Multivariate Functional Regression via Nested Reduced-Rank Regularization

Xiaokang Liu[1], Shujie Ma[2], Kun Chen[3*]
[1]*Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania*
[2]*Department of Statistics, University of California, Riverside*
[3]*Department of Statistics, University of Connecticut*

July 8, 2021

Abstract

We propose a nested reduced-rank regression (NRRR) approach in fitting a regression model with multivariate functional responses and predictors to achieve tailored dimension reduction and facilitate model interpretation and visualization. Our approach is based on a two-level low-rank structure imposed on the functional regression surfaces. A global low-rank structure identifies a small set of latent principal functional responses and predictors that drives the underlying regression association. A local low-rank structure then controls the complexity and smoothness of the association between the principal functional responses and predictors. The functional problem boils down to an integrated matrix approximation task through basis expansion, where the blocks of an integrated low-rank matrix share some common row space and/or column space. This nested reduced-rank structure also finds potential applications in multivariate time series modeling and tensor regression. A blockwise coordinate descend algorithm is developed. We establish the consistency of NRRR and show through non-asymptotic analysis that it can achieve at least a comparable error rate to that of the reduced-rank regression. Simulation studies demonstrate the effectiveness of NRRR. We apply the proposed methods in an electricity demand problem to relate daily electricity consumption trajectories with daily temperatures. Supplementary materials are available online.

Keywords: Dimension reduction; Matrix approximation; Multi-scale learning.

*Corresponding author; kun.chen@uconn.edu

1 Introduction

Multivariate functional data, which are generated when multiple variables are observed over a certain continuum, have become increasingly prevalent, partly due to the rapid advances in record keeping, inspection, and monitoring technologies in various fields. An object might be captured by cameras/scanners at a sequence of different angles/positions. As measured by various physiological indicators, the progression of a disease may be monitored frequently over time. With the richness of such data, it is often of interest to study the association between some multivariate functional responses and predictors. For example, with half-hourly observations on temperature and electricity consumption of the city Adelaide, the interest is to explore the predictive association between the daily electricity profiles and the daily temperature profiles for each day in a week simultaneously. Such a predictive model can then be used to infer future weekly power demand curves based on temperature forecasts to facilitate power supply and peak load management.

The aforementioned problem can be cast under the framework of functional regression, which has attracted considerable research efforts. Cardot et al. (1999, 2003) considered regressing a scalar response variable on a functional predictor, and James (2002) generalized it to the generalized linear regression setting. Faraway (1997) and Chiou et al. (2003) derived methods to model univariate functional response with scalar predictors. For relating a functional response and a functional predictor, Yao et al. (2005) considered a model based on functional principal component analysis (FPCA). He et al. (2010) studied a model which connects functional regression to functional canonical correlation analysis (FCCA). Ebaid (2008) imposed a low-rank structure on the coefficient surface and showed that low-rank regularization is closely connected to FPCA and FCCA. Extensions to the cases of multiple scalar or functional responses/predictors have been studied by various authors, e.g., Matsui et al. (2008), Zhu et al. (2017), and Krzyśko and Smaga (2017). Recently, He et al. (2018) proposed a multivariate varying-coefficient model to study the changing effects of predictors on responses, in which FPCA is used to reduce the number of unknown coefficient functions. As for the most general situation where both the response and the predictor are multivariate and functional, Ebaid (2008) considered imposing a low-rank structure on the coefficient

surface with basis expansion. Chiou et al. (2016) incorporated into their model the possible relationship between components of responses and predictors, respectively, by conducting multivariate FPCA to two sets of variables as the first step. For a comprehensive account of functional regression, see, e.g., Morris (2015) and Wang et al. (2016).

We consider the general scenario where both the response and the predictor are multivariate and functional. To formulate, let $\mathbf{y}(t) = [y_1(t), \ldots, y_d(t)]^{\mathrm{T}}$ be a d-dimensional vector of zero-mean functional response with $t \in \mathcal{T}$ and $\mathbf{x}(s) = [x_1(s), \ldots, x_p(s)]^{\mathrm{T}}$ be a p-dimensional vector of zero-mean functional predictor with $s \in \mathcal{S}$. We consider the multivariate functional linear regression model

$$\mathbf{y}(t) = \int_{\mathcal{S}} \mathbf{C}_0(s,t)\mathbf{x}(s)ds + \boldsymbol{\epsilon}(t), \qquad t \in \mathcal{T}, \tag{1}$$

where $\mathbf{C}_0(s,t) = [c_{k,l}(s,t)]_{d\times p}$ consists of unknown bivariate functions $c_{k,l}(s,t)$ assumed to be square integrable, i.e., $\int_{\mathcal{T}}\int_{\mathcal{S}} c_{k,l}^2(s,t)dsdt < \infty$, $k = 1, \ldots, d$, $l = 1, \ldots, p$, and $\boldsymbol{\epsilon}(t)$ is a d-dimensional zero-mean random error function. This formulation is a natural extension of the classical functional linear model (FLM) developed for univariate time-dependent responses. The key is to jointly estimate the many functional surfaces in Model (1) by utilizing the potential associations among the functional variables.

In this paper, our focus is on exploring the potentials of the *reduced-rank methodology* for fitting Model (1) with finite samples. In classical multivariate regression, low-rank models have been commonly applied to induce information sharing among the correlated responses and predictors in order to boost predictive performance and enhance model interpretation (Reinsel and Velu, 1998; Bunea et al., 2011; Chen et al., 2013). It appears straightforward to utilize this idea for functional regression, once a pragmatic basis expansion/truncation procedure (Ramsay and Silverman, 2005) is applied to transform the functional problem to finite dimensions. Imposing a low-rank structure on the resulting coefficient matrix is then a natural and somewhat generic choice for controlling model complexity (Ebaid, 2008). However, we argue that such a naive reduced-rank implementation does not take full advantage of the problem's multivariate and functional nature, and hence it can be inadequate in practice.

We innovate a *nested reduced-rank matrix representation*, to enable multi-scale learning in Model (1). At the global level, our method identifies latent principal functional factors that drive the functional association between the responses and the predictors. As such, dimension reduction is achieved when the number of latent responses is less than d and/or the number of latent predictors is less than p. This reduction can be quite effective in the presence of high-dimensional and highly-correlated functional variables. At the local level, the smaller-dimensional latent regression surface is assumed to be smooth and correspondingly its coefficient matrix derived through basis expansion is assumed to be of low rank, enabling another chance of dimension reduction. With these structures, the problem then boils down to a high-dimensional matrix decomposition and approximation task, where the nested reduced-rank structure implies that the blocks or submatrices of an integrated high-dimensional low-rank matrix share some common row space and/or column space. The applicability of the nested reduced-rank structure goes well beyond the functional setup; it also arises in vector autoregressive modeling of time series and tensor regression.

The paper is organized as follows. Section 2 introduces the nested reduced-rank formulation under Model (1), derives the model estimation procedure, and showcases the applicability of such nested reduced-rank matrix recovery in time series modeling, image compression, and tensor regression. Computational algorithms and rank selection methods are proposed, with details given in Supplementary Material A. In Section 3, we show the consistency of the proposed estimator and derive a non-asymptotic error bound. Simulation studies and the application on electricity demand are presented in Sections 4 and 5, respectively. In Section 6, we conclude with some remarks.

2 Nested Reduced-Rank Regression

2.1 Model Formulation

We propose a nested reduced-rank structure under Model (1), to appreciate both the multivariate and the functional natures of the problem.

Structure 1. *(Global reduced-rank structure)*

$$\mathbf{C}_0(s,t) = \mathbf{U}_0\mathbf{C}_0^*(s,t)\mathbf{V}_0^{\mathrm{T}}, \qquad s \in \mathcal{S}, t \in \mathcal{T},$$

where $\mathbf{U}_0 \in \mathbb{R}^{d\times r_y}$ *with* $r_y \le d$, $\mathbf{V}_0 \in \mathbb{R}^{p\times r_x}$ *with* $r_x \le p$, *and* $\mathbf{C}_0^*(s,t)$ *is an* $r_y \times r_x$ *latent regression surface. Without loss of generality, we assume* $\mathbf{U}_0^{\mathrm{T}}\mathbf{U}_0 = \mathbf{I}_{r_y}$ *and* $\mathbf{V}_0^{\mathrm{T}}\mathbf{V}_0 = \mathbf{I}_{r_x}$.

In Structure 1, $\mathbf{U}_0$ and $\mathbf{V}_0$ are designed to capture the "global" effects of the functional association, i.e., it implies that the association between $\mathbf{y}(t)$ and $\mathbf{x}(t)$ is driving by some lower-dimensional latent functional responses and latent predictors that are formed as some linear combinations of the original functional responses and predictors, respectively. That is, it implies that $\mathbf{y}^*(t) = \int_{\mathcal{S}} \mathbf{C}_0^*(s,t)\mathbf{x}^*(s)ds + \boldsymbol{\epsilon}^*(t)$, where $\mathbf{y}^*(t) = \mathbf{U}_0^{\mathrm{T}}\mathbf{y}(t)$, $\mathbf{x}^*(s) = \mathbf{V}_0^{\mathrm{T}}\mathbf{x}(s)$ and $\boldsymbol{\epsilon}^*(t) = \mathbf{U}_0^{\mathrm{T}}\boldsymbol{\epsilon}(t)$. When $r_y < d$ and/or $r_x < p$, our model achieves great dimensionality reduction and parsimony while retaining flexibility. The proposed structure is particularly helpful for simultaneously modeling a large number of functional responses and predictors that are highly correlated across s or t.

It is conventional to take a basis expansion and truncation approach to facilitate the modeling of the latent regression surface $\mathbf{C}_0^*(s,t) \in \mathbb{R}^{r_y\times r_x}$ (Ramsay and Silverman, 2005), for inducing its smoothness over s and t and converting the infinite-dimensional problem to be finite-dimensional. Specifically, we represent the latent regression surface $\mathbf{C}_0^*(s,t)$ as

$$\mathbf{C}_0^*(s,t) \approx (\mathbf{I}_{r_y} \otimes \boldsymbol{\Psi}^{\mathrm{T}}(t))\mathbf{C}_0^*(\mathbf{I}_{r_x} \otimes \boldsymbol{\Phi}(s)), \qquad \mathbf{C}_0^* \in \mathbb{R}^{(J_y r_y)\times(J_x r_x)}, \tag{2}$$

where $\mathbf{I}_a$ denotes the $a \times a$ identity matrix, $\boldsymbol{\Phi}(s) = [\phi_1(s), \ldots, \phi_{J_x}(s)]^{\mathrm{T}}$ consists of J_x basis functions with $\mathbf{J}_{\phi\phi} = \int_{\mathcal{S}} \boldsymbol{\Phi}(s)\boldsymbol{\Phi}^{\mathrm{T}}(s)ds$ being positive definite (p.d.), and similarly, $\boldsymbol{\Psi}(t) = [\psi_1(t), \ldots, \psi_{J_y}(t)]^{\mathrm{T}}$ consists of J_y basis functions with $\mathbf{J}_{\psi\psi} = \int_{\mathcal{T}} \boldsymbol{\Psi}(t)\boldsymbol{\Psi}^{\mathrm{T}}(t)dt$ being p.d. The matrix $\mathbf{C}_0^*$ on the right-hand side of (2) then collects all the coefficients to expand the $r_y \times r_x$ many bivariate functions in $\mathbf{C}_0^*(s,t)$ with the basis $\boldsymbol{\Phi}(s)$ and $\boldsymbol{\Psi}(t)$. Here we assume the basis is given, such as spline, wavelet, and Fourier basis, and is with a sufficiently large number of components.

With the expansion in (2), it boils down to consider the modeling of the high-dimensional

coefficient matrix $\mathbf{C}_0^*$. We further explore a potential low-rank structure in $\mathbf{C}_0^*$.

Structure 2. *(Local reduced-rank structure)*

$$rank(\mathbf{C}_0^*) \leq r,$$

for $r \leq \min(J_y r_y, J_x r_x)$*; that is,* $\mathbf{C}_0^* = \mathbf{A}_0^* \mathbf{B}_0^{*\mathrm{T}}$ *for some* $\mathbf{A}_0^* \in \mathbb{R}^{(J_y r_y) \times r}$, $\mathbf{B}_0^* \in \mathbb{R}^{(J_x r_x) \times r}$.

Since the above structure induces the dependency between the latent responses and the latent predictors through their basis-expanded representations, we achieve a finer dimension reduction at the "local" level.

The approximation error in (2) can be controlled under reasonable conditions. Assume that the $\lfloor \gamma \rfloor$th order derivative of each function in $\mathbf{C}_0^*(s,t)$ satisfies the Hölder condition of order $\gamma - \lfloor \gamma \rfloor$ with $\gamma > 1/2$, where $\lfloor \gamma \rfloor$ is the biggest integer strictly smaller than γ. This smoothness condition together with Structures 1–2 imply that the regression surface $\mathbf{C}_0(s,t)$ approximately admits a *nested reduced-rank representation*,

$$\sup_{s \in \mathcal{S}, t \in \mathcal{T}} |\mathbf{C}_0(s,t) - \mathbf{U}_0(\mathbf{I}_{r_y} \otimes \mathbf{\Psi}^{\mathrm{T}}(t)) \mathbf{A}_0^* \mathbf{B}_0^{*\mathrm{T}} (\mathbf{I}_{r_x} \otimes \mathbf{\Phi}(s)) \mathbf{V}_0^{\mathrm{T}}| = O(J_y^{-\gamma} + J_x^{-\gamma}). \tag{3}$$

We can choose the number of basis functions satisfying $J_y \to \infty$ and $J_x \to \infty$ as $n \to \infty$, so that the above approximation error vanishes. Indeed, this is allowed in our non-asymptotic analysis that provides a high-probability prediction error bound; see Section 3 for details.

Model (1) then becomes

$$\begin{aligned} \mathbf{y}(t) &\approx \int_{\mathcal{S}} \mathbf{U}_0(\mathbf{I}_{r_y} \otimes \mathbf{\Psi}^{\mathrm{T}}(t)) \mathbf{A}_0^* \mathbf{B}_0^{*\mathrm{T}} (\mathbf{I}_{r_x} \otimes \mathbf{\Phi}(s)) \mathbf{V}_0^{\mathrm{T}} \mathbf{x}(s) ds + \boldsymbol{\epsilon}(t) \\ &\approx (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))(\mathbf{U}_0 \otimes \mathbf{I}_{J_y}) \mathbf{A}_0^* \mathbf{B}_0^{*\mathrm{T}} (\mathbf{V}_0^{\mathrm{T}} \otimes \mathbf{I}_{J_x}) \left\{ \int_{\mathcal{S}} (\mathbf{I}_p \otimes \mathbf{\Phi}(s)) \mathbf{x}(s) ds \right\} + \boldsymbol{\epsilon}(t). \end{aligned} \tag{4}$$

We remark that $\mathbf{U}_0$, $\mathbf{V}_0$, $\mathbf{A}_0^*$ and $\mathbf{B}_0^*$ are not fully identifiable individually up to rotation or nonsingular transformation, similar to the settings in conventional reduced-rank estimation; nevertheless, the structure as a whole is well-defined and identifiable.

It is worthwhile to mention a few special cases. When the low-dimensional structures do not present at all, i.e., $r_x = p$, $r_y = d$ and $r = \min(J_x r_x, J_y r_y)$, the model becomes

$\mathbf{C}_0(s,t) = (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))\mathbf{C}_0^*(\mathbf{I}_p \otimes \mathbf{\Phi}(s))$, for which the least squares estimation is equivalent to separately regressing each response $y_k(t)$ on $\mathbf{x}(s)$ and hence there is no gain of conducting a multivariate analysis. When the global structure does not present, i.e., $r_x = p$ and $r_y = d$, the model reduces to a reduced-rank functional model as in Ebaid (2008).

2.2 Estimation

The model estimation at the population level can be conducted through minimizing the mean integrated squared error (MISE) with respect to $\mathbf{C}(s,t)$,

$$\mathbb{E}\int_{\mathcal{T}} \left\| \mathbf{y}(t) - \int_{\mathcal{S}} \mathbf{C}(s,t)\mathbf{x}(s)ds \right\|^2 dt, \tag{5}$$

where $\|\mathbf{a}\| = \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{a}}$ denotes the ℓ_2 norm. Define the integrated predictor and response as

$$\mathbf{x} = \int_{\mathcal{S}} (\mathbf{I}_p \otimes \mathbf{\Phi}(s))\mathbf{x}(s)ds, \quad \mathbf{y} = (\mathbf{I}_d \otimes \mathbf{J}_{\psi\psi}^{-\frac{1}{2}}) \int_{\mathcal{T}} (\mathbf{I}_d \otimes \mathbf{\Psi}(t))\mathbf{y}(t)dt, \tag{6}$$

and write $\mathbf{y}(t) = (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))(\mathbf{I}_d \otimes \mathbf{J}_{\psi\psi}^{-\frac{1}{2}})\mathbf{y} + (\mathbf{I}_d \otimes \mathbf{\Psi}_\perp^{\mathrm{T}}(t))(\mathbf{I}_d \otimes \mathbf{J}_{\psi_\perp\psi_\perp}^{-\frac{1}{2}})\mathbf{y}_\perp$, where $\mathbf{x} \in \mathbb{R}^{J_x p}$, $\mathbf{y} \in \mathbb{R}^{J_y d}$, $\mathbf{y}_\perp \in \mathbb{R}^{J_y d}$, and $\int \mathbf{\Psi}(t)\mathbf{\Psi}_\perp^{\mathrm{T}}(t)dt = 0$. Under the nested reduced-rank model in (4), the MISE in (5) becomes

$$\begin{aligned}
&\mathbb{E}\int_{\mathcal{T}} \left\| \mathbf{y}(t) - (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))(\mathbf{U} \otimes \mathbf{I}_{J_y})\mathbf{A}^*\mathbf{B}^{*\mathrm{T}}(\mathbf{V}^{\mathrm{T}} \otimes \mathbf{I}_{J_x})\mathbf{x} \right\|^2 dt \\
=&\mathbb{E}\int_{\mathcal{T}} \left\| (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))(\mathbf{I}_d \otimes \mathbf{J}_{\psi\psi}^{-\frac{1}{2}})\mathbf{y} - (\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t))(\mathbf{U} \otimes \mathbf{I}_{J_y})\mathbf{A}^*\mathbf{B}^{*\mathrm{T}}(\mathbf{V}^{\mathrm{T}} \otimes \mathbf{I}_{J_x})\mathbf{x} \right\|^2 dt + \text{const},
\end{aligned}$$

where "const" represents the constant term that is free of the model parameters. As a result, the estimation criterion becomes

$$\min_{\mathbf{U},\mathbf{V},\mathbf{A}^*,\mathbf{B}^*} \left\{ \mathbb{E} \left\| \mathbf{y} - (\mathbf{I}_d \otimes \mathbf{J}_{\psi\psi}^{\frac{1}{2}})(\mathbf{U} \otimes \mathbf{I}_{J_y})\mathbf{A}^*\mathbf{B}^{*\mathrm{T}}(\mathbf{V}^{\mathrm{T}} \otimes \mathbf{I}_{J_x})\mathbf{x} \right\|^2 \right\}. \tag{7}$$

This is a generalization of the reduced-rank regression criterion (Reinsel and Velu, 1998). Unlike the latter, however, (7) does not lead to an explicit analytic expression in general.

We now consider the corresponding sample estimation problem. Suppose the functional

responses and predictors are fully observed over their respective domains for n random subjects, i.e., $(\mathbf{y}_i(t), \mathbf{x}_i(s))$ for $t \in \mathcal{T}$, $s \in \mathcal{S}$, and $i = 1, \ldots, n$. The integrated predictors and responses for each subject i can then be computed according to (6),

$$\begin{aligned} x_{ilj} &= \int_{\mathcal{S}} \phi_j(s) x_{li}(s) ds, \qquad l = 1, \ldots, p; j = 1, \ldots, J_x, \\ y^0_{ikj} &= \int_{\mathcal{T}} \psi_j(t) y_{ki}(t) dt, \qquad k = 1, \ldots, d; j = 1, \ldots, J_y, \\ (y_{ik1}, \ldots, y_{ikJ_y})^{\mathrm{T}} &= \mathbf{J}_{\psi\psi}^{-\frac{1}{2}} (y^0_{ik1}, \ldots, y^0_{ikJ_y})^{\mathrm{T}}, \qquad k = 1, \ldots, d. \end{aligned} \tag{8}$$

In practice with discretely observed data, the integrals in (8) can be approximated by finite Riemann sums (Ramsay and Silverman, 2005); see Supplementary Material A.

Define $\mathbf{Y}_{\cdot j} = (y_{ikj})_{n \times d}$, for $j = 1, \ldots, J_y$, and let $\mathbf{Y} = (\mathbf{Y}_{\cdot 1}, \ldots, \mathbf{Y}_{\cdot J_y}) \in \mathbb{R}^{n \times (J_y d)}$. Similarly, define $\mathbf{X}_{\cdot j} = (x_{ilj})_{n \times p}$, and let $\mathbf{X} = (\mathbf{X}_{\cdot 1}, \ldots, \mathbf{X}_{\cdot J_x}) \in \mathbb{R}^{n \times (J_x p)}$. We write $\mathbf{A}^* = (\mathbf{A}_{1\cdot}^{\mathrm{T}}, \ldots, \mathbf{A}_{r_y\cdot}^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{A}_{h\cdot} \in \mathbb{R}^{J_y \times r}$ for $h = 1, \ldots, r_y$, and $\mathbf{B}^* = (\mathbf{B}_{1\cdot}^{\mathrm{T}}, \ldots, \mathbf{B}_{r_x\cdot}^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{B}_{h\cdot} \in \mathbb{R}^{J_x \times r}$ for $h = 1, \ldots, r_x$. Define $\widetilde{\mathbf{A}}_{h\cdot} = \mathbf{J}_{\psi\psi}^{\frac{1}{2}} \mathbf{A}_{h\cdot}$ and $\widetilde{\mathbf{A}}^* = (\widetilde{\mathbf{A}}_{1\cdot}^{\mathrm{T}}, \ldots, \widetilde{\mathbf{A}}_{r_y\cdot}^{\mathrm{T}})^{\mathrm{T}}$. Since $\mathbf{J}_{\psi\psi}$ is nonsingular, it suffices to consider the estimation of $\widetilde{\mathbf{A}}^*$ instead of $\mathbf{A}^*$. It is necessary to rearrange the rows of $\widetilde{\mathbf{A}}^*$ and $\mathbf{B}^*$, i.e., let $\mathbf{A} = (\mathbf{A}_{\cdot 1}^{\mathrm{T}}, \ldots, \mathbf{A}_{\cdot J_y}^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{A}_{\cdot j} \in \mathbb{R}^{r_y \times r}$ is formed by collecting the jth row of each $\widetilde{\mathbf{A}}_{h\cdot}$, and $\mathbf{B} = (\mathbf{B}_{\cdot 1}^{\mathrm{T}}, \ldots, \mathbf{B}_{\cdot J_x}^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{B}_{\cdot j} \in \mathbb{R}^{r_x \times r}$ is formed by collecting the jth row of each $\mathbf{B}_{h\cdot}$. Finally, these matrix notations allow us to write the sample MISE criterion as a nested reduced-rank regression (NRRR) problem,

$$\min_{\mathbf{C}} \|\mathbf{Y} - \mathbf{X}\mathbf{C}\|_{\mathrm{F}}^2, \qquad s.t.\, \mathbf{C} = (\mathbf{I}_{J_x} \otimes \mathbf{V}) \mathbf{B} \mathbf{A}^{\mathrm{T}} (\mathbf{I}_{J_y} \otimes \mathbf{U}^{\mathrm{T}}), \tag{9}$$

where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. Figure 1 shows a conceptual diagram of the nested reduced-rank structure in $\mathbf{C}$. The $\mathbf{U}$ and $\mathbf{V}$ can be regarded as two loading matrices for the responses and the predictors, respectively; see also the discussion after Structure 1. From matrix approximation point of view, they respectively capture the shared column and row spaces among the blockwise sub-matrices of $\mathbf{C}$. The score matrix $\mathbf{B}\mathbf{A}^{\mathrm{T}}$ is assembled from all the corresponding blockwise score matrices, and as a whole, it is of low rank.

The optimization problem in (9) is non-convex and has no closed-form solution, and

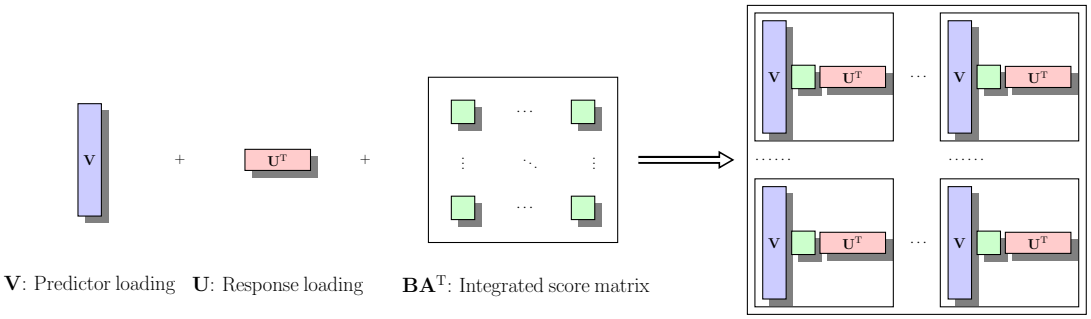


Figure 1: A diagram of the nested reduced-rank matrix structure in (9). The coefficient matrix $\mathbf{C}$ admits a block structure and is parameterized by two loading matrices $\mathbf{U}$ and $\mathbf{V}$, and an integrated score matrix $\mathbf{B}\mathbf{A}^{\mathrm{T}}$. The $\mathbf{U}$ and $\mathbf{V}$ respectively capture the shared column and row spaces among the blockwise sub-matrices of $\mathbf{C}$; the $\mathbf{B}\mathbf{A}^{\mathrm{T}}$ is assembled from all the corresponding blockwise score matrices, and as a whole it is of low rank.

thus we develop a blockwise coordinate descent algorithm to solve it. In the proposed algorithm, with fixed triplets of rank values (r, r_x, r_y), we alternatingly update each component $\mathbf{A}$, $\mathbf{B}$, $\mathbf{U}$ and $\mathbf{V}$ in the nested reduced-rank representation while keeping others fixed. The resulting sub-optimization problems are with explicit solutions. The objective function in (9) is monotone decreasing along the iterations, and consequently the convergence to a limiting point is guaranteed. We provide an initialization method to help achieve convergence in an efficient and stable manner; simulation studies demonstrate the robustness of the algorithm with respect to random perturbations of the initial values. The details of the algorithm, the initialization, and the robustness checks are in Supplementary Material A.

To choose an optimal set of rank values (r, r_x, r_y), the K-fold cross validation procedure can be used, which, however, can be quite computationally expensive. We propose a Bayesian Information Criterion (BIC) (Schwarz, 1978). Denote $\widehat{\mathbf{C}}(r, r_x, r_y)$ as the estimator of $\mathbf{C}$ by solving (9) with the rank values fixed at some (r, r_x, r_y) and write the sum of squared errors as $\mathrm{SSE}(r, r_x, r_y) = \|\mathbf{Y} - \mathbf{X}\widehat{\mathbf{C}}(r, r_x, r_y)\|_{\mathrm{F}}^2$. We define $\mathrm{BIC}(r, r_x, r_y) = ndJ_y \log\{\mathrm{SSE}(r, r_x, r_y)/(ndJ_y)\} + \log(ndJ_y)df(r, r_x, r_y)$, where $df(r, r_x, r_y)$ is the effective degrees of freedom of the model and is estimated by the number of free model parameters

$$\widehat{df}(r, r_x, r_y) = r_x\{r(\mathbf{X})/J_x - r_x\} + r_y(d - r_y) + (J_y r_y + J_x r_x - r)r. \tag{10}$$

When $r_y = d$, $r_x = r(\mathbf{X})/J_x$, the above formula gives $\widehat{df}(r, r(\mathbf{X})/J_x, d) = (J_y r_y + r(\mathbf{X}) - r)r$,

which is exactly the effective number of parameters in a rank-r reduced-rank regression model (Mukherjee et al., 2015). The difference in the number of parameters is $(J_y d - J_y r_y)(r - r_y/J_y) + (r(\mathbf{X}) - J_x r_x)(r - r_x/J_x)$.

2.3 Other Applications and Connection with Tensor Regression

The applicability of the nested reduced-rank estimation is beyond the functional setup. An interesting application is in high-dimensional vector autoregressive (VAR) modeling in multivariate time series analysis. Let $\mathbf{y}_t \in \mathbb{R}^p$ be the observed multivariate time series at time t. Consider a VAR model of order h,

$$\mathbf{y}_t = \mathbf{A}_1 \mathbf{y}_{t-1} + \ldots + \mathbf{A}_h \mathbf{y}_{t-h} + \mathbf{e}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{e}_t, \qquad t = 1, \ldots, T,$$

where $\mathbf{A}_i \in \mathbb{R}^{p \times p}$, $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_h) \in \mathbb{R}^{p \times hp}$, $\mathbf{x}_{t-1} = (\mathbf{y}_{t-1}^{\mathrm{T}}, \ldots, \mathbf{y}_{t-h}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{hp}$, and $\mathbf{e}_t \in \mathbb{R}^p$ is a zero-mean innovative process. Stationary reduced-rank VAR model was introduced in Luetkepohl (1993), where the coefficient matrix $\mathbf{A}$ is assumed to be of low rank. In high-dimensional scenarios, it is possible that (1) some linear combinations of the multivariate time series $\mathbf{y}_t$ are processes of pure noise, and (2) the dynamics of $\mathbf{y}_t$ is driven by its lags only through some linear combinations. This fact gives rise to a nested reduced-rank structure. Specifically, the global structure can be modeled as $\mathbf{A}_i = \mathbf{U}_0 \mathbf{A}_i^* \mathbf{V}_0^{\mathrm{T}}$, $i = 1, \ldots, h$, where $\mathbf{U}_0 \in \mathbb{R}^{p \times r_1}$ with $r_1 \leq p$, $\mathbf{V}_0 \in \mathbb{R}^{p \times r_2}$ with $r_2 \leq p$, satisfying $\mathbf{U}_0^{\mathrm{T}} \mathbf{U}_0 = \mathbf{I}_{r_1}$ and $\mathbf{V}_0^{\mathrm{T}} \mathbf{V}_0 = \mathbf{I}_{r_2}$. The local low-dimensional structure can be modeled by letting the matrix $(\mathbf{A}_1^*, \ldots, \mathbf{A}_h^*) \in \mathbb{R}^{r_1 \times (hr_2)}$ be of low rank. As such, $\mathbf{V}_0^{\mathrm{T}} \mathbf{y}_t$ gives the latent principal time series, and $\mathbf{U}_0^{\perp\mathrm{T}} \mathbf{y}_t$ are pure noise where $\mathbf{U}_0^{\perp} \in \mathbb{R}^{p \times (p - r_1)}$ and $\mathbf{U}_0^{\mathrm{T}} \mathbf{U}_0^{\perp} = \mathbf{0}$.

Another application is in surveillance video processing. In recent years, the sparse plus low-rank decomposition has been a popular method for surveillance video decoding, in which the low-rank component represents the background and the sparse component captures the moving objects. Since the surveillance video frames are usually with a static or gradually changed background, using a nested reduced-rank component with an extra global reduction scheme may improve the efficiency of background representation by dramatically reducing the temporal redundancy. These ideas will be further explored in our future work.

The proposed nested reduced-rank regression can also be regarded as a tensor-on-tensor regression (Kolda and Bader, 2009; Lock, 2018) with a delicately structured coefficient tensor. Specifically, the response matrix $\mathbf{Y} \in \mathbb{R}^{n\times(J_y d)}$ in (9) can be rearranged along three directions, i.e., subject, response component, and functional basis, into a 3rd-order response tensor $\mathcal{Y} \in \mathbb{R}^{n\times d\times J_y}$. Similarly, $\mathbf{X} \in \mathbb{R}^{n\times(J_x p)}$ can be rearranged into a 3rd-order predictor tensor $\mathcal{X} \in \mathbb{R}^{n\times p\times J_x}$. Consequently, the estimation criterion in (9) can be rewritten as $\|\mathcal{Y} - \langle \mathcal{X}, \mathcal{C}\rangle\|_{\mathrm{F}}^2$, where $\mathcal{C} \in \mathbb{R}^{p\times J_x\times d\times J_y}$ is a 4th-order coefficient tensor as a rearrangement of the coefficient matrix $\mathbf{C}$, and $\langle\cdot,\cdot\rangle$ is the contracted tensor product. Interestingly, the nested reduced-rank structure in $\mathbf{C}$ implies a novel tensor factorization structure in $\mathcal{C}$. The global reduced-rank structure implies a rank-restricted Tucker decomposition, i.e., $\mathcal{C} = [\mathcal{G}; \mathbf{V}, \mathbf{I}_{J_x}, \mathbf{U}, \mathbf{I}_{J_y}] = \mathcal{G} \times_1 \mathbf{V} \times_2 \mathbf{I}_{J_x} \times_3 \mathbf{U} \times_4 \mathbf{I}_{J_y}$, where $\times_i$ is the i-mode product, $\mathcal{G} \in \mathbb{R}^{r_x\times J_x\times r_y\times J_y}$ is the core tensor, and $\mathbf{U}$ and $\mathbf{V}$ are the loading matrices as in (9). The local reduced-rank structure is equivalent to impose a low-rank structure on a flattened matrix $\mathbf{G}$ of the core tensor $\mathcal{G}$, i.e., $\mathbf{G}$ is formed by arranging the slices $\mathbf{G}_{l,k} \in \mathbb{R}^{r_x\times r_y}$, $l = 1,\ldots,J_x$; $k = 1,\ldots,J_y$ along (l,k) as a blockwise matrix $\mathbf{G} = (\mathbf{G}_{l,k}) \in \mathbb{R}^{J_x r_x\times J_y r_y}$. To the best of our knowledge, existing studies mainly consider sparsity of the core tensor (Li et al., 2018), while we show that certain low-rank structure of the core tensor can be effective and interpretable. Our work thus offers a new venue for dimension reduction in tensor models.

3 Theoretical Analysis

Our theoretical analysis concerns the fundamental NRRR setup,

$$\mathbf{Y} = \mathbf{X}\mathbf{C}_0 + \mathbf{E}, \qquad s.t.\ \mathbf{C}_0 = (\mathbf{I}_{J_x} \otimes \mathbf{V}_0)\mathbf{B}_0\mathbf{A}_0^{\mathrm{T}}(\mathbf{I}_{J_y} \otimes \mathbf{U}_0^{\mathrm{T}}). \tag{11}$$

Accordingly, the objective function is $\mathbf{Q}_n(\mathbf{V},\mathbf{B},\mathbf{A},\mathbf{U}) = \|\mathbf{Y} - \mathbf{X}(\mathbf{I}_{J_x}\otimes\mathbf{V})\mathbf{B}\mathbf{A}^{\mathrm{T}}(\mathbf{I}_{J_y}\otimes\mathbf{U}^{\mathrm{T}})\|_{\mathrm{F}}^2$, and the NRRR estimator is obtained as $(\widehat{\mathbf{V}},\widehat{\mathbf{B}},\widehat{\mathbf{A}},\widehat{\mathbf{U}}) \in \arg\min_{\mathbf{V},\mathbf{B},\mathbf{A},\mathbf{U}} \mathbf{Q}_n(\mathbf{V},\mathbf{B},\mathbf{A},\mathbf{U})$. To facilitate the analysis, it is necessary to make the components $(\mathbf{V}_0,\mathbf{B}_0,\mathbf{A}_0,\mathbf{U}_0)$ identifiable individually; we defer the discussion until presenting the main results. Here, the integrated response and predictor matrices from functional data are treated as given, as the problem's

functional approximation aspect is not our focus. We have assumed that the rank values are known. Even so, the non-convexity of the NRRR problem, induced by the nested low-rank matrix decomposition, makes the theoretical analysis challenging. We need the following conditions on the model in (11) for our asymptotic analysis.

Assumption 1. $\mathbf{X}^{\mathrm{T}}\mathbf{X}/n \xrightarrow{a.s.} \mathbf{\Gamma}$ *as* $n \to \infty$, *where* $\mathbf{\Gamma}$ *is a fixed, positive-definite matrix.*

Assumption 2. *Each row* $\mathbf{e}_i$ *of* $\mathbf{E}$ *is independently and identically distributed with* $\mathbb{E}(\mathbf{e}_i) = \mathbf{0}$ *and* $cov(\mathbf{e}_i) = \mathbf{\Sigma}$, *where* $\mathbf{\Sigma}$ *is positive-definite.*

Theorem 1. *(Consistency) Suppose Assumptions 1 and 2 hold. Then there exists a local minimizer* $(\widehat{\mathbf{V}}, \widehat{\mathbf{B}}, \widehat{\mathbf{A}}, \widehat{\mathbf{U}})$ *of* $\mathbf{Q}_n(\mathbf{V}, \mathbf{B}, \mathbf{A}, \mathbf{U})$ *such that* $\|\widehat{\mathbf{V}} - \mathbf{V}_0\|_{\mathrm{F}} = O_p(n^{-\frac{1}{2}})$, $\|\widehat{\mathbf{B}} - \mathbf{B}_0\|_{\mathrm{F}} = O_p(n^{-\frac{1}{2}})$, $\|\widehat{\mathbf{A}} - \mathbf{A}_0\|_{\mathrm{F}} = O_p(n^{-\frac{1}{2}})$ *and* $\|\widehat{\mathbf{U}} - \mathbf{U}_0\|_{\mathrm{F}} = O_p(n^{-\frac{1}{2}})$.

Theorem 1 shows the consistency of the NRRR estimation in estimating the components of the nested low-rank structure, in the sense that there exists a local minimizer that is $\sqrt{n}-$consistent. For non-convex problems, such an asymptotic result is what to be expected (Chen et al., 2012). While the details of the proof are provided in Supplementary Material B.2, we briefly outline the main steps here. We first parameterize the coefficient matrix $\mathbf{C}_0$ such that the components in its nested low-rank structure, $(\mathbf{V}_0, \mathbf{B}_0, \mathbf{A}_0, \mathbf{U}_0)$, can be identifiable. Then a local neighborhood around the true value $\mathbf{C}_0$ with radius h is constructed, denoted as $\mathcal{N}(\mathbf{C}_0, h)$. We then show that for any given $\epsilon > 0$,

$$\mathbb{P}\Bigg\{ \inf_{\|\check{\mathbf{R}}^1\|_{\mathrm{F}}=\|\check{\mathbf{R}}^2\|_{\mathrm{F}}=\|\check{\mathbf{R}}^3\|_{\mathrm{F}}=\|\check{\mathbf{R}}^4\|_{\mathrm{F}}=h} \mathbf{Q}_n(\mathbf{V}_0 + \frac{1}{\sqrt{n}}\mathbf{R}^1, \mathbf{B}_0 + \frac{1}{\sqrt{n}}\mathbf{R}^2, \mathbf{A}_0 + \frac{1}{\sqrt{n}}\mathbf{R}^3, \mathbf{U}_0 + \frac{1}{\sqrt{n}}\mathbf{R}^4) \\ > \mathbf{Q}_n(\mathbf{V}_0, \mathbf{B}_0, \mathbf{A}_0, \mathbf{U}_0) \Bigg\} \geq 1 - \epsilon$$

with a large enough constant h. Here the infimum is taken over the perturbation matrices $\mathbf{R}^1, \mathbf{R}^2, \mathbf{R}^3, \mathbf{R}^4$ (one-to-one transformations of $\check{\mathbf{R}}^1, \check{\mathbf{R}}^2, \check{\mathbf{R}}^3, \check{\mathbf{R}}^4$) of $\mathbf{V}_0, \mathbf{B}_0, \mathbf{A}_0, \mathbf{U}_0$, respectively, with a fixed Frobenius norm h. That is, the objective function evaluated at any boundary point of the neighborhood of radius h is larger than that evaluated at the true value, with an arbitrarily large probability. It thus follows that a local minimizer must exist within the neighborhood with a $\sqrt{n}$ convergence rate.

We also attempt a non-asymptotic analysis to understand better the behavior of NRRR estimator in high-dimensional setups. Let's express the true functional regression surface as $\mathbf{C}_0(s,t) = \{\mathbf{I}_d \otimes \mathbf{\Psi}(t)^{\mathrm{T}}\}\{\mathbf{I}_d \otimes \mathbf{J}_{\psi\psi}^{-\frac{1}{2}}\}\widetilde{\mathbf{C}}_0^{\mathrm{T}}\{\mathbf{I}_p \otimes \mathbf{\Phi}(s)\}$, where $\widetilde{\mathbf{C}}_0$ is obtained by a rearrangement of the columns and rows of $\mathbf{C}_0$. Let $\widehat{\mathbf{C}} = (\mathbf{I}_{J_x} \otimes \widehat{\mathbf{V}})\widehat{\mathbf{B}}\widehat{\mathbf{A}}^{\mathrm{T}}(\mathbf{I}_{J_y} \otimes \widehat{\mathbf{U}}^{\mathrm{T}})$ be the NRRR estimator of $\mathbf{C}_0$, and $\widehat{\mathbf{C}}(s,t)$ is obtained by plugging in the corresponding components.

Theorem 2. *Suppose the random error matrix* $\mathbf{E}$ *has independent* $N(0,\sigma^2)$ *entries. With probability of at least* $1 - \exp\{-\theta^2(r(\mathbf{X}) + dJ_y)/2\}$*, we have*

$$\|\mathbf{X}\widehat{\mathbf{C}} - \mathbf{X}\mathbf{C}_0\|_{\mathrm{F}}^2 \lesssim (r(\mathbf{X}) + dJ_y)r,$$
$$\int_{\mathcal{T}}\int_{\mathcal{S}} \left\| \left(\widehat{\mathbf{C}}(s,t) - \mathbf{C}_0(s,t)\right)\mathbf{x}(s) \right\|^2 dsdt \lesssim (r(\mathbf{X}) + dJ_y)r,$$

where $\theta > 0$ *is a positive constant. Here* $\lesssim$ *means that the inequality holds up to some multiplicative numerical constants.*

Theorem 2 shows that the prediction error bounds of NRRR are at least comparable to those of reduced-rank regression (Bunea et al., 2011). The proof of Theorem 2 is in Supplementary Material B.3. This result provides support for using NRRR in problems with diverging dimensionality; indeed, we see from numerical studies that NRRR always outperforms RRR. We expect that the optimal rate for NRRR is faster than that given above, since the number of free parameters in a nested low-rank structure can be much smaller than that in a regular reduced-rank structure due to the global dimension reduction by $(\mathbf{V}_0, \mathbf{U}_0)$; see the formulation of the degrees of freedom in (10) and the discussion afterward. We will explore this conjecture in our future work.

4 Simulation

We compare the performance of the proposed NRRR methods with several competing methods, including the ordinary least squares method (OLS), the classical reduced-rank regression (RRR), and the reduced-rank ridge regression (RRS). For NRRR, besides the regular version, we consider a special case of setting $r_y = d$, denoted as NRRR-X, and the

nested reduced-rank ridge regression, denoted as NRRS, in which a ridge penalty is added to the NRRR criterion for inducing parameter shrinkage.

To generate synthetic data, we let $\mathbf{x}(s) = \{\mathbf{I}_p \otimes \mathbf{\Phi}^{\mathrm{T}}(s)\}\mathbf{x}$ and $\boldsymbol{\epsilon}(t) = \{\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t)\}\boldsymbol{\epsilon}$, where $\mathbf{x} \in \mathbb{R}^{J_x p}$, $\mathbf{y} \in \mathbb{R}^{J_y d}$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{J_y d}$ are random vectors, and $\mathbf{\Phi}(s)$ and $\mathbf{\Psi}(t)$ are the same two sets of B-spline basis functions used to expand $\mathbf{C}(s,t)$. The $\mathbf{y}(t)$ is then given according to (4), i.e., $\mathbf{y}(t) = \left\{\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t)\right\}\left\{(\mathbf{U}_0 \otimes \mathbf{I}_{J_y})\mathbf{A}_0^*\mathbf{B}_0^{*\mathrm{T}}(\mathbf{V}_0 \otimes \mathbf{I}_{J_x})^{\mathrm{T}}(\mathbf{I}_p \otimes \mathbf{J}_{\phi\phi})\mathbf{x} + \boldsymbol{\epsilon}\right\}$. Then, for each $i = 1, ..., n$, the discrete-time observations $(\mathbf{x}_i(s), \mathbf{y}_i(t))$ are generated as follows,

1. Generate $\mathbf{x}_i(s) = \left\{\mathbf{I}_p \otimes \mathbf{\Phi}^{\mathrm{T}}(s)\right\}\mathbf{x}_i$ for uniformly distributed points s_u, $u = 1, \ldots, g$ in $\mathcal{S} = [0,1]$, where $\mathbf{x}_i \in \mathbb{R}^{J_x p}$ is from $N(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma} = (\rho^{|i-j|})$ for some $0 < \rho < 1$.

2. Generate the entries of $\boldsymbol{\epsilon}_i \in \mathbb{R}^{J_y d}$ as independent samples from $N(0, \sigma^2)$.

3. Generate $\mathbf{y}_i(t) = \left\{\mathbf{I}_d \otimes \mathbf{\Psi}^{\mathrm{T}}(t)\right\}\left\{(\mathbf{U}_0 \otimes \mathbf{I}_{J_y})\mathbf{A}_0^*\mathbf{B}_0^{*\mathrm{T}}(\mathbf{V}_0 \otimes \mathbf{I}_{J_x})^{\mathrm{T}}(\mathbf{I}_p \otimes \mathbf{J}_{\phi\phi})\mathbf{x}_i + \boldsymbol{\epsilon}\right\}$ for uniformly distributed time points t_v, $v = 1, \ldots, m$ in $\mathcal{T} = [0,1]$.

The entries of $\mathbf{A}_0^* \in \mathbb{R}^{J_y r_y \times r}$ and $\mathbf{B}_0^* \in \mathbb{R}^{J_x r_x \times r}$ are independent samples from $N(0,1)$, and $\mathbf{U}_0 \in \mathbb{R}^{d \times r_y}$ and $\mathbf{V}_0 \in \mathbb{R}^{p \times r_x}$ are generated by orthogonalizing random matrices of independent $N(0,1)$ entries via QR decomposition.

Two settings of model dimensions are considered:

Setting 1 : $n = 100$, $m = g = 60$, $p = 10$, $d = 10$, $r = 5$, $r_x = 3$, $j_x = 8$, $r_y = 3$, $j_y = 8$.

Setting 2 : $n = 100$, $m = g = 100$, $p = 20$, $d = 20$, $r = 3$, $r_x = 3$, $j_x = 8$, $r_y = 3$, $j_y = 8$.

In Setting 1, the model dimensions, $pj_x = 80, dj_y = 80$ are comparable and slightly smaller than the sample size, but the number of unknowns, 80×80, is already very large. In Setting 2, the model dimensions are much higher than the sample size, i.e., $pj_x = 160, dj_y = 160$, and the total number of unknowns is four times that in Setting 1. For each setting, we try different signal to noise ratios (SNR $\in \{1, 2, 4\}$), defined as the ratio between the standard deviation of all the elements in the response matrix $(\mathbf{U}_0 \otimes \mathbf{I}_{J_y})\mathbf{A}_0^*\mathbf{B}_0^{*\mathrm{T}}(\mathbf{V}_0 \otimes \mathbf{I}_{J_x})^{\mathrm{T}}(\mathbf{I}_p \otimes \mathbf{J}_{\phi\phi})(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ and the noise level σ, and different design correlations ($\rho \in \{0.1, 0.5, 0.9\}$). The ranks and tuning parameters (if there is any) for other methods are selected by 10-fold cross validation. For methods with nested reduced-rank structure,

we also experiment with the proposed BIC criterion to select ranks; see Supplementary Material A.3 for details. The experiment is replicated 300 times for each setting.

To evaluate the performance of different methods, we compute for each method the trimmed mean squared prediction error (MSPE) from all runs (the smallest and largest 20 observations are deleted from 300 runs) based on an independent testing set of size $n_{te} = 500$, i.e, $\text{MSPE}(\widehat{\mathbf{C}}, \mathbf{C}_0) = \|\mathbf{Y}_{te} - \mathbf{X}_{te}\widehat{\mathbf{C}}\|_{\text{F}}^2/n_{te}$ where $\mathbf{Y}_{te}$ and $\mathbf{X}_{te}$ are the integrated response and predictor matrices. Similarly, to evaluate the estimation of the functional responses, we compute the trimmed mean squared functional prediction error (MSFPE) defined as $\sum_{i=1}^{n_{te}}\sum_{v=1}^{m}\|\mathbf{y}_{te,i}(t_v) - \widehat{\mathbf{y}}_{te,i}(t_v)\|_{\text{F}}^2/n_{te}$.

Tables 1 and 2 present the prediction errors (MSPE) under Settings 1 and 2, respectively. The results from OLS are omitted as they are much worse than those of the other methods. Among the five methods presented, RRR has the worst performance. The performance of NRRR is slightly better than that of NRRR-X. RRS substantially improves its corresponding counterpart RRR by incorporating ℓ_2 shrinkage estimation. In general, the improvement is more substantial when the SNR is low and/or the design correlation is high. In contrast, in most scenarios, NRRS only slightly outperforms or has comparable performance to NRRR. This is because NRRR has already considered a more delicate low-dimensional structure so that the extra shrinkage becomes less effective. Due to space limitations, we present the results on estimating r, r_x and r_y in Supplementary Material C.1. RRR usually leads to an underestimation of r; this is expected as RRR tries to use an overall low-rank structure to mimic the finer or even lower-dimensional nested low-rank structure. NRRR performs well in rank estimation in general. The results confirm that NRRR can produce a more interpretable model with improved predictive accuracy.

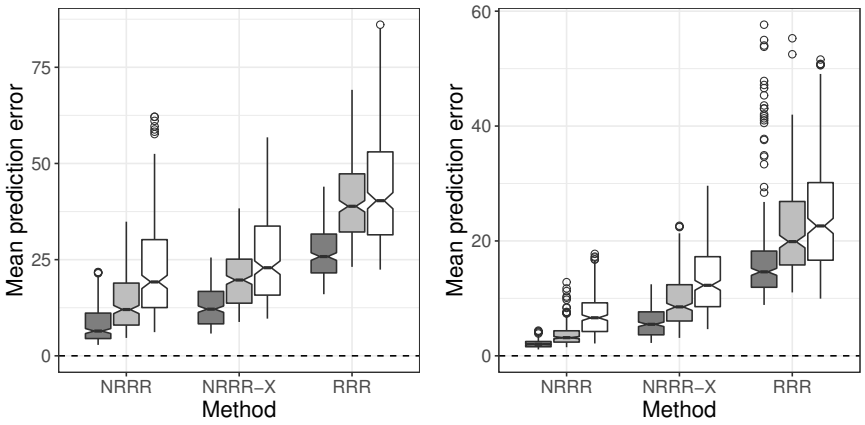
To visualize the effects of nested low-rank dimension reduction, Figure 2 displays the boxplots of MSFPE for NRRR, NRRR-X, and RRR under Settings 1 and 2 with SNR = 1, and Figure 3 draws two particular sets of the true and predicted curves by NRRR, RRR, and OLS from the simulation. The efficacy of the nested dimension reduction is apparent. The results under other settings deliver the same message and hence are omitted. Except for RRR and RRS, all the above results are obtained using BIC to select the model ranks.

Table 1: Simulation results on model estimation under Setting 1. Reported are the mean MSPE values with their standard deviations in parentheses. To improve the presentation, all values are multiplied by 10.

	ρ	NRRR	NRRR-X	RRR	RRS	NRRS
	0.1	11.43 (2.64)	12.16 (2.81)	14.47 (3.16)	11.34 (2.46)	10.97 (2.50)
SNR = 1	0.5	18.14 (4.28)	19.07 (4.33)	22.42 (4.92)	17.46 (3.81)	17.61 (4.18)
	0.9	26.20 (9.17)	26.58 (9.01)	29.56 (9.92)	23.87 (8.04)	25.6 (8.92)
	0.1	2.68 (0.56)	2.84 (0.59)	3.84 (0.80)	3.08 (0.59)	2.77 (0.55)
SNR = 2	0.5	4.18 (1.01)	4.47 (1.10)	5.91 (1.40)	4.56 (1.06)	4.20 (0.99)
	0.9	6.42 (2.19)	6.79 (2.31)	8.26 (2.58)	6.48 (2.09)	6.29 (2.06)
	0.1	0.65 (0.14)	0.68 (0.15)	0.92 (0.21)	0.96 (0.19)	0.77 (0.17)
SNR = 4	0.5	1.04 (0.26)	1.08 (0.27)	1.47 (0.38)	1.31 (0.29)	1.14 (0.27)
	0.9	1.52 (0.52)	1.61 (0.55)	2.11 (0.70)	1.69 (0.53)	1.59 (0.51)

Table 2: Simulation results on model estimation under Setting 2. The layout is the same as in Table 1.

	ρ	NRRR	NRRR-X	RRR	RRS	NRRS
	0.1	6.20 (1.47)	6.56 (1.55)	7.82 (1.98)	6.97 (1.60)	6.30 (1.50)
SNR = 1	0.5	9.76 (3.15)	10.33 (3.38)	11.82 (3.91)	10.55 (3.29)	9.76 (3.12)
	0.9	14.44 (5.72)	15.06 (5.87)	16.21 (6.40)	14.88 (5.76)	14.28 (5.67)
	0.1	1.56 (0.41)	1.58 (0.41)	2.40 (1.11)	2.07 (0.49)	1.61 (0.43)
SNR = 2	0.5	2.46 (0.74)	2.51 (0.74)	3.16 (0.98)	3.08 (0.88)	2.49 (0.73)
	0.9	3.28 (1.22)	3.40 (1.28)	3.86 (1.46)	3.91 (1.66)	3.34 (1.22)
	0.1	0.37 (0.10)	0.38 (0.10)	1.05 (1.01)	0.78 (0.18)	0.42 (0.16)
SNR = 4	0.5	0.61 (0.19)	0.62 (0.19)	0.95 (0.28)	0.91 (0.23)	0.63 (0.19)
	0.9	0.88 (0.35)	0.90 (0.36)	1.05 (0.41)	1.07 (0.44)	0.89 (0.37)



(a) Setting 1 (b) Setting 2

Figure 2: Boxplots of MSFPE under Settings 1 & 2. Each set of three boxplots for $\rho = 0.1, 0.5, 0.9$ is showing in black, grey, and white colors from left to right.

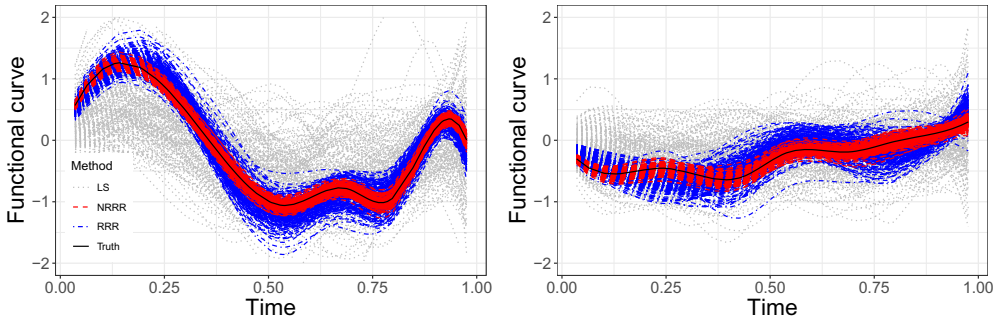


Figure 3: Comparison of the true curves and the predicted curves in two simulation runs under Setting 1 with SNR $= 2$ and $\rho = 0.5$.

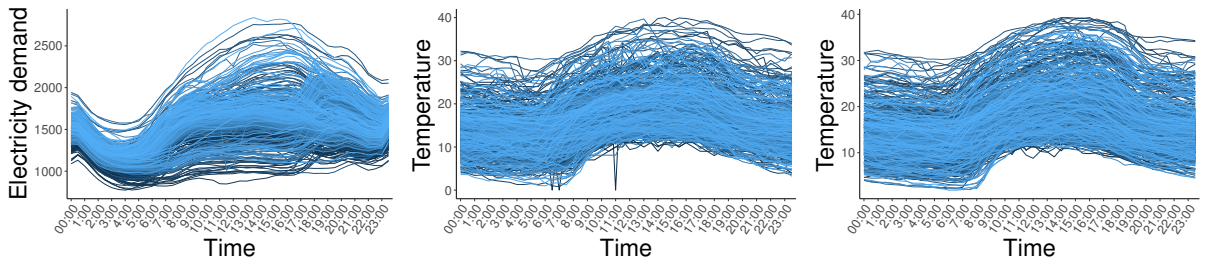
The results obtained from using 10-fold cross validation for all methods are similar and presented in Supplementary Material C.2.

5 Application to Adelaide Electricity Demand Data

Adelaide is the capital city of the state of South Australia. The city has a Mediterranean climate, with warm-dry summers and cool-mild winters. In the summertime, the cooling mainly depends on air conditioning, which makes the electricity demand highly dependent on the weather conditions, and large volatility in temperature throughout the day could make stable electricity supply challenging. Therefore, it is of great interest to understand the dependence and predictive association between the electricity demand and the temperature for facilitating electricity supply management (Magnano, 2007; Magnano et al., 2008; Fan and Hyndman, 2015). Here we apply NRRR to perform a multivariate functional regression analysis between daily half-hour electricity demand profiles for the seven days of a week and the corresponding temperature profiles for the seven days of the same week.

Half-hourly temperature records at two locations, Adelaide Kent town and Adelaide airport, are available between 7/6/1997 and 3/31/2007. Also available are the half-hourly electricity demand records of Adelaide for the same period. As such, for each day during the period, there are three observed functional curves, each with 48 half-hourly observations. As an illustration, Figure 4 plots the temperature and electricity demand profiles of all the Mondays from 7/6/1997 to 3/31/2007. Since our primary focus is on studying the general

association between the within-day demand and temperature trajectories in a week, we center the 48 discrete observations of each daily curve to remove the between-day trend and seasonality of the data. Each week is then treated as a replication.



(a) (b) (c)

Figure 4: Adelaide electricity demand analysis: (a) electricity demand in Adelaide, (b) temperature in Kent town, and (c) temperature at the airport. Plotted are the half-hourly observed profiles for all Mondays.

After data pre-processing, we use the daily half-hour electricity demand as the multivariate functional response with $d = 7$ (corresponding to 7 days in a week from Monday to Sunday), and as for the predictors, we consider two settings. In the first setting, we only use the half-hour temperature data from Kent as the multivariate functional predictors, so that $p = 7$; in the second setting, we also include temperature data from the airport to make $p = 14$. Not surprisingly, the two sets of temperature data are extremely highly correlated, so the second setting is meant to test for the behaviors of different methods in the presence of high collinearity. In either setting, the total sample size is $n = 508$, equaling the number of weeks in the study period. To leave sufficient flexibility in estimating the regression surface, we use B-spline with 30 degrees of freedom to convert the discrete observations to the integrated form according to (6).

First, we compare different methods using an out-of-sample random splitting procedure. Each time, we randomly select 400 samples as the training set and the remaining 108 samples as the test set. The model is fitted using the training data, and the relative mean squared prediction error (RMSPE) is then computed based on the test data, i.e., $\text{RMSPE}(\widehat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n_{te}} \sum_{i=1}^{n_{te}} \int \|\mathbf{y}_{te,i}(t) - \widehat{\mathbf{y}}_{te,i}(t)\|^2 dt / \int \|\mathbf{y}_{te,i}(t)\|^2 dt$. The procedure is repeated 100 times, and the results are reported in Table 5. NRRR and NRRS perform very well in

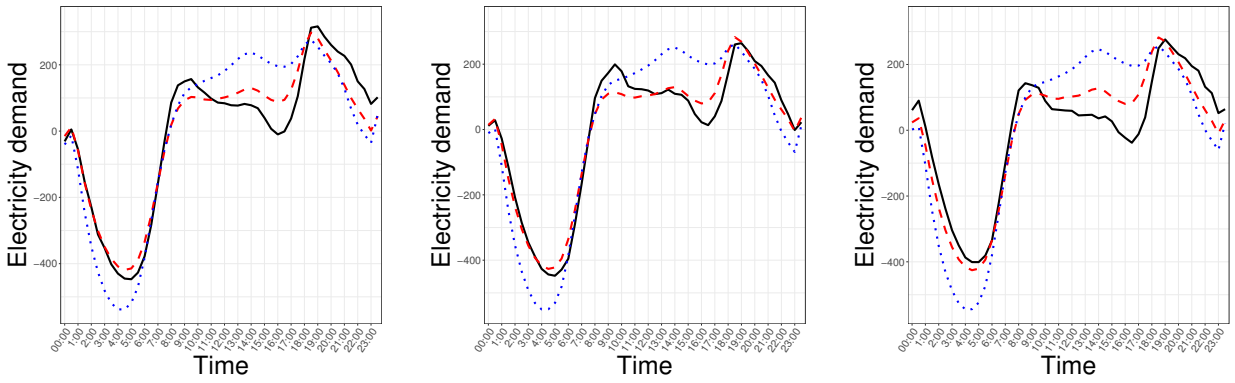


Figure 5: Adelaide electricity demand analysis: selected examples of observed demand curves (solid) and out-of-sample predicted curves by RRR (dotted) and NRRR (dashed).

both settings, and their predicted curves can account for about 74% of the total variation in the observed demand curves. The results show a dramatic global dimension reduction of the functional predictors, as r_x is estimated to be only 1 most of the time. As r_y is often close to the number of original functional responses, this indicates that each daily electricity demand curve has its own pattern, and thus there is not much room for a global dimension reduction. In contrast, RRR and RRS perform much worse in prediction, and RRR even fails in Setting 2. To visualize, Figure 5 plots some randomly selected observed and predicted curves under Setting 1; the superiority of NRRR is apparent. These results clearly show the power and necessity of global dimension reduction, especially in the presence of a high correlation among the functional predictors.

Table 3: Adelaide electricity demand analysis: out-of-sample performance of different methods. Reported are the means and standard deviations (in parenthesis) of RMSPE, r, r_x, and r_y over 100 simulation runs.

	Methods	RRR	NRRR	RRS	NRRS
Setting 1	RMSPE	0.42 (0.04)	0.27 (0.02)	0.38 (0.03)	0.26 (0.02)
	r	1.53 (0.63)	4.06 (0.28)	3.60 (0.70)	4.09 (0.35)
	r_x		1.00 (0.00)		1.01 (0.10)
	r_y		5.70 (1.47)		5.73 (1.43)
Setting 2	RMSPE	1.08 (0.19)	0.26 (0.02)	0.55 (0.05)	0.26 (0.02)
	r	0.26 (0.44)	4.45 (0.89)	1.00 (0.00)	4.40 (0.80)
	r_x		1.00 (0.00)		1.01 (0.10)
	r_y		6.72 (0.57)		6.75 (0.52)

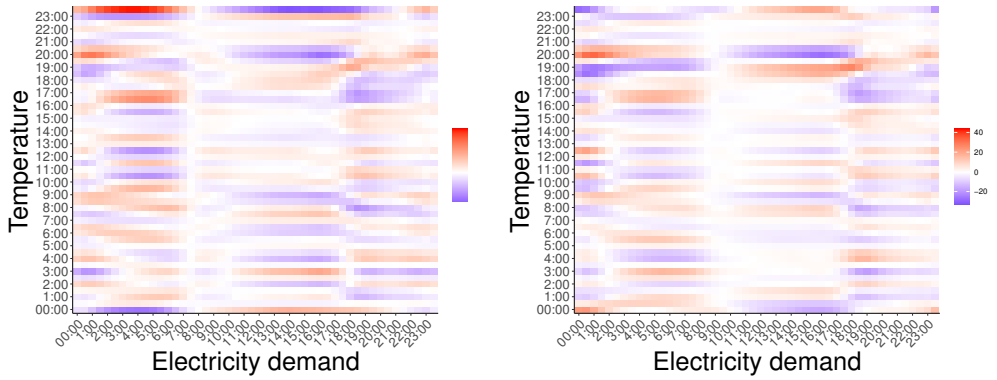
We then use all data to fit a final NRRR model with only the temperature observations

from Kent. The estimated rank values are $\widehat{r} = 4$, $\widehat{r}_x = 1$, and $\widehat{r}_y = 5$. The estimated loading matrix for the predictors is $\widehat{\mathbf{V}} = (0.22, 0.39, 0.46, 0.52, 0.43, 0.28, 0.25)^{\mathrm{T}}$. This shows that only one latent functional predictor is driving the electronic demands patterns, and this factor can be roughly explained as the averaged daily temperature profile of the week. It appears that the days closer to the middle of the week load higher. There is not much global reduction on the response side, as the estimated loading matrix $\widehat{\mathbf{U}}$ is of rank 5. To make sense of $\widehat{\mathbf{U}}$, it may be more convenient to examine the two basis vectors of its orthogonal complement, i.e., the first two singular vectors of $\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\mathrm{T}}$, which give the latent response factors that are not related to the temperatures at all. While the first loading vector $(-0.52, 0.36, 0.28, 0.25, -0.56, 0.34, -0.18)^{\mathrm{T}}$ is hard to interpret, the second loading vector $(0.00, -0.68, 0.73, 0.00, -0.04, 0.05, -0.04)^{\mathrm{T}}$ clearly indicates that the difference between the electronic demand profiles of Tuesday and Wednesday is mostly a noise process. In other words, the demand profiles of these two days are related to the temperature process in almost the same way.

Let $\widetilde{\mathbf{u}}_k$ be the kth row of $\widehat{\mathbf{U}}$. Then Model (4) shows that the estimated regression surface

$$\widetilde{c}_k(s,t) = \widetilde{\mathbf{u}}_k^{\mathrm{T}}(\mathbf{I}_{\widehat{r}_y} \otimes \mathbf{\Psi}^{\mathrm{T}}(t))\widehat{\mathbf{A}}^{*}\widehat{\mathbf{B}}^{*\mathrm{T}}(\mathbf{I}_{\widehat{r}_x} \otimes \mathbf{\Phi}(s)), \qquad k = 1, \ldots, d, \tag{12}$$

would indicate how the response $y_k(t)$ is related to the latent predictor $\widehat{\mathbf{V}}^{\mathrm{T}}\mathbf{x}(s)$ over s and t. In the context of this application, $\widetilde{c}_k(s,t)$ shows how the electricity demand trajectory on the kth day of a week is related to the trajectory of the week's average temperature. We therefore plot the heatmaps of these surfaces to visualize. Figure 6 displays the plots for Tuesday and Saturday. While the patterns of the association are hard to comprehend in general, some observations can be made. First, there are three association regimes throughout each day, i.e., night hours from about midnight to 7:30, daylight hours from about 7:30 to 18:00, and the rest hours from about 18:00 to midnight. This corresponds well with the general patterns of daily electricity demand, and the three regimes are separated by the "Morning ramp", i.e., the transition from relatively lower loads to higher loads in the morning, and the peak load time around 18:00. Noticeably, the electricity demand in daylight hours is the least associated with the temperature. Another observation is that



(a) Tuesday (b) Saturday

Figure 6: Adelaide electricity demand analysis: heatmaps of estimated regression surfaces defined in (12).

temperatures between about 19:00 to 20:30 and 23:00 to 00:00 in general have the largest effects on the electricity demand. This may be related to household and entertainment activities. Lastly, we observe that the association patterns on the workdays are similar and are slightly different from those on the weekends.

6 Discussion

There are many research directions that stem from the proposed nested reduced-rank estimation framework. Our method can be extended to the historical functional regression, i.e., when s and t are both on the same domain such as time, it is required that $\mathbf{C}_0(s,t) = 0$ for any $s > t$, so that the future dynamics of $\mathbf{x}(s)$ is not used in the modeling of the current or past dynamics of $\mathbf{y}(t)$. Another interesting direction is to consider sparse and low-rank estimation. For example, to enable the selection of the functional predictors, we could assume that $\mathbf{V}_0$ is a row-sparse matrix and utilize group-wise regularization such as group lasso in estimation. We have taken a basis expansion approach to deal with the functional aspect of the problem, where we assume the basis functions are given. An alternative is to take a functional principal component analysis or functional canonical correlation analysis, in which the basis functions are obtained as the eigenfunctions of the covariance operators of $\mathbf{y}(t)$ and $\mathbf{x}(s)$. While with any given number of components such a data-driven basis

expansion has the advantage of explaining most of the variations in the ℓ_2 sense, the analysis is much more complicated as it then involves the estimation of the unknown basis. For the non-convex NRRR problem, the proposed algorithm enjoys the monotone descending property and works well empirically, and it will be interesting to fully explore its converges properties. In view of Gorski et al. (2007) and Mishra et al. (2017), it is expected that the algorithm can at least converge to some coordinatewise minimum point. On the theoretical side, it is pressing to study the non-asymptotic behavior of NRRR under reasonable conditions on the integrated design matrix originated from the functional setup. Last but not least, we will further explore the nested reduced-rank structure, or even more generally, a multi-resolution reduced-rank structure in other statistical problems such as time series analysis and large-scale matrix and tensor approximation tasks.

Acknowledgment

The research of Ma was supported in part by the U.S. NSF grants DMS-17-12558 and DMS-20-14221, and a UCR Academic Senate CoR Grant. Chen's research was partially supported by U.S. NSF grants DMS-1613295 and IIS-1718798.

SUPPLEMENTARY MATERIAL

Technical details and additional simulation results: The pdf file contains three sections of contents, including A. Computation: details of the computational algorithm and procedures; B. Proofs: proofs of Theorems 1 and 2; and C. Additional simulation: simulation results on rank estimation and cross validation tuning. (NRRR-supp.pdf)

Codes and data: The proposed methods are implemented in an R package named `NRRR`, which is publicly available at `https://github.com/xliu-stat/NRRR`. The scripts for numerical studies are included. The electricity demand dataset is available in the R package `fds` on CRAN. Please read the file README contained in the zip file for more details. (NRRR-supp.zip, zip archive)

References

Bunea, F., She, Y., and Wegkamp, M. H. (2011), "Optimal Selection of Reduced Rank Estimators of High-Dimensional Matrices," *The Annals of Statistics*, 39, 1282–1309.

Cardot, H., Ferraty, F., and Sarda, P. (1999), "Functional Linear Model," *Statistics & Probability Letters*, 45, 11–22.

— (2003), "Spline Estimators for the Functional Linear Model," *Statistica Sinica*, 13, 571–591.

Chen, K., Chan, K.-S., and Stenseth, N. C. (2012), "Reduced Rank Stochastic Regression with a Sparse Singular Value Decomposition," *Journal of the Royal Statistical Society: Series B*, 74, 203–221.

Chen, K., Dong, H., and Chan, K.-S. (2013), "Reduced Rank Regression via Adaptive Nuclear Norm Penalization," *Biometrika*, 100, 901–920.

Chiou, J.-M., Muller, H.-G., and Wang, J.-L. (2003), "Functional Quasi-Likelihood Regression Models with Smooth Random Effects," *Journal of the Royal Statistical Society: Series B*, 65, 405–423.

Chiou, J.-M., Yang, Y.-F., and Chen, Y.-T. (2016), "Multivariate Functional Linear Regression and Prediction," *Journal of Multivariate Analysis*, 146, 301–312.

Ebaid, R. (2008), "Reduced-Rank Regression of Functional Data," Ph.D. thesis, Temple University.

Fan, S. and Hyndman, R. J. (2015), "Forecasting Long-Term Peak Half-Hourly Electricity Demand for South Australia," The Australian Energy Market Operator.

Faraway, J. J. (1997), "Regression Analysis for a Functional Response," *Technometrics*, 39, 254–261.

Gorski, J., Pfeuffer, F., and Klamroth, K. (2007), "Biconvex Sets and Optimization with Biconvex Functions: a Survey and Extensions," *Mathematical Methods of Operations Research*, 66, 373–407.

He, G., Müller, H.-G., Wang, J.-L., and Yang, W. (2010), "Functional Linear Regression via Canonical Analysis," *Bernoulli*, 16, 705–729.

He, K., Lian, H., Ma, S., and Huang, J. Z. (2018), "Dimensionality Reduction and Variable Selection in Multivariate Varying-Coefficient Models with a Large Number of Covariates," *Journal of the American Statistical Association*, 113, 746–754.

James, G. M. (2002), "Generalized Linear Models with Functional Predictors," *Journal of the Royal Statistical Society: Series B*, 64, 411–432.

Kolda, T. G. and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500.

Krzyśko, M. and Smaga, Ł. (2017), "An Application of Functional Multivariate Regression Model to Multiclass Classification," *Statistics in Transition*, 18, 433–442.

Li, X., Xu, D., Zhou, H., and Li, L. (2018), "Tucker Tensor Regression and Neuroimaging Analysis," *Statistics in Biosciences*, 10, 520–545.

Lock, E. F. (2018), "Tensor-on-Tensor Regression," *Journal of Computational and Graphical Statistics*, 27, 638–647.

Luetkepohl, H. (1993), *Introduction to Multiple Time Series Analysis*, New York: Springer Verlag.

Magnano, L. (2007), "Mathematical Models for Temperature and Electricity Demand," Ph.D. thesis, University of South Australia.

Magnano, L., Boland, J., and Hyndman, R. (2008), "Generation of Synthetic Sequences of Half-Hourly Temperature," *Environmetrics*, 19, 818–835.

Matsui, H., Araki, Y., and Konishi, S. (2008), "Multivariate Regression Modeling for Functional Data," *Journal of Data Science*, 6, 313–331.

Mishra, A., Dey, D. K., and Chen, K. (2017), "Sequential Co-Sparse Factor Regression," *Journal of Computational and Graphical Statistics*, 26, 814–825.

Morris, J. S. (2015), "Functional Regression," *Annual Review of Statistics and Its Application*, 2, 321–359.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015), "On the Degrees of Freedom of Reduced-Rank Estimators in Multivariate Regression," *Biometrika*, 102, 457–477.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis*, New York: Springer.

Reinsel, G. C. and Velu, P. (1998), *Multivariate Reduced-Rank Regression: Theory and Applications*, New York: Springer.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Wang, J.-L., Chiou, J.-M., and Muller, H.-G. (2016), "Functional Data Analysis," *Annual Review of Statistics and Its Application*, 3, 257–295.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Linear Regression Analysis for Longitudinal Data," *The Annals of Statistics*, 33, 2873–2903.

Zhu, H., Morris, J. S., Wei, F., and Cox, D. D. (2017), "Multivariate Functional Response Regression, with Application to Fluorescence Spectroscopy in a Cervical Pre-Cancer Study," *Computational Statistics and Data Analysis*, 111, 88–101.