chen, and Zhang 2022).

In Step 2, we obtain estimates of $\boldsymbol{\beta}^{(g)}$, $g \in \mathcal{O}$ based on two disjoint sets of samples. Then we estimate Γ_t by a least-square type of estimate $\widehat{\Gamma}_t$ using the definition of Γ_t below (12). In Step 3, the whole tensor is assembled according to (12) with the aforementioned estimates. Thanks to the structural missing patterns, we are able to achieve completion in one round. Moreover, each step involves convex optimizations, which are computationally efficient.

Our method exhibits two primary distinctions from existing tensor completion methods, such as the Cross method proposed in Zhang (2019), which also employs decomposition (8). First, in the standard tensor completion problem, the sample for each element of the oracle tensor is independent and unbiased. In contrast, in the domain generalization problem under consideration, we need to fit regression models for each domain. The produced OLS estimates therefore exhibit correlated errors, which need to be calibrated as detailed in Algorithm 2. Second,

Algorithm 2: SVD for mode t.

```
Input: $\{\widetilde{X}^{(g)}, \tilde{\boldsymbol{y}}^{(g)}\}_{g\in\mathcal{O}}$.
Output: Estimated rank $\tilde{r}_t$ and basis $\widetilde{V}_t \in \mathbb{R}^{\omega_t\times\tilde{r}_t}$.
For each $g \in \mathcal{O}$, let $\tilde{n}^{(g)} = |I_1^{(g)}|$ and compute $\widetilde{\Sigma}^{(g)} = (\widetilde{X}^{(g)})^\top\widetilde{X}^{(g)}/\tilde{n}^{(g)}$,
    $$\tilde{\boldsymbol{\beta}}^{(g)} = \{\widetilde{\Sigma}^{(g)}\}^{-1}(\widetilde{X}^{(g)})^\top\tilde{\boldsymbol{y}}^{(g)}/\tilde{n}^{(g)} \text{ and } \tilde{\sigma}_g^2 = \frac{\|\tilde{\boldsymbol{y}}^{(g)} - \tilde{X}^{(g)}\tilde{\boldsymbol{\beta}}^{(g)}\|_2^2}{\tilde{n}^{(g)} - p}. \quad (13)$$
- If $t = 0$, let $\widetilde{B}_0 = \mathcal{M}_0^\top[\tilde{\boldsymbol{\beta}}(\mathcal{O})] \in \mathbb{R}^{|\mathcal{O}|\times p}$. Perform eigen decomposition for
    $$\widetilde{\Theta}_0 = \frac{1}{|\mathcal{O}|}\widetilde{B}_0^\top\widetilde{B}_0 - \frac{1}{|\mathcal{O}|}\sum_{g\in\mathcal{O}}(\widetilde{\Sigma}^{(g)})^{-1}\frac{\tilde{\sigma}_g^2}{\tilde{n}^{(g)}} = \mathring{V}_0\mathring{\Lambda}_0\mathring{V}_0^\top \in \mathbb{R}^{p\times p}.$$
  Let $\tilde{r}_0 = \sum_{k=1}^p \mathbb{1}((\mathring{\Lambda}_0)_{k,k} \geq \lambda_0)$ and $\widetilde{V}_0 = \{\mathring{V}_0\}_{.,1:\tilde{r}_0} \in \mathbb{R}^{p\times\tilde{r}_0}$, where $\lambda_0 = C_\lambda\sqrt{\|\widetilde{\Theta}_0\|_2(p + \log\bar{n})/(\bar{n}|\mathcal{O}|)}$ for
  $\bar{n} = \sum_{g\in\mathcal{O}}\tilde{n}^{(g)}/|\mathcal{O}|$.
- If $t \geq 1$, let $\widetilde{B}_t = \mathcal{M}_t^\top[\tilde{\boldsymbol{\beta}}(\Omega)] \in \mathbb{R}^{|\Omega_{-t}|p\times\omega_t}$. Define
    $$\widetilde{\Theta}_t = \frac{1}{|\Omega_{-t}|}\widetilde{B}_t^\top\widetilde{B}_t - \frac{1}{|\Omega_{-t}|}\mathrm{Diag}(\tilde{\boldsymbol{v}}) = \mathring{V}_t\mathring{\Lambda}_t\mathring{V}_t^\top \in \mathbb{R}^{\omega_t\times\omega_t},$$
  where $\mathrm{Diag}(\tilde{\boldsymbol{v}})$ is a diagonal matrix with $\{\mathrm{Diag}(\tilde{\boldsymbol{v}})\}_{j,j} = \tilde{v}_j$ and $\tilde{\boldsymbol{v}}_j = \sum_{g\in\Omega, g_t=(\Omega_t)_j}\mathrm{Tr}(\{\widetilde{\Sigma}^{(g)}\}^{-1})\frac{\tilde{\sigma}_g^2}{\tilde{n}^{(g)}}$. Let
  $\tilde{r}_t = \sum_{k=1}^{\omega_t}\mathbb{1}((\mathring{\Lambda}_t)_{k,k} \geq \lambda_t)$ and $\widetilde{V}_t = \{\mathring{V}_t\}_{.,1:\tilde{r}_t} \in \mathbb{R}^{\omega_t\times\tilde{r}_t}$, where $\lambda_t = C_\lambda\sqrt{\|\widetilde{\Theta}_t\|_2(\omega_t + \log\bar{n})/(\bar{n}|\Omega_{-t}|)}$.
```

for tensor $\boldsymbol{\beta}(\mathcal{G})$, the 0th mode is not exchangeable with other q group modes. This distinction arises because $\boldsymbol{\beta}^{(g)}$ represents the smallest unit of interest in domain generalization, and it is either observed or missed in its entirety. Consequently, the dimension reduction approach we adopt for the 0th mode differs from the strategies employed for the remaining q modes. Another related work is Simchowitz, Gupta, and Zhang (2023), which studies combinatorial distribution shifts, that is, the training and test data distributions differ significantly due to differences in the combinations of features. The difference is that they have one source and one target distribution but we have multiple source and target distributions. Moreover, we allow $\boldsymbol{x}_i^{(g)}$ to be continuous, making the analysis very different.

4. Theoretical Properties

In this section, we provide theoretical guarantees for Algorithm 1. We first state the main assumptions.

Condition 3 (Distribution of observed data). For each $g \in \mathcal{O}$, $\boldsymbol{x}_i^{(g)}$, $i = 1, \ldots, n^{(g)}$, is independent sub-Gaussian with mean zero and covariance matrix $\Sigma^{(g)}$, where $c_1 \leq \min_{g\in\mathcal{O}}\Lambda_{\min}(\Sigma^{(g)}) \leq \max_{g\in\mathcal{O}}\Lambda_{\max}(\Sigma^{(g)}) \leq c_2$ for some positive constants c_1 and c_2. For each $g \in \mathcal{O}$, $n^{(g)} \asymp n$ and the noise $\epsilon_i^{(g)}$ is independent sub-Gaussian with mean zero and variance σ_g^2 and $\max_{g\in\mathcal{O}}\sigma_g^2 \leq \bar{\sigma}^2 < \infty$. For each $g \in \mathcal{G}\backslash\mathcal{O}$, $\boldsymbol{x}_i^{(g)}$, $i = 1, \ldots, n^{(g)}$, is independent sub-Gaussian with mean zero and covariance matrix $\Sigma^{(g)}$ with $\max_{g\in\mathcal{G}\backslash\mathcal{O}}\Lambda_{\max}(\Sigma^{(g)}) \leq c_2$.

Condition 3 assumes sub-Gaussian designs and sub-Gaussian errors. Heterogeneous distributions for both $\boldsymbol{x}_i^{(g)}$ and $\epsilon_i^{(g)}$ are allowed. The assumption that $n^{(g)} \asymp n$ is a simplified scenario for technical convenience and has been commonly considered in the multi-task learning literature (Guo et al. 2011; Du et al. 2020; Tripuraneni, Jin, and Jordan 2021). The sub-Gaussian assumption on the covariates $X^{(g)}$, $g \in \mathcal{G} \setminus \mathcal{O}$ are only used to establish prediction error bounds for the unseen groups and they are not used in the training phase. Condition 3 also assumes that $n^{(g)} \asymp n$ for simplicity. In practice, it could be the case that $n_{\min} := \min_{g\in\mathcal{O}} n^{(g)} \ll \bar{n}$. In this case, the performance of TensorDG can depend on $n_{\min}$. Hence, one can remove the groups with only a few samples from $\mathcal{O}$ for a faster convergence rate. Specifically, let $\tilde{\mathcal{O}}$ denote the set of groups with $n^{(g)} > 0$. We would like to find $\mathcal{O} \subseteq \tilde{\mathcal{O}}$ such that $\mathcal{O}$ contains important structural information so that Condition 2 is satisfied and $n_{\min}$ is not too small.

Our analysis allows (p_t, r_t), $t = 0, \ldots, q$ to go to infinity but $p \leq c_1 n$ for some small enough constant c_1. This low-dimensional assumption guarantees the regularity of the least square estimates of each group. In Section E of the supplements, we discuss possible extensions of the proposed methods to the high-dimensional setting. For $t = 1, \ldots, q$, define

$$\Theta_t = \frac{1}{|\Omega_{-t}|}\{\mathcal{M}_t[\boldsymbol{\beta}(\Omega)]\}^\top\mathcal{M}_t[\boldsymbol{\beta}(\Omega)] \in \mathbb{R}^{\omega_t\times\omega_t},$$
$$\Theta_0 = \frac{1}{|\mathcal{O}|}\{B_0^{(jo)}\}^\top B_0^{(jo)} \in \mathbb{R}^{p\times p}. \quad (14)$$

Define $e_* = \min_{0\leq t\leq q}\Lambda_{r_t}(\Theta_t)$ and $e^* = \max_{0\leq t\leq q}\Lambda_{\max}(\Theta_t)$.

Condition 4 (Eigenvalue conditions). For Θ_t defined in (14), assume that for some large enough constant $C > 2C_\lambda$, $e_* \geq C\bar{\sigma}\max\{\max_{1\leq t\leq q}\sqrt{\frac{e^*(\omega_t+\log n)}{n|\Omega_{-t}|}}, \sqrt{\frac{e^* p}{n|\mathcal{O}|}}\}$. Moreover, assume that $\Lambda_{r_0}(B_0^{(jo)}) \geq C'\bar{\sigma}\sqrt{\frac{p+\log n}{n}}$ and $\Lambda_{r_t}(B_t^{(jo)}) \geq C'\bar{\sigma}\max\{\sqrt{\frac{r_t+\log n}{n}}, \Lambda_{\max}^{1/2}(B_t^{(jo)})(\frac{r_t+\log n}{n})^{1/4}\}$, $t = 1,\ldots,q$, for some large enough constant C'.

Condition 4 can be viewed as an assumption on the condition numbers for $B_t^{(jo)}$ and Θ_t, $t = 0,\ldots,q$. Note that the condition numbers are allowed to go to infinity as $n \to \infty$.

4.1. Upper Bounds for the Generalization Errors

We first establish the estimation accuracy of subspace estimation in Step 1 of Algorithm 1. For two orthonormal matrices $A \in \mathbb{R}^{p_1\times p_2}$ and $B \in \mathbb{R}^{p_1\times p_2}$, let $\sin\theta(A,B) := \|\mathrm{Diag}(\sin(\arccos(\lambda_1)),\ldots,\sin(\arccos(\lambda_r)))\|_2$, where $\lambda_1 \geq \cdots \geq \lambda_{p_1\wedge p_2}$ are the singular values of $A^\top B$.

Lemma 1 (Subspace estimation for each mode). Suppose that Conditions 1–4 hold. If $p(\log|\mathcal{O}| + \log n) \leq c_1 n$ with small enough constant c_1, then for any $1 \leq t \leq q$,

$$\sin\theta(\widetilde{V}_t, V_t) \leq \frac{c_2\bar{\sigma}}{e_*}\sqrt{\frac{e^*(\omega_t + \log n)}{n|\Omega_{-t}|}},\ \ \tilde{r}_t = r_t$$

with probability at least $1 - \exp\{-c_2\log n\}$. Moreover, with probability at least $1 - \exp\{-c_2\log n\}$,

$$\sin\theta(\widetilde{V}_0, V_0) \leq \frac{c_2\bar{\sigma}}{e_*}\sqrt{\frac{e^*(p + \log n)}{n|\mathcal{O}|}},\ \ \tilde{r}_0 = r_0.$$

Lemma 1 provides the convergence rate of $\widetilde{V}_t$ for V_t, $t = 0,\ldots,q$ under given conditions. For each $1 \leq t \leq q$, we use $n|\Omega_{-t}|$ samples to estimate V_t, loosely speaking. The condition $p(\log|\mathcal{O}| + \log n) \leq c_1 n$ guarantees that $\min_{g\in\mathcal{O}}\Lambda_{\min}(\widetilde{\Sigma}^{(g)})$ is bounded away from zero with high probability.

In the following theorem, we provide formal upper bounds for our proposal. For $t = 0,\ldots,q$, let $\widetilde{\Gamma}_t = (V_t^\top\widetilde{V}_t)^{-1}\Gamma_t$. Let $C_R = \max_{1\leq t\leq q}\|R_t\|_{2,\infty} \vee 1$ and $\bar{C}_R = \max_{1\leq t\leq q}\|R_t\|_2 \vee 1$. Let $\rho_t = \|\mathcal{M}_t[\boldsymbol{\beta}(\Omega)]\|_2/\Lambda_{r_t}(B_t^{(jo)})$ for $t = 1,\ldots,q$ and $\rho_0 = \min_{1\leq t\leq q}\|\mathcal{M}_0[\boldsymbol{\beta}(\Omega_{-t}\circ_t[p_t])]\|_2/\Lambda_{r_0}(\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{A}_t\circ_t[p_t])])$. In the simple case where $\mathcal{A}_t = \Omega_{-t}$, ρ_t is the condition number for the matrix $\mathcal{M}_t[\boldsymbol{\beta}(\Omega)]$ for $t = 1,\ldots,q$ and ρ_0 is the minimum condition number for the matrix $\mathcal{M}_0[\boldsymbol{\beta}(\Omega_{-t}\circ_t[p_t])]$, $t = 1,\ldots,q$.

Theorem 1 (Domain generalization bounds in tensor Frobenius norm). Assume Conditions 1–4. Suppose that $p(\log|\mathcal{O}| + \log n) \leq c_1 n$, $\omega_t \leq c_1\sqrt{n}|\Omega_{-t}|$, $\sum_{t=1}^q\sqrt{\frac{p_t+\log n}{\Lambda_{r_t}^2(B_t^{(jo)})n}} \leq c_1$ for some small enough constant c_1, and $\max_{0\leq t\leq q}\rho_t \leq C < \infty$. Then for any fixed $\eta \lesssim \log n$, with probability at least $1 - \exp\{-c_2\log n\} - c_3\exp\{-c_4\eta\}$

$$\begin{aligned}\|\hat{\boldsymbol{\beta}}(\mathcal{G}) - \boldsymbol{\beta}(\mathcal{G})\|_{\ell_2} &\vee \sqrt{\sum_{g\in\mathcal{G}}\frac{1}{n^{(g)}}\|X^{(g)}(\hat{\boldsymbol{\beta}}^{(g)} - \boldsymbol{\beta}^{(g)})\|_2^2}\\ &\lesssim \bar{C}_R^{q-1}\sum_{t=1}^q\sqrt{\frac{(\bar{C}_R^2 r_t + p_t + \bar{C}_R^2\eta)r_t}{n}}\\ &+ \bar{C}_R^{q-1}\sqrt{\frac{(p+\eta)r_0}{n}}\\ &+ \bar{C}_R^{q}\sqrt{\frac{\prod_{t=0}^q r_t + \eta}{n}}.\end{aligned} \tag{15}$$

Theorem 1 provides an upper bound for the estimation errors in all the environments including the unseen ones. The generalization error is decomposed into three main sources. The first component is the estimation error of $\widehat{\Gamma}_t$ for $t = 1,\ldots,q$. The second component comes from the estimation of $\widehat{\Gamma}_0$, which is the dimension reduction for the 0th mode and has no missingness. The third term comes from the noise in the estimated core tensor.

To further understand this result, if $\bar{C}_R \leq c_1 < \infty$, $\eta \lesssim \min_{0\leq t\leq q}p_t \wedge (\prod_{t=0}^q r_t)$, and $r_t \leq c_2 p_t$ for some $c_2 < 1$, and, then the upper bound in (15) can be rewritten as

$$\sum_{t=1}^q\sqrt{\frac{r_t(p_t - r_t)}{n}} + \sqrt{\frac{(p - r_0)r_0}{n}} + \sqrt{\frac{\prod_{t=0}^q r_t}{n}}. \tag{16}$$

It is known that the degree of freedom for a tensor with rank $(r_0, r_1,\ldots,r_q)$ is $\sum_{t=0}^q r_t(p_t - r_t) + \prod_{t=0}^q r_t$. Loosely speaking, the upper bound in (16) shows that the estimation error of $\widehat{\boldsymbol{\beta}}(\mathcal{G})$ is equivalent to estimating all the essential parameters with the observed samples. From the perspective of tensor completion, Theorem 2 in Zhang (2019) considers the case that each observed element in the tensor has one independent sample and its upper bound involves the raw noises. The result (15) is a probabilistic bound and it is more general because, in the current case, each observed element in the tensor corresponds to a regression model with $n^{(g)} \asymp n$ independent samples.

In the following, we present a counterpart of Theorem 1 with respect to the max norm. Let $\rho_t' = \frac{\max_{g\in\mathcal{G}}\|\boldsymbol{\beta}^{(g)}\|_2}{\Lambda_{r_t}(B_t^{(jo)})/\sqrt{|\mathcal{A}_t|}}$ for $t = 1,\ldots,q$ and $\rho_0' = \frac{\max_{g\in\mathcal{G}}\|\boldsymbol{\beta}^{(g)}\|_2}{\Lambda_{r_0}(B_0^{(jo)})/\sqrt{|\mathcal{O}|}}$.

Theorem 2 (Domain generalization bounds in max norm). Assume Conditions 1–4 hold. Assume that $p(\log|\mathcal{G}| + \log n) \leq c_1 n$, $\omega_t \leq c_1\sqrt{n}|\Omega_{-t}|$, $\sum_{t=1}^q\sqrt{\frac{r_t+\log n}{n\Lambda_{r_t}^2(B_t^{(jo)})}} \leq c_1\sqrt{\frac{r_0+\log n}{\prod_{t=0}^q r_t+\log n}} \wedge \frac{1}{\max_{1\leq t\leq q}\sqrt{\omega_{-t}}}$, and $\max_{0\leq t\leq q}\rho_t' \leq C$, then with probability at least $1 - \exp\{-c_2\log n\}$,

$$\begin{aligned}\max_{g\in\mathcal{G}}\|\hat{\boldsymbol{\beta}}^{(g)} - \boldsymbol{\beta}^{(g)}\|_2 &\vee \max_{g\in\mathcal{G}}\frac{1}{\sqrt{n^{(g)}}}\|X^{(g)}(\hat{\boldsymbol{\beta}}^{(g)} - \boldsymbol{\beta}^{(g)})\|_2\\ &\lesssim \sum_{t=1}^q\sqrt{\frac{\omega_t(C_R^2 r_t + \log n)}{a_t n}}\\ &+ \sqrt{\frac{p + \log n}{|\mathcal{O}|n}} + C_R^q\sqrt{\frac{r_0 + \log|\mathcal{G}| + \log n}{n}}.\end{aligned} \tag{17}$$

The results in max norm (17) can be more useful in the domain generalization setting. The three terms in the upper bound also correspond to the three sources of errors as in Theorem 1. The term $\log|\mathcal{G}|$ appears in the upper bound as we take maximum over all the groups in $\mathcal{G}$. The quantities ρ_0'

and ρ_t' are constants if $\boldsymbol{\beta}^{(g)}$, $g \in \Omega$ are independent sub-Gaussian and $\|\boldsymbol{\beta}^{(g)}\|_2 \asymp \|\boldsymbol{\beta}^{(g')}\|_2$ for all $g, g' \in \mathcal{G}$. We see that the rate in max norm can be slower than the rate in tensor Frobenius norm divided by $\sqrt{\prod_{t=1}^q p_t}$. Especially, the last term in (17) does not show the aggregation effects across multiple groups. In noisy matrix completion setting, the optimal entry-wise estimation rate has no aggregation effect either (Chen et al. 2019) and our tensor completion results can be understood analogously. Nevertheless, the rate in (17) can still be faster than the meta-learning method (Tripuraneni, Jin, and Jordan 2021) if $n \gg n^{(g^*)}$.

4.2. Minimax Lower Bound

In this section, we provide minimax lower bound results for the current problem. Let $\boldsymbol{\beta}(\mathcal{G}) \in \mathbb{R}^{p\times p_1\times\cdots\times p_q}$ be the coefficient tensor corresponding to a set of group $\mathcal{G}$. We consider the following parameter space

$$\Theta(\boldsymbol{r}, \bar{C}_R) = \Big\{\boldsymbol{\beta}(\mathcal{G}) \in \mathbb{R}^{p\times p_1\times\cdots\times p_q} : \mathrm{rank}(\mathcal{M}_k[\boldsymbol{\beta}(\mathcal{G})]) = \mathrm{rank}(\mathcal{M}_k[\boldsymbol{\beta}(\Omega)]) \le r_k, k = 0, \ldots, q, \max_{1\le t\le q} \|R_t\|_2 \le \bar{C}_R, \max_{0\le t\le q} \rho_t \le C\Big\},$$

where $\boldsymbol{r} = (r_0, \ldots, r_q)$ and C is some large enough constant. We present the minimax lower bound result below.

Theorem 3 (Minimax lower bound in tensor Frobenius norm). Assume Conditions 2 and 3, $p_t \ge 3r_t$, $4r_t \le \prod_{t'\neq t} r_{t'}$, $r_t \ge 2$ for $t = 0, \ldots, q$, and q is finite. There exists some positive constant c_1 such that

$$\inf_{\widehat{\boldsymbol{\beta}}(\mathcal{G})} \sup_{\boldsymbol{\beta}(\mathcal{G})\in\Theta(\boldsymbol{r},\bar{C}_R)} \mathbb{P}\left(\|\hat{\boldsymbol{\beta}}(\mathcal{G}) - \boldsymbol{\beta}(\mathcal{G})\|_{\ell_2} \ge c_1\bar{C}_R^{q-1}\sum_{t=1}^q \sqrt{\frac{p_t r_t}{n}} + c_1\bar{C}_R^{q-1}\sqrt{\frac{pr_0}{n}} + c_1\bar{C}_R^{q-1}\sqrt{\frac{\prod_{t=0}^q r_t}{n}}\right) \ge 1/4.$$

Compared with the rate in (15) of Theorem 1, we see that the proposed algorithm is minimax rate optimal in the parameter space $\Theta(\boldsymbol{r}, \bar{C}_R)$ when $\bar{C}_R = O(1)$ or $\bar{C}_R \prod_{t=1}^q r_t = O(p)$. A simple example in which $\bar{C}_R = O(1)$ is that $R_t = V_t$ for $V_t \in \mathbb{O}_{p_t,r_t} = \{V \in \mathbb{R}^{p_t\times r_t} : V^\top V = I_{r_t}\}$.

To summarize, TensorDG enjoys both computational efficiency and estimation accuracy if the low-rank tensor model holds. However, our proposal could fail if Condition 2 is not satisfied. In practice, Condition 2 can be diagnosed to some extent. For instance, one can determine whether the rank of $\mathcal{M}_t[\boldsymbol{\beta}(\mathcal{A}_t \circ_t \Omega_t)]$ equals the rank of $\mathcal{M}_t[\boldsymbol{\beta}(\mathcal{A}_t \circ_t [p_t])]$ based on the information criterion or the eigen-ratio criterion (Han, Chen, and Zhang 2022). If they are not equal, then Condition 2 is violated. Such tests can answer the important question whether the model is generalizable based on the observed data, which concerns the reliability of a domain generalization method. In contrast, there seems no direct way to verify the generalizability of some other frameworks such as the invariant causal models.

5. Extensions

In this section, we extend the main methodology to handle transfer learning tasks. We present the extensions to high-dimensional scenarios in Section E of the supplements.

TensorDG has demonstrated the capability to estimate $\boldsymbol{\beta}^{(g)}$ even if $n^{(g)} = 0$ by leveraging the low-rank tensor structure. In real-world scenarios, it is also common to have a limited number of samples available from the target domain. Specifically, for a given target domain g^*, it often holds that $0 < n^{(g^*)} \ll \sum_{g\in\mathcal{O}} n^{(g)}$. In this situation, we would like to harness the $n^{(g^*)}$ samples from the target domain. Li, Cai, and Li (2022), Tian and Feng (2022), and Li et al. (2023) have studied transfer learning with sparsity-based similarity characterizations among many others. Du et al. (2020), Tripuraneni, Jin, and Jordan (2021), Chua, Lei, and Lee (2021), Duan and Wang (2023), Tian, Gu, and Feng (2023) and other works consider transfer learning with low-rank similarity assumptions on $\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{G})]$. Ghosh et al. (2020) and Kong et al. (2020) consider clustered multi-task learning in the setting where different groups or sub-populations may come from the same task, which can also be modeled using the low-rank similarity assumptions on $\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{G})]$.

In transfer learning, avoiding negative transfer is a critical concern. Negative transfer occurs when the assumed similarity between the source and target tasks fails, which could lead to a deterioration in the performance of transfer learning compared to only using target data. To address this concern, we relax (2) to account for an additional level of model heterogeneity. Specifically, we assume

$$\begin{aligned} y_i^{(g)} &= (\boldsymbol{x}_i^{(g)})^\top \boldsymbol{\beta}^{(g)} + \epsilon_i^{(g)}, \ i \in [n^{(g)}], \ \forall g \in \mathcal{O}\setminus\{g^*\} \\ y_i^{(g^*)} &= (\boldsymbol{x}_i^{(g^*)})^\top \boldsymbol{\gamma}^{(g^*)} + \epsilon_i^{(g^*)}, \ i \in [n^{(g^*)}], \\ \boldsymbol{\gamma}^{(g^*)} &= \boldsymbol{\beta}^{(g^*)} + \boldsymbol{\delta}^{(g^*)}, \end{aligned} \tag{18}$$

where $\{\boldsymbol{\beta}^{(g)}\}_{g\in\mathcal{O}}$ belongs to the tensor $\boldsymbol{\beta}(\mathcal{G})$ satisfying Condition 2 and $\boldsymbol{\delta}^{(g^*)} \in \mathbb{R}^p$ represents a unique direction of $\boldsymbol{\beta}^{(g^*)}$. The magnitude of $\boldsymbol{\delta}^{(g^*)}$ also denotes the level of misspecification of the low-rank tensor model. To estimate $\boldsymbol{\gamma}^{(g^*)}$, we use the oracle Trans-Lasso (Li, Cai, and Li 2022) based on the TensorDG estimate of $\boldsymbol{\beta}^{(g^*)}$.

Theorem 4 (Estimation and prediciton errors of TensorTL). Assume the Conditions of Theorem 2 and model (18). For $\lambda_{g^*} \ge c_0\sqrt{\log p/n^{(g^*)}}$ with a large enough constant c_0, it holds that with probability at least $1-\exp\{-c_1\log p\}-\exp\{-c_2\log n\}$,

$$\begin{aligned} &\|\hat{\boldsymbol{\gamma}}^{(g^*)} - \boldsymbol{\gamma}^{(g^*)}\|_2 \vee \frac{1}{\sqrt{n^{(g^*)}}}\|X^{(g^*)}(\hat{\boldsymbol{\gamma}}^{(g^*)} - \boldsymbol{\gamma}^{(g^*)})\|_2 \\ &\quad\lesssim \sum_{t=1}^q \sqrt{\frac{C_R^2 r_t + \log n}{a_t n}} \\ &\qquad + \sqrt{\frac{p+\log n}{|\mathcal{O}|n}} + C_R^q\sqrt{\frac{r_0+\log n}{n}} + \sqrt{\frac{\|\boldsymbol{\delta}^{(g^*)}\|_0 \log p}{n^{(g^*)}}}. \end{aligned}$$

In Theorem 4, we establish the convergence rate of Algorithm 3. The first three terms in the right-hand-side is the upper bound for the TensorDG estimate $\|\hat{\boldsymbol{\beta}}^{(g^*)} - \boldsymbol{\beta}^{(g^*)}\|_2$. The last term comes from the estimation of the bias $\boldsymbol{\delta}^{(g^*)}$. Recall that

Algorithm 3: TensorTL: Transfer learning based on Algorithm 1

```
Input: TensorDG estimate $\hat{\boldsymbol\beta}^{(g^*)} = \{\widehat{\boldsymbol\beta}(\mathcal{G})\}_{\cdot g_1^*,\ldots,g_q^*}$ and target samples $(X^{(g^*)}, \boldsymbol{y}^{(g^*)})$.
Output: $\hat{\boldsymbol\gamma}^{(g^*)}$.
Step 1: For a tuning parameter $\lambda_{g^*} > 0$, compute
    $$\hat{\boldsymbol\delta}^{(g^*)} = \arg\min_{\boldsymbol\delta \in \mathbb{R}^p} \left\{ \frac{1}{n^{(g^*)}} \|\boldsymbol{y}^{(g^*)} - X^{(g^*)}\hat{\boldsymbol\beta}^{(g^*)} - X^{(g^*)}\boldsymbol\delta\|_2^2 + \lambda_{g^*}\|\boldsymbol\delta\|_1 \right\},$$
Step 2: Output $\hat{\boldsymbol\gamma}^{(g^*)} = \hat{\boldsymbol\beta}^{(g^*)} + \hat{\boldsymbol\delta}^{(g^*)}$.
```

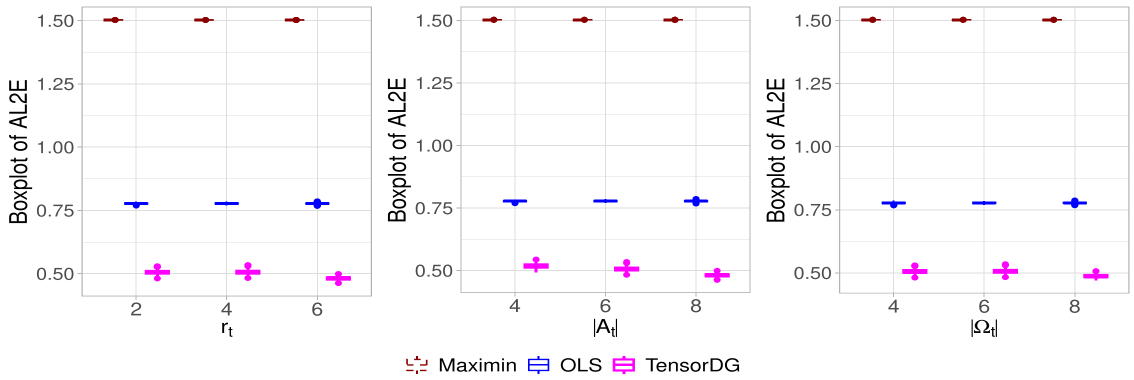


Figure 2. Boxplots of the square-root of AL2E based on Maximin (dashed brown), OLS (solid blue), and TensorDG (bold solid magenta) in experiments (a), (b), and (c).

the single-task OLS estimator has a convergence rate of order $p/n^{(g^*)}$. Under the mild conditions that $C_R = O(1)$ and $r_0 + \log n + \|\boldsymbol\delta^{(g^*)}\|_0 \log p \ll p$, the TensorTL estimate $\hat{\boldsymbol\gamma}^{(g^*)}$ has a faster convergence rate than the OLS. This condition indeed requires that the misspecified parameter $\boldsymbol\delta^{(g^*)}$ is sparse, that is, the misspecification level is relatively low. We can further compare TensorTL with the methods based on the low-rank similarity assumptions without tensor structures. For instance, Tripuraneni, Jin, and Jordan (2021) considers $\boldsymbol\delta^{(g^*)} = 0$ and the convergence rate of their method is of order $\sqrt{pr_0/(n|\mathcal{O}|)} + \sqrt{r_0/n^{(g^*)}}$ using our notation. We see that when $n^{(g^*)} \ll n$, the TensorTL estimate can have faster rate of convergence for $\boldsymbol\delta^{(g^*)} = 0$.

6. Numerical Experiments

We evaluate the performance of our proposals in domain generalization and transfer learning in comparison to some existing methods.

6.1. Domain Generalization Performance

We first evaluate the dependence of domain generalization errors on $|\mathcal{A}_t|$, $|\Omega_t|$, and r_t. For a generic estimator $\widehat{\boldsymbol\beta}(\mathcal{G})$, define its Average ℓ_2-Error (AL2E) as $\|\widehat{\boldsymbol\beta}(\mathcal{G}) - \boldsymbol\beta(\mathcal{G})\|_{\ell_2}/\sqrt{|\mathcal{G}|}$ and its Average Domain Generalization Errors (ADGE) as $\text{ADGE} = \sqrt{\sum_{g\in\mathcal{O}^c} \|\hat{\boldsymbol\beta}^{(g)} - \boldsymbol\beta^{(g)}\|_2^2/|\mathcal{O}^c|}$.

We compare the performance of TensorDG, single-task OLS, and Maximin estimator. It is known that sample splitting can result in inefficient use of samples. We evaluate different versions of sample splitting and find that the most effective version is to all the samples in all the steps of Algorithm 1. To compute single-task OLS, we generate $n^{(g)}$ samples for group g if $g \notin \mathcal{O}$. In contrast, TensorDG and Maximin only use samples in $\mathcal{O}$.

The default setting in our simulation is $n^{(g)} = 300$ for each $g \in \mathcal{O}$, $q = 2$, $(p_0, p_1, p_2) = (80, 10, 15)$, $r_t = 4$, $r_0 = 2r_t$, and $|\mathcal{A}_t| = |\Omega_t| = 6$ for $t = 1, 2$. In experiment (a), we consider $r_t \in \{2, 4, 6\}$ and set other parameters as default. In experiment (b), we consider $|\mathcal{A}_t| \in \{4, 6, 8\}$ and set other parameters as default. In experiment (c), we consider $|\Omega_t| \in \{4, 6, 8\}$ and set other parameters as default. The average ℓ_2-error of TensorDG, single-task OLS, and Maximin estimator are given in Figure 2. Each setting is replicated with 500 Monte Carlo experiments. We see that the average estimation error of TensorDG increases as r_t increases and decreases as $|\mathcal{A}_t|$ or $|\Omega_t|$ increases, which aligns with our theoretical analysis. In Figure 3, we see that TensorDG has the smallest average domain generalization errors. The single-task OLS has larger errors as it only uses $n^{(g)}$ samples from group g and its estimation accuracy is invariant to r_t, $|\mathcal{A}_t|$, and $|\Omega_t|$. The Maximin estimator has the largest domain generalization errors in these settings. One reason is that the success of Maximin requires that the parameters of test domains fall in the simplex formed by the parameters of the training domains. However, such assumptions may not be true in the current setting.

6.2. Transfer Learning Performance

Next, we evaluate the performance of TensorTL for transfer learning tasks. For comparison, we consider a modification of Meta-LM-MoM proposed in Tripuraneni, Jin, and Jordan

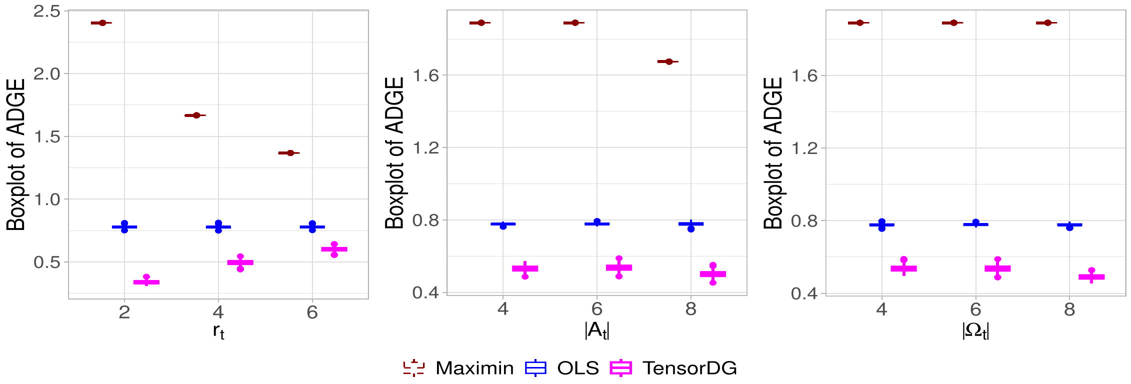


Figure 3. Boxplots of the square-root of ADGE based on Maximin (dashed brown), OLS (solid blue), and TensorDG (bold solid magenta) in experiments (a), (b), and (c).

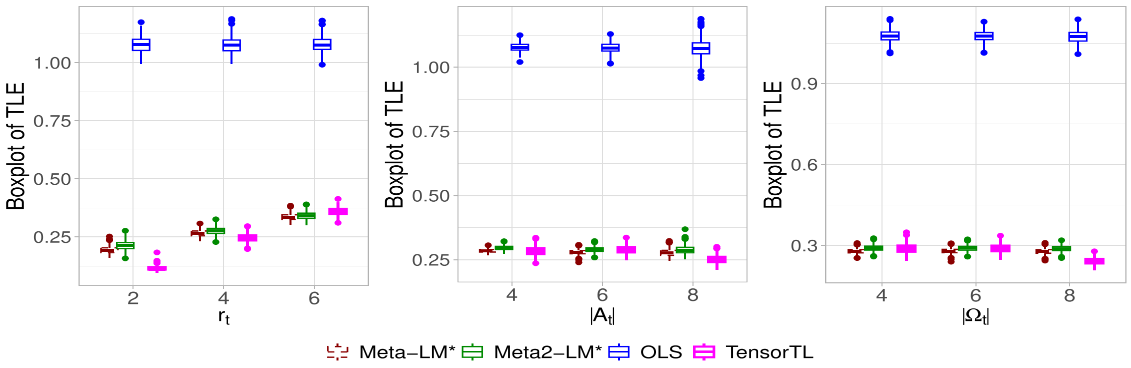


Figure 4. Boxplots of TLE based on Meta-LM* (dashed brown), Meta2-LM* (solid green), OLS(solid blue), and TensorTL (bold solid magenta) in experiments (a), (b), and (c) with $\|\boldsymbol{\delta}^{(g^*)}\|_0 = 0$.

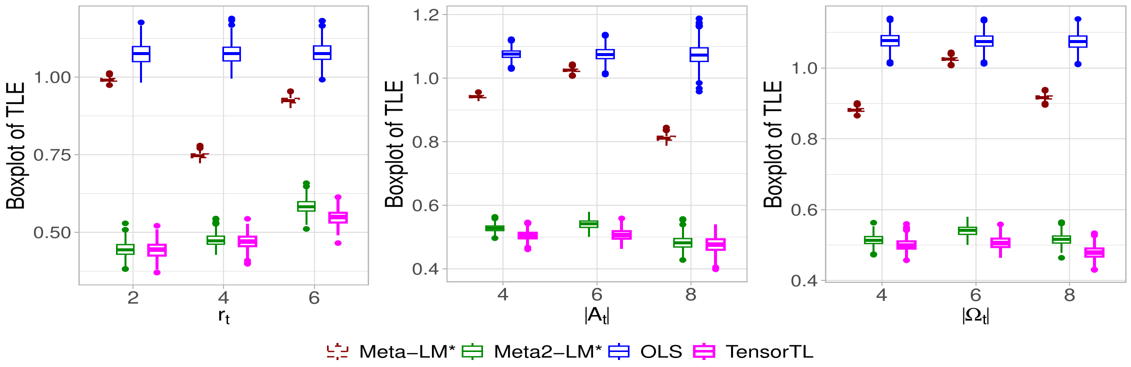


Figure 5. Boxplots of TLE based on Meta-LM* (dashed brown), Meta2-LM* (solid green), OLS (solid blue), and TensorTL (bold solid magenta) in experiments (a), (b), and (c) with $\|\boldsymbol{\delta}^{(g^*)}\|_0 = 3$.

(2021), which estimates the linear subspace of $\mathcal{M}_0[\boldsymbol{\beta}(\mathcal{G})]$ based on Method-of-Moments (MoM). We modify the original Meta-LM-MoM by changing the MoM step to our proposed Algorithm 2 because we find that our proposal can significantly improve the estimation accuracy over MoM. We call this transfer learning method “Meta-LM*”. We also consider “Meta2-LM*” which adds a bias-correction step to “Meta-LM*”. The detailed implementation is given in Section G.1 in the supplementary files. We consider same setting as in Section 6.1 except that $n^{(g)} = 150$ for $g \in \mathcal{O}^c$. This setup is due to in transfer learning settings, the data from the target domain is always very limited. For $\boldsymbol{\gamma}^{(g^*)}$ defined in (18), we consider $\|\boldsymbol{\delta}^{(g^*)}\|_0 \in \{0, 3\}$, respectively. If $\delta_j^{(g^*)} \neq 0$, we simulate $\delta_j^{(g^*)} \sim N(0, 0.25)$ independently. We report the boxplot of its Transfer Learning Error (TLE) $\|\hat{\boldsymbol{b}}^{(g^*)} - \boldsymbol{\gamma}^{(g^*)}\|_2$ for all $g^* \in \mathcal{O}^c$ for a generic estimate of $\boldsymbol{\gamma}^{(g^*)}, \hat{\boldsymbol{b}}^{(g^*)}$.

From Figure 4, we see that Meta-LM*, Meta2-LM*, and TensorTL improve over single-task OLS when the low-rank tensor model is correctly specified. Meta-LM* and Meta2-LM* have slightly larger estimation errors than TensorTL in most settings. This is because its accuracy relies on $n^{(g^*)}$ which is relatively small in these experiments. In contrast, the performance of TensorTL is not limited by $n^{(g^*)}$ when $\boldsymbol{\delta}^{(g^*)} = 0$ by Theorem 4. In Figure 5, we consider the case where the low-rank tensor model is mis-specified. We see that Meta-LM* is worse than Meta2-LM* and TensorTL. This demonstrates the effectiveness of the bias-correction step in TensorTL.

7. Real Data Application

We apply the proposed methods to the Diabetes Health Indicators Dataset, which contains the data collected from the Behavioral Risk Factor Surveillance System, a health-related telephone survey, collected by the U.S. Centers for Disease Control and Prevention in 2015. This dataset has 70,692 samples and 22 covariates. The goal is to predict whether an individual has diabetes or not. The data is publicly available at *https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset*.

We consider two-dimensional group indices $\boldsymbol{g} = (j, k)$, where $j = 1, \ldots, 12$ denotes the age group of an individual and $k = 1, \ldots, 5$ denotes the education level of an individual. The original data has 13 age groups and 6 education levels. As the lowest age group and the lowest education group contain few observations, we combine the first and second age group and combine the lowest two education levels. This formulation leads to a multi-task classification problem with $|\mathcal{G}| = 12 \times 5 = 60$ and $p = 19$. We define $\mathcal{O}$ as the set of groups whose group size is larger than 100. The missing patterns are displayed in Figure 6.

For each $g \in \mathcal{G} \setminus \mathcal{O}$, we use a random half of the samples, denoted as $\mathcal{N}_{tr}^{(g)}$, for training if necessary and the other half of the samples $\mathcal{N}_{te}^{(g)}$ as the test data. For the domain generalization task, we compare the performance of single-task logistic regression, Maximin, and TensorDG. The single-task logistic performs ℓ_1-penalized logistic regression based on samples in $\mathcal{N}_{tr}^{(g)}$ for each $g \notin \mathcal{O}$. For Maximin and TensorDG, we replace the least squares with ℓ_1-penalized logistic regression and they are trained solely based on the samples in $\mathcal{O}$. For the transfer learning task, we compare the performance of single-task logistic, Meta-Logistic, Meta2-Logistic, and TensorTL, where Meta-Logistic and Meta2-Logistic are analogous to Meta-LM* and Meta2-LM*, respectively, except that the linear regression is replaced by ℓ_1-penalized logistic regression.

For each estimate $\boldsymbol{b}$ and group g, we evaluate the average classification error over the domains $g \in \mathcal{G} \setminus \mathcal{O}$ based on the test samples.

$$\mathrm{ACE}(\boldsymbol{b}, g) = \frac{1}{|\mathcal{N}_{te}^{(g)}|} \sum_{i \in \mathcal{N}_{te}^{(g)}} |y_i^{(g)} - \mathbb{1}(\boldsymbol{b}^\top \boldsymbol{x}_i^{(g)} > 0.5)|.$$

The classification results are reported in Figure 7. We see that TensorDG has significant improvement over Maximin for domain generalization. It is also better than the benchmark method, single-task logistic regression. TensorTL also improves Meta-Logistic and Meta2-Logistic for transfer learning. We also diagnose the low-rank assumption (Condition 2) in the supplement (Section H) and the results show that the is no obvious violation of the low-rank tensor assumption.

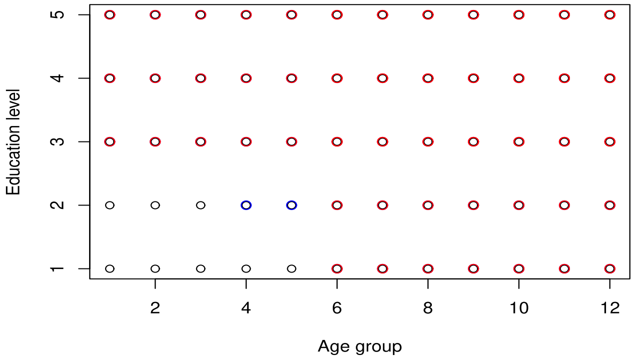


Figure 6. Group structure in the diabetes prediction dataset. Red circles: groups in the arm and body sets. Blue circles: groups in $\mathcal{O}$ but not used in the arm and body sets. Black circles: groups not in $\mathcal{O}$. Specifically, the body set is $\Omega = \{6, 7, 8, 9, 10, 11, 12\} \times \{3, 4, 5\}$. The arm sets are $\mathcal{A}_1 = \{3, 4, 5\}$ and $\mathcal{A}_2 = \{6, 7, 8, 9, 10, 11, 12\}$.

8. Discussion

We study domain generalization and transfer learning with multi-dimensional group indices in linear models. Based on a low-rank tensor model, we develop rate optimal methods for domain generalization. The proposed framework can be extended to deal with binary or categorical outcomes based on other machine learning methods. As deep neural networks have shown significant successes in practice, a direction of interest is to extend the current model to deep neural nets where each layer of neural networks across different domains forms a low-rank tensor. The technical tools developed in this article can potentially apply to these cases to facilitate developing domain generalization in neural networks and other machine learning methods with provable guarantees.

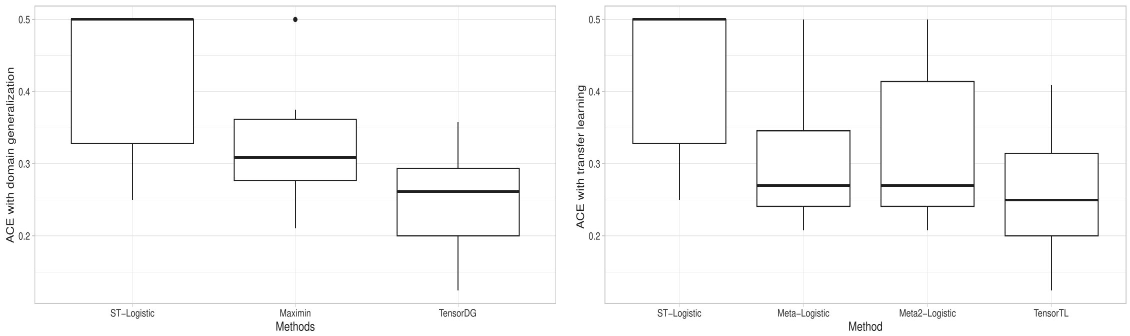


Figure 7. Boxplots of ACEs for groups in $\mathcal{O}^c$ in the domain generalization setting (left) and transfer learning setting (right) based on different methods for diabetes prediction.

Supplementary Materials

Supplement to "Multi-dimensional domain generalization with low-rank structures". In the Supplementary Materials, we provide the proofs of theorems and further results on simulations and data applications.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

SL's research was supported in part by the National Natural Science Foundation of China (grant no. 12201630). LZ's research was supported in part by NSF grant DMS-2340241.

ORCID

Sai Li http://orcid.org/0000-0002-6362-3593

References

Adomavicius, G., and Tuzhilin, A. (2010), "Context-Aware Recommender Systems," in *Recommender Systems Handbook*, pp. 217–253, Boston: Springer. [2]

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019), "Invariant Risk Minimization," arXiv preprint arXiv:1907.02893. [2]

Baktashmotlagh, M., Harandi, M. T., Lovell, B. C., and Salzmann, M. (2013), "Unsupervised Domain Adaptation by Domain Invariant Projection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 769–776. [2]

Bi, X., Tang, X., Yuan, Y., Zhang, Y., and Qu, A. (2021), "Tensors in Statistics," *Annual Review of Statistics and its Application*, 8, 345–368. [2]

Bühlmann, P. (2020), "Invariance, Causality and Robustness," *Statistical Science*, 35, 404–426. [2]

Chen, H., Raskutti, G., and Yuan, M. (2019), "Non-Convex Projected Gradient Descent for Generalized Low-Rank Tensor Regression," *The Journal of Machine Learning Research*, 20, 172–208. [2]

Chen, R., Yang, D., and Zhang, C.-H. (2022), "Factor Models for High-Dimensional Tensor Time Series," *Journal of the American Statistical Association*, 117, 94–116. [5]

Chen, Y., and Bühlmann, P. (2020), "Domain Adaptation Under Structural Causal Models," arXiv preprint arXiv:2010.15764. [2]

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019), "Inference and Uncertainty Quantification for Noisy Matrix Completion," *Proceedings of the National Academy of Sciences*, 116, 22931–22937. [8]

Chua, K., Lei, Q., and Lee, J. D. (2021), "How Fine-Tuning Allows for Effective Meta-Learning," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 8871–8884. [8]

Curtis, D. (2018), "Polygenic Risk Score for Schizophrenia is More Strongly Associated with Ancestry than with Schizophrenia," *Psychiatric Genetics*, 28, 85–89. [1]

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. (2020), "Few-Shot Learning via Learning the Representation, Provably," arXiv preprint arXiv:2002.09434. [2,3,6,8]

Duan, Y., and Wang, K. (2023), "Adaptive and Robust Multi-Task Learning," *The Annals of Statistics*, 51, 2015–2039. [8]

Fan, J., Fang, C., Gu, Y., and Zhang, T. (2023), "Environment Invariant Linear Least Squares," arXiv preprint arXiv:2303.03092. [2]

Feng, Z., Han, S., and Du, S. S. (2021), "Provable Adaptation Across Multiway Domains via Representation Learning," in *International Conference on Learning Representations*. [2]

Ghosh, A., Chung, J., Yin, D., and Ramchandran, K. (2020), "An Efficient Framework for Clustered Federated Learning," in *Advances in Neural Information Processing Systems* (Vol. 33), pp. 19586–19597. [8]

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011), "Joint Estimation of Multiple Graphical Models," *Biometrika*, 98, 1–15. [6]

Guo, Z. (2023), "Statistical Inference for Maximin Effects: Identifying Stable Associations Across Multiple Studies," *Journal of the American Statistical Association* (just-accepted), 1–32. [2]

Han, Y., Chen, R., Yang, D., and Zhang, C.-H. (2020), "Tensor Factor Model Estimation by Iterative Projection," arXiv preprint arXiv:2006.02611. [5]

Han, Y., Chen, R., and Zhang, C.-H. (2022), "Rank Determination in Tensor Factor Model," *Electronic Journal of Statistics*, 16, 1726–1803. [2,5,8]

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. (2021), "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349. [1]

Hitchcock, F. L. (1927), "The Expression of a Tensor or a Polyadic as a Sum of Products," *Journal of Mathematics and Physics*, 6, 164–189. [3]

Kong, W., Somani, R., Song, Z., Kakade, S., and Oh, S. (2020), "Meta-Learning for Mixed Linear Regression," in *International Conference on Machine Learning*, pp. 5394–5404, PMLR. [8]

Kumar, A., Ma, T., and Liang, P. (2020), "Understanding Self-Training for Gradual Domain Adaptation," i *International Conference on Machine Learning*, pp. 5468–5479, PMLR. [2]

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. (2021), "Predicting What You Already Know Helps: Provable Self-Supervised Learning," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 309–323. [2]

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017), "Deeper, Broader and Artier Domain Generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5542–5550. [2]

Li, S., Cai, T. T., and Li, H. (2022), "Transfer Learning for High-Dimensional Linear Regression: Prediction, Estimation and Minimax Optimality," *Journal of the Royal Statistical Society*, Series B, 84, 149–173. [2,8]

Li, S., Zhang, L., Cai, T. T., and Li, H. (2023), "Estimation and Inference for High-Dimensional Generalized Linear Models with Knowledge Transfer," *Journal of the American Statistical Association*, 119, 1274–1285. [8]

Lotfollahi, M., Dony, L., Agarwala, H., and Theis, F. (2021), "Out-of-Distribution Prediction with Disentangled Representations for Single-Cell RNA Sequencing Data," bioRxiv, 2021–09. [1]

Meinshausen, N., and Bühlmann, P. (2015), "Maximin Effects in Inhomogeneous Large-Scale Data," *The Annals of Statistics*, 43, 1801–1830. [2]

Montanari, A., and Sun, N. (2018), "Spectral Algorithms for Tensor Completion," *Communications on Pure and Applied Mathematics*, 71, 2381–2425. [2]

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014), "Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery," in *International Conference on Machine Learning*, pp. 73–81, PMLR. [2]

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. (2021), "Stabilizing Variable Selection and Regression," *The Annals of Applied Statistics*, 15, 1220–1246. [2]

Raskutti, G., Yuan, M., and Chen, H. (2019), "Convex Regularization for High-Dimensional Multiresponse Tensor Regression," *The Annals of Statistics*, 47, 1554–1584. [2]

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018), "Invariant Models for Causal Transfer Learning," *The Journal of Machine Learning Research*, 19, 1309–1342. [2]

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. (2020), "The Risks of Invariant Risk Minimization," in *International Conference on Learning Representations*. [2]

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019), "Distributionally Robust Neural Networks," in *International Conference on Learning Representations*. [2]

Sharifi-Noghabi, H., Harjandi, P. A., Zolotareva, O., Collins, C. C., and Ester, M. (2021), "Out-of-Distribution Generalization from Labelled and Unlabelled Gene Expression Data for Drug Response Prediction," *Nature Machine Intelligence*, 3, 962–972. [1]

Simchowitz, M., Gupta, A., and Zhang, K. (2023), "Tackling Combinatorial Distribution Shift: A Matrix Completion Perspective," in *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3356–3468, PMLR. [6]

Tian, Y., and Feng, Y. (2022), "Transfer Learning Under High-Dimensional Generalized Linear Models," *Journal of the American Statistical Association*, 118, 2684–2697. [8]

Tian, Y., Gu, Y., and Feng, Y. (2023), "Learning from Similar Linear Representations: Adaptivity, Minimaxity, and Robustness," arXiv preprint arXiv:2303.17765. [8]

Tripuraneni, N., Jin, C., and Jordan, M. (2021), "Provable Meta-Learning of Linear Representations," in *International Conference on Machine Learning*, pp. 10434–10443. PMLR. [3,6,8,9,10]

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. (2018), "Generalizing to Unseen Domains via Adversarial Data Augmentation," in *Advances in Neural Information Processing Systems* (Vol. 31). [2]

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Yu, P. (2022), "Generalizing to Unseen Domains: A Survey on Domain Generalization," *IEEE Transactions on Knowledge and Data Engineering*, 35, 8052–8072. [1]

Wimalawarne, K., Sugiyama, M., and Tomioka, R. (2014), "Multitask Learning Meets Tensor Factorization: Task Imputation via Convex Optimization," in *Advances in Neural Information Processing Systems* (Vol. 27). [2]

Woodward, A. A., Urbanowicz, R. J., Naj, A. C., and Moore, J. H. (2022), "Genetic Heterogeneity: Challenges, Impacts, and Methods through an Associative Lens," *Genetic Epidemiology*, 46, 555–571. [1]

Xia, D., Yuan, M., and Zhang, C.-H. (2021), "Statistically Optimal and Computationally Efficient Low Rank Tensor Completion from Noisy Entries," *The Annals of Statistics*, 49, 76–99. [2,4]

Zhang, A. (2019), "Cross: Efficient Low-Rank Tensor Completion," *The Annals of Statistics*, 47, 936–964. [2,5,7]

Zhang, A., and Xia, D. (2018), "Tensor svd: Statistical and Computational Limits," *IEEE Transactions on Information Theory*, 64, 7311–7338. [5]

Zhang, A. R., Luo, Y., Raskutti, G., and Yuan, M. (2020), "Islet: Fast and Optimal Low-Rank Tensor Regression via Importance Sketching," *SIAM Journal on Mathematics of Data Science*, 2, 444–479. [2,5]

Zhang, R., Xu, Q., Yao, J., Zhang, Y., Tian, Q., and Wang, Y. (2023), "Federated Domain Generalization with Generalization Adjustment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3954–3963. [1]

Zhou, H., Li, L., and Zhu, H. (2013), "Tensor Regression with Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [2]

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022), "Domain Generalization: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 4396–4415. [1]

Zhou, X., Lin, Y., Zhang, W., and Zhang, T. (2022), "Sparse Invariant Risk Minimization," in *International Conference on Machine Learning*, pp. 27222–27244, PMLR. [2]