Transfer Learning Meets Functional Linear Regression: No Negative Transfer Under Posterior Drift

Xiaoyu Hu[1], **Zhenhua Lin**[2]

[1]School of Mathematics and Statistics, Xi'an Jiaotong University
[2]Department of Statistics and Data Science, National Univeristy of Singapore
huxiaoyu2019@gmail.com, linz@nus.edu.sg

Abstract

Posterior drift refers to changes in the relationship between responses and covariates while the distributions of the covariates remain unchanged. In this work, we explore functional linear regression under posterior drift with transfer learning. Specifically, we investigate when and how auxiliary data can be leveraged to improve the estimation accuracy of the slope function in the target model when posterior drift occurs. We employ the approximated least square method together with a lasso penalty to construct an estimator that transfers beneficial knowledge from source data. Theoretical analysis indicates that our method avoids negative transfer under posterior drift, even when the contrast between slope functions is quite large. Specifically, the estimator is shown to perform at least as well as the classical estimator using only target data, and it enhances the learning of the target model when the source and target models are sufficiently similar. Furthermore, to address scenarios where covariate distributions may change, we propose an adaptive algorithm using aggregation techniques. This algorithm is robust against non-informative source samples and effectively prevents negative transfer. Simulation and real data examples are provided to demonstrate the effectiveness of the proposed algorithm.

1 Introduction

Functional data analysis (FDA) has gained increasing attention over the past two decades. Two monographs, Ramsay and Silverman (2005) and Hsing and Eubank (2015), provide comprehensive treatments on methodologies and theories of FDA. In particular, functional linear regression has emerged as a crucial tool and has been extensively studied in the literature, including Yao, Müller, and Wang (2005); Cai and Hall (2006); Hall and Horowitz (2007); Li and Hsing (2007); Crambes, Kneip, and Sarda (2009); Yuan and Cai (2010). The aforementioned methodologies are developed and can be successful when there is sufficient training data for the target task. However, in many real-world applications, the available training data is often limited, leading to unsatisfactory estimation results. Fortunately, samples from different but related sources can provide beneficial information to boost performance on the target problem. The process of transferring knowledge from additional data to improve the task on the target data is a popular topic known as transfer learning (Pan and Yang 2009; Zhuang et al. 2020).

Transfer learning has been widely applied to various tasks, including text classification (Xue et al. 2008), recommendation systems (Pan and Yang 2013), and medical diagnosis (Hajiramezanali et al. 2018). Although many methodologies have been developed in the machine learning community, less attention is paid to the statistical properties and theoretical guarantees. Recently, some works have begun to explore transfer learning algorithms in different statistical models (Cai and Wei 2021; Li, Cai, and Li 2022; Tian and Feng 2023; Tian, Weng, and Feng 2022; Zhang and Zhu 2022; Li et al. 2024; Li, Cai, and Li 2023; Jin et al. 2024). However, in the context of functional linear regression, the estimation performance under transfer learning remains unclear. For prediction, Lin and Reimherr (2024) considered using reproducing kernel Hilbert spaces (RKHS) and derived the error bound on the excess prediction risk under transfer learning. In contrast, estimation in functional linear regression is a more difficult problem than prediction, as measured by the minimax optimal rate (Cai and Hall 2006; Hall and Horowitz 2007). The impact on slope estimation, which matters for understanding the effect of functional explanatory variables on the response, has not yet been investigated in the context of transfer learning.

In this paper, we study the slope estimation problem in functional linear regression under the transfer learning setting. Specifically, we focus on the case of posterior drift, where the relationship between responses and covariates changes while the covariate distributions remain unchanged (Kouw and Loog 2018; Cai and Wei 2021; Li, Cai, and Li 2022; Maity et al. 2024). Notably, we can relax the condition of unchanged covariate distributions to aligned eigenspaces of the functional covariates, see Sections 2 and 3 for details. Since the slope function is intrinsically infinite-dimensional, truncation is indispensable to balance bias and variance during estimation. Our main idea involves constructing the estimator using the approximated least square method. We project all functional covariates onto the common eigenspace estimated from the auxiliary data and define the transformed variables as scores. We use the scores from the source data to obtain an initial estimator, then correct the bias with a lasso penalty using the target data (Tibshirani 1996).

Moreover, we derive the convergence rate of the proposed estimator and demonstrate that there is no negative transfer in the presence of posterior drift. When the contrast between slope functions is sufficiently small, the proposed algorithm effectively improves estimation performance on the target model by transferring knowledge from source data. It is noteworthy that the truncation parameter depends not only on the smoothness of the slope function and the decay rate of the eigenvalues of covariance functions, but also on the relatedness between target and source models.

The theoretical challenges mainly lie in the following aspects. First, the projection scores of the target data are correlated, differing from the conventional case in Hall and Horowitz (2007). Second, the truncation level is potentially larger than the target sample size due to additional information from auxiliary data, which brings new theoretical issues. To tackle these challenges, we employ the oracle inequality to quantify the error bounds and generalize the results regarding the restricted eigenvalue condition in high-dimensional regression to the functional data setting (Raskutti, Wainwright, and Yu 2010; Negahban et al. 2012).

In practice, we may not know whether the covariate distributions are equal or share an aligned eigenspace. To avoid performance deterioration, we provide an adaptive algorithm utilizing aggregation techniques. We demonstrate its robustness and effectiveness in alleviating negative transfer caused by non-informative source samples through numerical studies.

The main contributions of this paper are summarized as follows:

- We propose an effective algorithm to enhance the learning of the target slope function by transferring knowledge from source samples under posterior drift.
- We theoretically demonstrate the absence of negative transfer in slope estimation for functional linear regression under posterior drift.
- Under scenarios where covariate distributions may differ substantially, we propose an adaptive algorithm through sparse aggregation to prevent performance degradation.

Notations. We use bold letters to denote vectors. For a given p-dimensional vector $\boldsymbol{w} = (w_1, \ldots, w_p)^{\mathrm{T}} \in \mathbb{R}^p$, the l_q norm is given by $\|\boldsymbol{w}\|_q = (\sum_{j=1}^p |w_j|^q)^{1/q}, q > 0$ and $\|\boldsymbol{w}\|_\infty = \max_j |w_j|$. For a function $f : [0,1] \to \mathbb{R}$, let $\|f\|_2^2 = \int_0^1 f^2(t)dt$. For a random variable ξ, define the ψ_p-Orlicz norm by $\|\xi\|_{\psi_p} = \inf\{c > 0 : E\exp(|\xi|^p/c^p) \le 2\}$ for $p \ge 1$. For $a, b \in \mathbb{R}$, define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For two deterministic and non-negative sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$, we use $a_n \ll b_n$ or $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. And $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $\sup_n a_n/b_n < \infty$. Let $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For two random sequences $\{x_n\}_{n=1}^\infty$ and $\{y_n\}_{n=1}^\infty$, let $x_n = O_P(y_n)$ denote $P(|x_n/y_n| \le c) \to 1$ for some finite constant $c > 0$ and $x_n = o_P(y_n)$ denote $P(|x_n/y_n| > c) \to 0$ for any constant $c > 0$. Unless otherwise stated, let $c, c_1, c_2, \ldots$ and $C, C_1, C_2, \ldots$ denote positive constants, not depending on the sample sizes $n, n_1, \ldots, n_L$. We allow c and C to be different at different appearances.

2 Methodology Under Posterior Drift

Models

Given the target distribution Q, the observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent and identically distributed (i.i.d.) from Q, where $X_i \in L^2(\mathcal{T})$ is the random covariate function in the space of square integrable functions on a compact interval $\mathcal{T}$ and Y_i is the scale response. Without loss of generality, let $\mathcal{T} = [0,1]$. The target model is

$$Y_i - EY = \int_{\mathcal{T}} b(t)\big(X_i(t) - \mu(t)\big)dt + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $b(t)$ is an unknown slope function, $\mu(t) = \mathbb{E}\{X(t)\}$ and ϵ_i is random noise with $\mathbb{E}(\epsilon_i|X_i) = 0$ and variance σ^2, independent of the covariates.

In the context of transfer learning, we observe additional data from source distributions $P^{(l)}$, where $l = 1, \ldots, L$. Denote $\mathcal{A} = \{1, \ldots, L\}$. The independent random samples $(X_1^{(l)}, Y_1^{(l)}), \cdots, (X_{n_l}^{(l)}, Y_{n_l}^{(l)}) \overset{i.i.d.}{\sim} P^{(l)}$ are generated by the following source models,

$$Y_i^{(l)} - \mathbb{E}(Y^{(l)}) = \int_{\mathcal{T}} w^{(l)}(t)\big(X_i^{(l)}(t) - \mu^{(l)}(t)\big)dt + \epsilon_i^{(l)}, \tag{2}$$

$i = 1, \ldots, n_l$, where $w^{(l)}(t)$ is an unknown slope function, $\mu^{(l)}(t) = \mathbb{E}\{X^{(l)}(t)\}$ and $\epsilon_i^{(l)}$ is random noise with $\mathbb{E}(\epsilon_i^{(l)}|X_i^{(l)}) = 0$ and variance $(\sigma^2)^{(l)}$, independent of the covariates.

Contrast Under Posterior Drift

Denote the covariance functions by $K(s,t) = \mathrm{cov}(X(s), X(t))$, $K^{(l)}(s,t) = \mathrm{cov}(X^{(l)}(s), X^{(l)}(t))$. Under posterior drift where $K^{(l)}(s,t)$ and $K(s,t)$ share an aligned eigenspace, the spectral expansion is given by

$$K(s,t) = \sum_k \lambda_k \phi_k(s)\phi_k(t),$$

$$K^{(l)}(s,t) = \sum_k \lambda_k^{(l)} \phi_k(s)\phi_k(t),$$

where $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$, $\lambda_1^{(l)} \ge \lambda_2^{(l)} \ge \cdots \ge 0$ and $\int_{\mathcal{T}} \phi_k(t)\phi_l(t)dt = \mathbb{1}(k = l)$. Consequently, the Karhunen-Loève expansion is

$$X_i(t) = \mu(t) + \sum_k \xi_{i,k}\phi_k(t), \;\; i = 1, \ldots, n,$$

$$X_i^{(l)}(t) = \mu^{(l)}(t) + \sum_k \xi_{i,k}^{(l)}\phi_k(t), \;\; i = 1, \ldots, n_l,$$

where $\xi_{i,k} = \int_{\mathcal{T}} \big(X_i(t) - \mu(t)\big)\phi_k(t)dt$ and $\xi_{i,k}^{(l)} = \int_{\mathcal{T}} \big(X_i^{(l)}(t) - \mu^{(l)}(t)\big)\phi_k(t)dt$. Given the complete orthonormal basis $\phi_1, \phi_2, \ldots$, write

$$w^{(l)}(t) = \sum_{k=1}^\infty w_k^{(l)}\phi_k(t), \quad b(t) = \sum_{k=1}^\infty b_k\phi_k(t).$$

In general, the slope functions $w^{(l)}(t)$ are different from $b(t)$. Let $\delta_k^{(l)} = b_k - w_k^{(l)}$ and $\|\boldsymbol{\delta}^{(l)}\|_1 = \sum_{k=1}^{\infty} |\delta_k^{(l)}|$ for $l = 1, \ldots, L$. Note that the contrast $\|\boldsymbol{\delta}^{(l)}\|_1$ measures the relatedness between the target model and source models. Clearly, the smaller the contrast is, the more information can be transferred from the source data. Intuitively, if we expect to improve the target estimation by borrowing information from source data, the source models should be sufficiently close to the target model, that is, the contrast should be sufficiently small. This will be discussed further in Section 3.

Transfer Learning Under Posterior Drift

To handle the infinite dimensionality, we first perform functional principal component analysis (FPCA) on the source data to estimate the eigenfunctions and obtain the projected score variables for both source and target data. Denote

$$\hat{K}^{(l)}(s,t) = \frac{1}{n_l - 1} \sum_{i=1}^{n_l} \left\{ X_i^{(l)}(s) - \bar{X}^{(l)}(s) \right\} \left\{ X_i^{(l)}(t) - \bar{X}^{(l)}(t) \right\},$$

$$\hat{K}^{\mathcal{A}}(s,t) = \sum_{l=1}^{L} \pi_l \hat{K}^{(l)}(s,t), \tag{3}$$

where $\bar{X}^{(l)}(t) = n_1^{-1} \sum_{i=1}^{n_1} X_i^{(l)}(t)$ and $\pi_l = n_l/N$, $N = \sum_{l=1}^{L} n_l$. Note that

$$\hat{K}^{\mathcal{A}}(s,t) = \sum_{k} \hat{\lambda}_k^{\mathcal{A}} \hat{\phi}_k^{\mathcal{A}}(s) \hat{\phi}_k^{\mathcal{A}}(t), \tag{4}$$

where $\hat{\lambda}_1^{\mathcal{A}} \geq \hat{\lambda}_2^{\mathcal{A}} \geq \cdots$ are eigenvalues and $\hat{\phi}_1^{\mathcal{A}}, \hat{\phi}_2^{\mathcal{A}}, \cdots$ are corresponding eigenfunctions. Let

$$\begin{aligned} \hat{\xi}_{i,k}^{(l)} &= \int_{\mathcal{T}} \big(X_i^{(l)}(t) - \bar{X}^{(l)}(t) \big) \hat{\phi}_k^{\mathcal{A}}(t) dt, \ i = 1, \ldots, n_l, \\ \hat{\xi}_{i,k} &= \int_{\mathcal{T}} \big(X_i(t) - \bar{X}(t) \big) \hat{\phi}_k^{\mathcal{A}}(t) dt, \ i = 1, \ldots, n, \end{aligned} \tag{5}$$

where $\bar{X}(t) = n^{-1} \sum_{i=1}^{n} X_i(t)$.

We introduce the following notations for vectors and matrices. Let $\boldsymbol{Y}^{(l)} = (Y_1^{(l)}, \ldots, Y_{n_l}^{(l)})^{\mathrm{T}}$ and $\bar{\boldsymbol{Y}}^{(l)}$ be an n_l-dimensional vector with each element equal to $\bar{Y}^{(l)} = n_l^{-1} \sum_{i=1}^{n_l} Y_i^{(l)}$. Define $\boldsymbol{Y}$ and $\bar{\boldsymbol{Y}}$ similarly for the target sample. Denote

$$\hat{\Xi}^{(l)} = \begin{pmatrix} \hat{\xi}_{1,1}^{(l)} & \cdots & \hat{\xi}_{1,m}^{(l)} \\ \hat{\xi}_{2,1}^{(l)} & \cdots & \hat{\xi}_{2,m}^{(l)} \\ \vdots & \ddots & \vdots \\ \hat{\xi}_{n_l,1}^{(l)} & \cdots & \hat{\xi}_{n_l,m}^{(l)} \end{pmatrix},$$

where the truncation parameter m is allowed to grow with the sample sizes. Define $\hat{\Xi} \in \mathbb{R}^{n \times m}$ analogously.

We use the approximated least square method to transfer knowledge from the source data and obtain an initial estimator. Since the slope functions of the target model and the source models are generally different, this initial estimator may be biased. To correct the bias, we use the target data and apply the approximated least square with a lasso penalty. The proposed algorithm is presented in Algorithm 1.

Algorithm 1: The transfer learning algorithm under posterior drift.

Input: Target data $(X_i, Y_i), i = 1, \ldots, n$ and auxiliary data $(X_i^{(l)}, Y_i^{(l)}), i = 1, \ldots, n_l; l = 1, \ldots, L$.

Compute $\hat{K}^{\mathcal{A}}(s,t)$ as defined in (3) and obtain the estimates $\hat{\lambda}_k^{\mathcal{A}}, \hat{\phi}_k^{\mathcal{A}}$ in (4).

Obtain the score variables $\hat{\xi}_{i,k}$ and $\hat{\xi}_{i,k}^{(l)}$ in (5), $k = 1, \ldots, m$.

Step 1: An initial estimator.

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^m}{\arg\min} \sum_{l=1}^{L} \pi_l (n_l - 1)^{-1} \|\boldsymbol{Y}^{(l)} - \bar{\boldsymbol{Y}}^{(l)} - \hat{\Xi}^{(l)} \boldsymbol{w}\|_2^2.$$

Step 2: Bias correction.

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta} \in \mathbb{R}^m}{\arg\min} \frac{1}{2n} \|\boldsymbol{Y} - \bar{\boldsymbol{Y}} - \hat{\Xi} \hat{\boldsymbol{w}} - \hat{\Xi} \boldsymbol{\delta}\|_2^2 + \tau \|\boldsymbol{\delta}\|_1,$$

where $\tau \geq 0$. Let $\hat{b}_k = \hat{w}_k + \hat{\delta}_k$.

Output: $\hat{b}(t) = \sum_{k=1}^{m} \hat{b}_k \hat{\phi}_k^{\mathcal{A}}(t)$.

In Step 1, the true parameter of interest is $\boldsymbol{w} = \big(\sum_{l=1}^{L} \pi_l \Sigma^{(l)} \big)^{-1} \big(\sum_{l=1}^{L} \pi_l \mathbb{E}(\boldsymbol{\xi}^{(l)} Y^{(l)}) \big)$, where $\Sigma^{(l)}$ is a diagonal matrix with elements $\lambda_1^{(l)}, \ldots, \lambda_m^{(l)}$ and $\boldsymbol{\xi}^{(l)} = (\xi_1^{(l)}, \ldots, \xi_m^{(l)})^{\mathrm{T}}$. In step 2, the targeted parameter is $\boldsymbol{\delta} = \boldsymbol{b} - \boldsymbol{w} = \big(\sum_{l=1}^{L} \pi_l \Sigma^{(l)} \big)^{-1} \sum_{l=1}^{L} (\pi_l \Sigma^{(l)} \boldsymbol{\delta}^{(l)})$, where $\boldsymbol{b} = (b_1, \ldots, b_m)^{\mathrm{T}}$ and $\boldsymbol{\delta}^{(l)} = (\delta_1^{(l)}, \ldots, \delta_m^{(l)})$. By computation, we obtain

$$\hat{\boldsymbol{w}} = \left(\sum_{l=1}^{L} \pi_l (n_l - 1)^{-1} \hat{\Xi}^{(l)\mathrm{T}} \hat{\Xi}^{(l)} \right)^{-1} \left\{ \sum_{l=1}^{L} \pi_l (n_l - 1)^{-1} \hat{\Xi}^{(l)\mathrm{T}} (\boldsymbol{Y}^{(l)} - \bar{\boldsymbol{Y}}^{(l)}) \right\}.$$

Based on the estimation procedure for $\hat{K}^{\mathcal{A}}$ and $\hat{\phi}_k^{\mathcal{A}}$, it follows that $\sum_{l=1}^{L} \pi_l (n_l - 1)^{-1} \hat{\Xi}^{(l)\mathrm{T}} \hat{\Xi}^{(l)}$ is a diagonal matrix. However, the $\hat{\xi}_{i,k}$'s in Step 2 may be correlated, meaning that $\hat{\Xi}^{\mathrm{T}} \hat{\Xi}$ is not necessarily a diagonal matrix, which differs from the conventional case in Hall and Horowitz (2007). Furthermore, there is no explicit solution for $\hat{\boldsymbol{\delta}}$ due to the lasso penalty.

3 Theoretical Properties

We investigate the theoretical properties of the proposed estimator. To begin with, we provide some necessary conditions. Assumption 1 states that the score variables of the target distribution are sub-Gaussian, which is common in the literature on functional and nonparametric analysis (Lin and Lin 2024+; Tian and Feng 2023). In Assumption 2, we

assume that the covariance functions share common eigenfunctions (Dai, Müller, and Yao 2017). Moreover, Assumption 3 characterizes the decay rate of the eigenvalues of different covariance functions using the parameter α, which simplifies the exposition. Assumptions 2 and 3 are milder than assuming equal covariate distributions under posterior drift. Assumption 4 concerns the slope parameters of interest and the contrast between the target model and source models. Assumption 5 about the moments of the source data is quite standard (Hall and Horowitz 2007).

Assumption 1 *The score vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_m)^{\mathrm{T}}$ is sub-Gaussian, i.e., $\|\boldsymbol{v}^{\mathrm{T}}\boldsymbol{\xi}\|_{\psi_2} \leq K(\boldsymbol{v}^{\mathrm{T}}\Sigma_{\boldsymbol{\xi}}\boldsymbol{v})^{1/2}$, for some constant $K > 0$, any vector $v \in \mathbb{R}^m$ and any integer $m > 0$, where $\Sigma_{\boldsymbol{\xi}}$ is the covariance of $\boldsymbol{\xi}$.*

Assumption 2 *The covariance functions $K(s,t)$ and $K^{(l)}(s,t)$ share common eigenfunctions.*

Assumption 3 *Foe some universal constant $c_1 > 0$, the eigenvalues satisfy $\lambda_k \leq c_1 k^{-\alpha}, \lambda_k - \lambda_{k+1} \geq c_1^{-1}k^{-\alpha-1}$ and $\lambda_k^{(l)} \leq c_1 k^{-\alpha}, \lambda_k^{(l)} - \lambda_{k+1}^{(l)} \geq c_1^{-1}k^{-\alpha-1}$ for $l = 1, \ldots, L$ and $\alpha > 1$.*

Assumption 4 *For some universal constant $c_2 > 0$, assume $b_k \leq c_2 k^{-\beta}, \beta > \alpha/2 + 1$ and $\sum_k |\delta_k^{(l)}| \leq h$ for $l = 1, \ldots, L$.*

Assumption 5 *For some universal constant $c_3 > 0$, assume $X^{(l)}$ has finite fourth moment, $\int_{\mathcal{T}} \mathbb{E}\{\big(X^{(l)}(t)\big)^4\}dt \leq c_3 < \infty$, and $\mathbb{E}(\xi_k^{(l)})^4 \leq c_3\{\mathbb{E}(\xi_k^{(l)})^2\}^2$ for all k. Moreover, $\mathbb{E}(\epsilon^{(l)})^4 \leq c_3 < \infty$.*

In the context of transfer learning, the truncation parameter m can potentially be much larger than n, and the sample correlations between $\hat{\xi}_{ik}$'s are nonzero. These features distinguish the problem from the classical functional linear regression. To tackle these issues, we establish the restricted eigenvalue property under the scenario of functional data, and leverage the oracle inequalities to obtain error bounds.

Denote $\Xi = (\xi_{i,k}) \in \mathbb{R}^{n\times m}$ and D is a diagonal matrix with elements $\lambda_1^{1/2}, \lambda_2^{1/2}, \ldots, \lambda_m^{1/2}$. If $m > n$, the smallest eigenvalue of $\Xi^{\mathrm{T}}\Xi/n$ is 0. This means that small perturbations in $\|\Xi\boldsymbol{v}\|_2/n$ can turn into large changes in $\|\boldsymbol{v}\|_2$ for $\boldsymbol{v} \in \mathbb{R}^m$, leading to unstable solutions. In Proposition 1, we establish the connection between $n^{-1}\|\Xi\boldsymbol{v}\|_2^2$ and $\|D\boldsymbol{v}\|_2^2$ in the functional data setting, which is known as the restricted eigenvalue property in high-dimensional regression analysis (Raskutti, Wainwright, and Yu 2010).

Proposition 1 *Under Assumption 1, for any vector $\boldsymbol{v} \in \mathbb{R}^m$, there exists some constant c_K depending on the sub-Gaussian parameter K, such that*

$$\frac{1}{n}\|\Xi\boldsymbol{v}\|_2^2 \geq \frac{1}{4}\|D\boldsymbol{v}\|_2^2 - \frac{c_K\|D\|_F}{n^{1/2}}\|\boldsymbol{v}\|_1\|D\boldsymbol{v}\|_2, \quad (6)$$

with probability at least $1 - c_4\exp(-c_5 n)$ for some constants $c_4, c_5 > 0$.

With the result in Proposition 1, we quantify the estimation error via oracle inequalities in Theorem 1. Theorem 1 reveals an interesting phenomenon indicating that the sparsity parameter τ plays a crucial role. If we take $\tau \asymp n^{-1/2}$, the rate of convergence consists of several components. First, the error term $m^{1-2\beta}$ represents the bias caused by truncation. Second, the bound $(m^{1+\alpha} + m^3h^2)N^{-1}$ quantifies the estimation error of the initial estimator to its probabilistic limit in Step 1. These two components are standard in functional linear regression (Hall and Horowitz 2007) with $h = 0$. The extra term $\left(n^{-1/2}m^{\alpha}h\right) \wedge h^2$ characterizes the error in Step 2. If we set $\tau = 0$, the rate of convergence is consistent with that in classical functional linear regression using only target data.

Theorem 1 *Suppose Assumptions 1-5 hold. If $\tau \asymp n^{-1/2}$, $N^{-1}m^{2(\alpha+1)} = o(1)$ and $h = O(1)$, then*

$$\|\hat{b}-b\|_2^2 = O_P\left(\frac{m^{\alpha}h}{n^{1/2}} \wedge h^2 + \frac{m^{1+\alpha}}{N} + m^{1-2\beta} + \frac{m^3h^2}{N}\right).$$

If $\tau = 0$, $n^{-1}m^2 = o(1)$ and $N^{-1}m^{\alpha+3} = o(1)$, then

$$\|\hat{b}-b\|_2^2 = O_P\left(\frac{m^{1+\alpha}}{n} + \frac{m^{1+\alpha}}{N} + m^{1-2\beta}\right).$$

The truncation parameter m plays an important role in the final convergence rate, which is determined by the bias-variance trade-off. In Corollary 1, we elucidate the choice of the parameter m and the corresponding rate of convergence under different bias levels h. Compared to the minimax rate $n^{-(2\beta-1)/(\alpha+2\beta)}$ in the conventional functional linear regression, we identify when the transfer learning algorithm improves estimation performance. When $N \gg n$ and $h \ll n^{-(\beta-1/2)/(\alpha+2\beta)}$, the obtained rate of convergence is faster than that of the classical estimator using only target data. This is intuitive, as effective information transfer from the source data is possible only when there are sufficient source data and the source models are sufficiently close to the target model. Note that the truncation level m can be larger than n by taking advantage of the additional information from the source data.

Corollary 1 *Suppose Assumptions 1-5 hold. Assume $n \lesssim N$.*

- *If $h \lesssim N^{-\frac{2\beta-1}{2(\alpha+2\beta)}}$, we take $\tau \asymp n^{-1/2}$ and $m \asymp N^{1/(\alpha+2\beta)}$, then*
$$\|\hat{b} - b\|_2^2 = O_P\big(N^{-\frac{2\beta-1}{\alpha+2\beta}}\big).$$
- *If $N^{-\frac{2\beta-1}{2(\alpha+2\beta)}} \lesssim h \lesssim n^{-\frac{2\beta-1}{2(\alpha+2\beta)}}$, we take $\tau \asymp n^{-1/2}$ and $h^{-2/(2\beta-1)} \lesssim m \lesssim (Nh^2)^{1/(1+\alpha)}$, $m^{2(\alpha+1)}N^{-1} = o(1)$, then*
$$\|\hat{b} - b\|_2^2 = O_P(h^2).$$
- *If $h \gtrsim n^{-\frac{2\beta-1}{2(\alpha+2\beta)}}$, we take $\tau = 0$ and $m \asymp n^{1/(\alpha+2\beta)}$, then*
$$\|\hat{b} - b\|_2^2 = O_P\big(n^{-\frac{2\beta-1}{\alpha+2\beta}}\big).$$

Corollary 1 implies that if the bias h is sufficiently small, the lasso penalty can leverage this and help achieve a faster convergence rate. Otherwise, the ordinary least square suffices for bias correction. Notably, there is no negative

transfer under posterior drift in functional linear regression, which contrasts with other high-dimensional regression problems (Li, Cai, and Li 2022; Tian and Feng 2023). In practice, the tuning parameters m and τ can be determined through cross validation.

4 Adaptive Estimation

In some applications, we may not know whether the functional covariates share aligned eigenspace. If the condition is violated, our estimator may be subject to performance deterioration due to the unaligned eigenfunctions. Similar phenomena have been observed for the FPCA-based approaches in the conventional functional linear regression (Yuan and Cai 2010; Cai and Yuan 2012).

To avoid potential performance degradation when covariate distributions have different eigenspaces, we propose an adaptive algorithm consisting of two main steps. First, we construct a collection of candidate source sets and obtain candidate estimators using Algorithm 1 along with auxiliary samples from these source sets. Second, we perform a sparse aggregation step on these candidate estimators (Gaïffas and Lecué 2011). The aggregated estimator is expected to be not much worse than the best candidate estimator under consideration.

We randomly split the target data into two subsets $\mathcal{D}_1 = \{(X_i, Y_i), i \in \mathcal{I}_1\}$ and $\mathcal{D}_2 = \{(X_i, Y_i), i \in \mathcal{I}_2\}$, where $\tilde{n} = |\mathcal{I}_1|, \bar{n} = |\mathcal{I}_2|$. We use $\mathcal{D}_1$ to construct candidate sets and candidate estimators, and $\mathcal{D}_2$ for aggregation.

Candidate Source Sets

To avoid the unstable inverse of covariance operators and truncation operations during the construction of candidate sets, we model the discrepancy between cross covariance functions, defined as $\zeta^{(l)} = \|g^{(l)} - g\|_2^2$, where $g^{(l)}(t) = \mathrm{cov}\big(X^{(l)}(t), Y^{(l)}\big)$ and $g(t) = \mathrm{cov}\big(X(t), Y\big)$. The statistic $\zeta^{(l)}$ is a viable choice as it typically increases with the contrast $\|\boldsymbol{\delta}^{(l)}\|_1$ and has the potential to characterize differences in eigenspaces. The empirical estimates are given by

$$\hat{\zeta}^{(l)} = \int_0^1 \left\{ \frac{1}{n_l} \sum_{i=1}^{n_l} X_i^{(l)}(t)(Y_i^{(l)} - \bar{Y}^{(l)}) - \frac{1}{|\tilde{n}|} \sum_{i \in \mathcal{I}_1} X_i(t)(Y_i - \bar{Y}) \right\}^2 dt. \tag{7}$$

Then, we construct the candidate sets as

$$\hat{\mathcal{A}}_l = \{1 \le k \le L : \hat{\zeta}^{(k)} \text{is among the first } l \text{ smallest of all}\}, \tag{8}$$

for $l = 1, \ldots, L$.

Sparse Aggregation

For each constructed candidate set $\hat{\mathcal{A}}_l$, we obtain the candidate estimator $\hat{b}_l(t)$ using $\mathcal{D}_1 \cup \hat{\mathcal{A}}_l$ with Algorithm 1 for $l = 0, 1, \ldots, L$. Motivated by the principle of model aggregation, which aims to obtain an estimator not much worse than the best estimator under consideration, we use sparse aggregation to achieve adaptivity and prevent negative transfer. To reduce computational cost, we adopt sparse aggregation without the preselection step as in Gaïffas and Lecué (2011).

Algorithm 2: The adaptive transfer learning algorithm for functional linear regression.

Input: Target data $(X_i, Y_i), i = 1, \ldots, n$ and auxiliary data $(X_i^{(l)}, Y_i^{(l)}), i = 1, \ldots, n_l$.
Randomly split the target data into two sub-samples $\mathcal{D}_1 = \{(X_i, Y_i), i \in \mathcal{I}_1\}$ and $\mathcal{D}_2 = \{(X_i, Y_i), i \in \mathcal{I}_2\}$.
Construct the $L+1$ candidate sets $\hat{\mathcal{A}}_0, \hat{\mathcal{A}}_1, \ldots, \hat{\mathcal{A}}_L$ such that $\hat{\mathcal{A}}_0 = \emptyset$ using (7) and (8).
for $l = 0, \ldots, L$ **do**
Obtain the candidate estimator $\hat{b}_l(t)$ using data $\mathcal{D}_1 \cup \{(X_i^{(l)}, Y_i^{(l)}), i = 1, \ldots, n_l; l \in \hat{\mathcal{A}}_l\}$ with Algorithm 1.
end for
Sparse aggregation:

$$\hat{b}_{sagg}(t) = \hat{\lambda}\hat{b}_{l_{1,\star}}(t) + (1 - \hat{\lambda})\hat{b}_{l_{2,\star}}(t),$$

where $l_{1,\star}, l_{2,\star}$ and $\hat{\lambda}$ are obtained from (9) and (10).
Output: $\hat{b}_{sagg}(t)$.

Define $R_{n,2}(b) = \sum_{i \in \mathcal{I}_2} \{Y_i - \bar{Y}_2 - \int_0^1 b(t)\big(X_i(t) - \bar{X}_2(t)\big)dt\}^2/\bar{n}$, where $\bar{Y}_2 = \sum_{i \in \mathcal{I}_2} Y_i/\bar{n}$ and $\bar{X}_2(t) = \sum_{i \in \mathcal{I}_2} X_i(t)/\bar{n}$. Denote

$$l_{1,\star} = \underset{l=0,1,\ldots,L}{\arg\min}\, R_{n,2}(\hat{b}_l). \tag{9}$$

Moreover,

$$\hat{\lambda}, l_{2,\star} = \underset{\lambda \in [0,1], l=0,\ldots,L}{\arg\min} R_{n,2}(\lambda \hat{b}_{l_{1,\star}} + (1-\lambda)\hat{b}_l). \tag{10}$$

The sparse aggregate estimator is $\hat{b}_{sagg}(t) = \hat{\lambda}\hat{b}_{l_{1,\star}}(t) + (1 - \hat{\lambda})\hat{b}_{l_{2,\star}}(t)$.

Remark 1 *In addition to sparse aggregation, other methods such as exponential aggregation (Rigollet and Tsybakov 2011) and Q-aggregation (Dai, Rigollet, and Zhang 2012) can be utilized for adaptive estimation. However, their aggregate performance heavily depends on the critical temperature parameter. In our numerical experiments, we compare their performance by tuning the temperature parameter through 5-fold cross-validation against the performance of sparse aggregation. Based on numerical results, we advocate for sparse aggregation because of its competitive, robust performance and computational efficiency, which does not require parameter tuning.*

5 Synthetic Data

We conduct several experiments to demonstrate the performance of the transfer learning algorithm. Let $\mathcal{T} = [0, 1]$, $n = 150$, $n_l = 100$ and the total number of source samples $L = 20$. We take $X_i(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik} \phi_k(t)$,

$i = 1, \ldots, n$, where $\lambda_k = k^{-\alpha}$, $\phi_k(t) = 2^{1/2}\cos(k\pi t)$ for $k \geq 1$, and Z_{ik}'s are uniformly distributed on $[-3^{1/2}, 3^{1/2}]$. The target slope function is $b(t) = \sum_{k=1}^{50} b_k \phi_k(t)$ with $b_k = 4k^{-\beta}(-1)^{k+1}$. For the slope functions in the auxiliary samples, we have $w^{(l)}(t) = \sum_{k=1}^{50} w_k^{(l)} \phi_k(t)$ for $l = 1, \ldots, L$. The generation mechanisms of $w_k^{(l)}$ and $X^{(l)}$ will be provided later. The responses Y and $Y^{(l)}$ are generated from (1) and (2), respectively, with $\mathbb{E}Y = \mathbb{E}Y^{(l)} = 0$, $\mu(t) = \mu^{(l)}(t) = 0$ and the errors $\epsilon, \epsilon^{(l)} \sim N(0, \sigma_\epsilon^2)$ where $\sigma_\epsilon = 0.5$. Denote $\mathcal{A}_h = \{1 \leq l \leq L : \|\boldsymbol{\delta}^{(l)}\|_1 \leq h\}$.

Transfer Learning on $\mathcal{A}_h$ with Aligned Eigenspace

We explore the numerical performance of Algorithm 1 when the eigenspaces of the target sample and source samples are aligned.

(I) For $l \in \mathcal{A}_h$, let $X_i^{(l)}(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik}^{(l)} \phi_k(t)$ for $i = 1, \ldots, n_l$ and $w_k^{(l)} = b_k - R_k h/s$ for $k = 1, \ldots, 50$, where $Z_{ik}^{(l)}$'s are generated from $N(0,1)$, s is some positive integer within $[1, 50]$ and R_k's are independent Rademacher random variables.

Let $K = |\mathcal{A}_h|$. We consider different combinations of $s = 1, 5, 20, 50$ and $h = 2, 20, 200$, respectively. The proposed method in Algorithm 1 is denoted by "$\mathcal{A}_h$ TL-FLR". For comparison purposes, we include the FPCA-based estimation (Hall and Horowitz 2007) using only target data, which is denoted by "FLR". We evaluate the performance using the mean integrated squared error (MISE), i.e., $\int_0^1 (\hat{b}(t) - b(t))^2 dt$, which is approximated on a grid of 100 equally spaced points on $[0, 1]$. The tuning parameters of all considered approaches are chosen by 5-fold cross-validation.

As shown in Figure 1, the "$\mathcal{A}_h$ TL-FLR" consistently outperforms "FLR" across all considered scenarios, even when the contrast is extremely large. The numerical results align with theoretical findings in Theorem 1 and Corollary 1. In functional linear regression, when the eigenspaces of the target sample and source samples are aligned, the proposed transfer learning algorithm avoids negative transfer, meaning its performance is at least as good as that of the "FLR" using only target data.

Adaptive Transfer Learning in General Settings

We further include the adaptive method in Algorithm 2, denoted by "Agg TL-FLR", and the naive transfer learning method using all source samples, denoted by "Naive TL-FLR" for comparison under various data generation mechanisms.

(II) For $l = 1, \ldots, L$, let $X_i^{(l)}(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik}^{(l)} \phi_k(t)$, $i = 1, \ldots, n_l$, where $Z_{ik}^{(l)}$'s are generated from $N(0,1)$. For $l \in \mathcal{A}_h$, $w_k^{(l)} = b_k - R_k h/50$; for $l \in \mathcal{A}_h^c$, $w_k^{(l)} = b_k - 40R_k$, where R_k's are independent Rademacher random variables.

(III) For $l \in \mathcal{A}_h$, let $X_i^{(l)}(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik}^{(l)} \phi_k(t)$; for $l \in \mathcal{A}_h^c$, $X_i^{(l)}(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik}^{(l)} \psi_k(t)$, where ψ_k's are

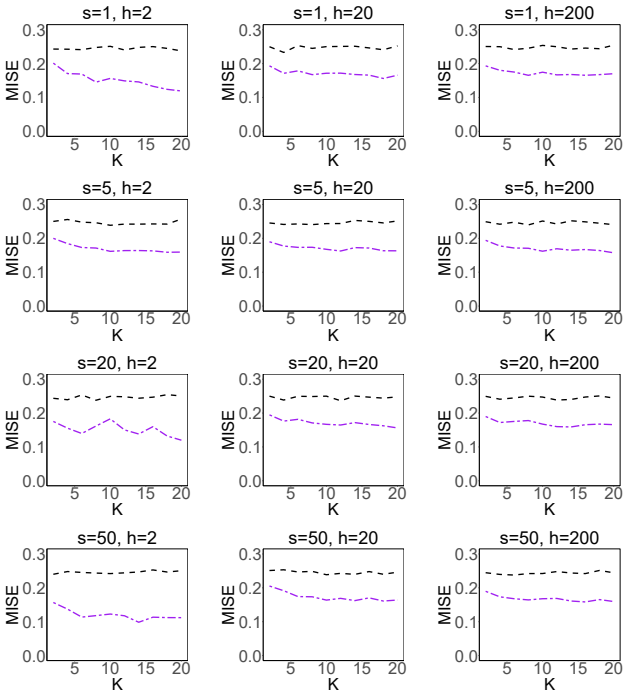

Figure 1: Estimation errors of different methods under Model (I) over 1000 repetitions. FLR: black, dashed; $\mathcal{A}_h$ TL-FLR: purple, dot-dashed.

Haar functions, i.e.,

$$\psi_{2^j+\ell}(t) = \begin{cases} 2^{j/2}, & t \in [\frac{\ell}{2^j}, \frac{\ell+0.5}{2^j}), \\ -2^{j/2}, & t \in [\frac{\ell+0.5}{2^j}, \frac{\ell+1}{2^j}], \\ 0, & \text{otherwise}, \end{cases}$$

for $j = 0, 1, \ldots$ and $\ell = 0, 1, \ldots$. The regression coefficients $w_k^{(l)}$ are the same as in Model (II).

(IV) For $l = 1, \ldots, L$, let $X_i^{(l)}(t) = \sum_{k=1}^{50} \sqrt{\lambda_k} Z_{ik}^{(l)} \psi_k(t)$, $i = 1, \ldots, n_l$, where ψ_k's are the same as in Model (III). The regression coefficients $w_k^{(l)}$ are the same as in Model (II).

Model (II) corresponds to the case where all the source samples share the same eigenspace as the target sample. In contrast, the eigenspaces for each source sample in Model (IV) are completely unaligned with the eigenspace of the target data. In addition, Model (III) represents an intermediate case where the eigenspaces for source samples in $\mathcal{A}_h$ are aligned with the eigenspace of the target data.

The results are depicted in Figure 2. In Model (II) where the eigenspaces are perfectly aligned, the three transfer learning methods consistently outperform "FLR", as there is no negative transfer under this model. The "Agg TL-FLR" exhibits the ability to adaptively select the most informative source samples for estimation, leading to the best performance. In Model (III), the method "Agg TL-FLR" achieves comparable performance to "$\mathcal{A}_h$ TL-FLR". However, the "Naive TL-FLR" leads to inferior performance when K is small, due to the inclusion of the source samples in $\mathcal{A}_h^c$

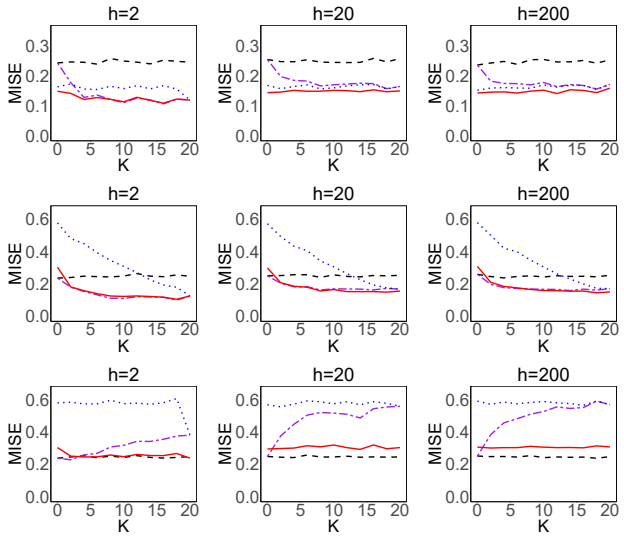

Figure 2: Estimation errors of different methods over 500 repetitions. Top row: Model (II); middle row: Model (III); bottom row: Model (IV). FLR: black, dashed; $\mathcal{A}_h$ TL-FLR: purple, dot-dashed; Naive TL-FLR: blue, dotted; Agg TL-FLR: red, solid.

whose eigenspaces are misaligned with the eigenspace of the target data. In the most challenging case, i.e., Model (IV), as expected, both "Naive TL-FLR" and "$\mathcal{A}_h$ TL-FLR" perform worse than the "FLR", since the eigenspaces in all source samples are different from the eigenspace of the target data. Even the contrast between $b(t)$ and $w^{(l)}(t)$ for $l \in \mathcal{A}_h$ is relatively small, the FPCA based estimation fails because of the inaccurate estimation of the leading eigenspace of the target data. In contrast, the adaptive method "Agg TL-FLR" still performs decently in this extremely challenging scenario, comparable to or a bit worse than the "FLR".

In summary, the proposed algorithm "Agg TL-FLR" adaptively selects the truly informative source samples for estimation, effectively alleviating negative transfer even in the case where all source samples are not informative at all.

6 Real Data

We demonstrate the performance of the proposed method using stock price data from the period 05/01/2021-09/30/2021 (Lin and Reimherr 2024). The dataset consists of 60, 58, 31, 30, 104, 55, 70, 68, 46, 103, 41 stocks from 11 sectors, including basic industries (BI), capital goods (CG), consumer durable (CD), consumer non-durable (CND), consumer services (CS), energy (En), finance (Fin), health care (HC), public utility (PU), technology (Tech) and transportation (Trans), respectively.

We aim to predict the monthly return of the second month using the monthly cumulative return of the first month for a specific sector, leveraging information from other sectors via transfer learning. For a given stock, the daily prices are $\{v(t_0), v(t_1), \ldots, v(t_T)\}$ in the first month and are $\{v'(t_0), v'(t_1), \ldots, v'(t_T)\}$ in the second month. Therefore,

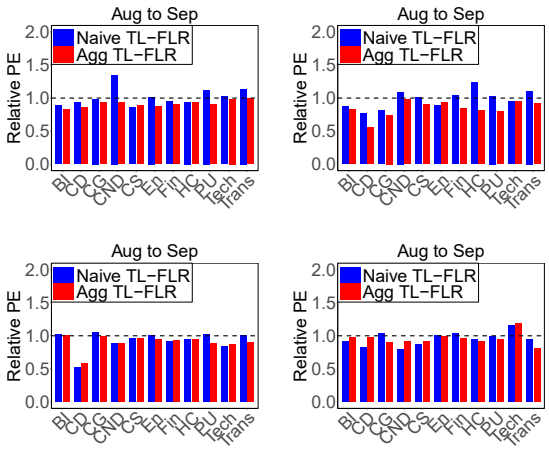

Figure 3: Average relative prediction errors of different methods for each target sector over 500 repetitions.

the covariates and responses are defined as

$$X(t) = \frac{v(t) - v(t_0)}{v(t_0)}, \quad Y = \frac{v'(t_T) - v'(t_0)}{v'(t_0)}.$$

We treat the data from each sector as a target sample, with all other sectors serving as source samples. The target data is randomly split into 80% training data and 20% testing data. The prediction errors are evaluated on the testing data over 500 independent repetitions. We compare the performance of "Naive TL-FLR", "Agg TL-FLR" and "FLR" by reporting the average relative prediction errors, i.e., the prediction errors of "Naive TL-FLR" or "Agg TL-FLR" divided by the prediction errors of "FLR". As illustrated in Figure 3, the "Agg TL-FLR" consistently outperforms the "FLR" in terms of prediction accuracy, with relative prediction errors significantly below 1 in almost all cases. In contrast, the "Naive TL-FLR" often results in much poorer performance compared to the "FLR". These observations support our claim that the adaptive algorithm "Agg TL-FLR" effectively alleviates negative transfer.

7 Conclusion

In this work, we focus on slope estimation in functional linear regression under transfer learning. When the covariate distributions share an aligned eigenspace, we propose an algorithm that avoids negative transfer even when the contrast is large. Our algorithm enhances the learning of the target model, provided that the sample size of the source data is sufficiently large and the contrast between the source and target models is sufficiently small. To mitigate performance degradation when the aligned eigenspace condition is violated, we introduce an adaptive algorithm via constructing potential candidate estimators and performing sparse aggregation. The adaptive algorithm performs well in both synthetic and real data examples.

Acknowledgments

Dr. Lin's research was partially supported by MOE AcRF Tier 1 grant (A-8002518-00-00).

References

Cai, T. T.; and Hall, P. 2006. Prediction in functional linear regression. *The Annals of Statistics*, 34(5): 2159–2179.

Cai, T. T.; and Wei, H. 2021. Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1): 100–128.

Cai, T. T.; and Yuan, M. 2012. Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association*, 107(499): 1201–1216.

Crambes, C.; Kneip, A.; and Sarda, P. 2009. Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37(1): 35–72.

Dai, D.; Rigollet, P.; and Zhang, T. 2012. Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3): 1878–1905.

Dai, X.; Müller, H.-G.; and Yao, F. 2017. Optimal Bayes classifiers for functional data and density ratios. *Biometrika*, 104(3): 545–560.

Gaïffas, S.; and Lecué, G. 2011. Hyper-sparse optimal aggregation. *The Journal of Machine Learning Research*, 12: 1813–1833.

Hajiramezanali, E.; Zamani Dadaneh, S.; Karbalayghareh, A.; Zhou, M.; and Qian, X. 2018. Bayesian multi-domain learning for cancer subtype discovery from next-generation sequencing count data. *Advances in Neural Information Processing Systems*, 31.

Hall, P.; and Horowitz, J. L. 2007. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1): 70–91.

Hsing, T.; and Eubank, R. 2015. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. New York: Wiley.

Jin, J.; Yan, J.; Aseltine, R. H.; and Chen, K. 2024. Transfer learning with large-scale quantile regression. *Technometrics*, 1–13.

Kouw, W. M.; and Loog, M. 2018. An introduction to domain adaptation and transfer learning. *arXiv preprint arXiv:1812.11806*.

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: Statistical Methodology*, 84(1): 149–173.

Li, S.; Cai, T. T.; and Li, H. 2023. Transfer learning in large-scale gaussian graphical models with false discovery rate control. *Journal of the American Statistical Association*, 118(543): 2171–2183.

Li, S.; Zhang, L.; Cai, T. T.; and Li, H. 2024. Estimation and inference for high-dimensional generalized linear models with knowledge transfer. *Journal of the American Statistical Association*, 119(546): 1274–1285.

Li, Y.; and Hsing, T. 2007. On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98(9): 1782–1804.

Lin, H.; and Reimherr, M. 2024. On Hypothesis Transfer Learning of Functional Linear Models. In *International conference on machine learning*.

Lin, Y.; and Lin, Z. 2024+. Hypothesis Testing for Functional Linear Models via Bootstrapping. *Bernoulli*, in press.

Maity, S.; Dutta, D.; Terhorst, J.; Sun, Y.; and Banerjee, M. 2024. A linear adjustment-based approach to posterior drift in transfer learning. *Biometrika*, 111(1): 31–50.

Negahban, S. N.; Ravikumar, P.; Wainwright, M. J.; and Yu, B. 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4): 538–557.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10): 1345–1359.

Pan, W.; and Yang, Q. 2013. Transfer learning in heterogeneous collaborative filtering domains. *Artificial Intelligence*, 197: 39–55.

Ramsay, J. O.; and Silverman, B. W. 2005. *Functional data analysis*. Springer.

Raskutti, G.; Wainwright, M. J.; and Yu, B. 2010. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research*, 11: 2241–2259.

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

Tian, Y.; and Feng, Y. 2023. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544): 2684–2697.

Tian, Y.; Weng, H.; and Feng, Y. 2022. Unsupervised multitask and transfer learning on gaussian mixture models. *arXiv preprint arXiv:2209.15224*.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Xue, G.-R.; Dai, W.; Yang, Q.; and Yu, Y. 2008. Topic-bridged PLSA for cross-domain text classification. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 627–634.

Yao, F.; Müller, H.-G.; and Wang, J.-L. 2005. Functional linear regression analysis for longitudinal data. *The Annals of Statistics*, 33(6): 2873–2903.

Yuan, M.; and Cai, T. T. 2010. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6): 3412–3444.

Zhang, Y.; and Zhu, Z. 2022. Transfer learning for high-dimensional quantile regression via convolution smoothing. *arXiv preprint arXiv:2212.00428*.

Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; and He, Q. 2020. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1): 43–76.