Nonparametric Matrix Response Regression with Application to Brain Imaging Data Analysis

Wei Hu
Department of Statistics, University of California, Irvine
Dehan Kong
Department of Statistical Sciences, University of Toronto
and
Weining Shen
Department of Statistics, University of California, Irvine
April 2, 2019

Abstract

With the rapid growth of neuroimaging technologies, a great effort has been dedicated recently to investigate the dynamic changes in brain activity. Examples include time course calcium imaging and dynamic brain functional connectivity. In this paper, we propose a novel nonparametric matrix response regression model to characterize the association between 2D image outcomes and predictors such as time and patient information. Our estimation procedure can be formulated as a nuclear norm regularization problem, which can capture the underlying low-rank structures of the dynamic 2D images. We develop an efficient algorithm to solve the optimization problem and introduce a Bayesian information criterion for our model to select the tuning parameters. Asymptotic theories including the risk bound and rank consistency are derived. We finally evaluate the empirical performance of our method using numerical simulations and real data applications from a calcium imaging study and an electroencephalography study.

Keywords: Calcium imaging, Electroencephalography, Nonparametric regression, Matrix data, Nuclear norm, Low rank
1 Introduction

Large-scale neuroimaging studies have received increased attention in statistics literature, with applications including calcium imaging, electroencephalography (EEG), magnetic resonance imaging and functional magnetic resonance imaging. The obtained neuroimaging data often takes complex structure, in the form of two-dimensional matrices. For example, the EEG data can be represented as a two-dimensional matrix, where voltage values were measured from multiple electrodes placed on the subject’s scalp for consecutive time points.

With recent explosive development of neuroimaging technologies, in many applications, instead of observing one 2D image, one can observe a sequence of 2D imaging data objects. It is usually of interest to model the dynamic change process of these 2D images (over the study period) and study their associations with other predictors (e.g., health information). Moreover, these image observations often possess additional structural information such as spatial/temporal correlation, low rankness, and sparsity, which may provide useful scientific insight but also imposes additional challenges to statistical analysis. Here we discuss two relevant motivating examples. The first example is related to the fluorescent calcium imaging, which is a popular technique for observing the spiking activity of large neuronal populations. The locations of neurons and the times at which they fire can be observed via a sequence of 2D images taken over the time, typically using two-photon microscopy [Helmchen and Denk, 2005, Petersen et al., 2018]. The scientific question is to identify the major neurons and model their spiking activity over the time. As shown in Figure 2(a), the images are quite noisy and structured-sparse (in the sense that the signal level is low at the boundary for most images), yet contains rich information (the video clip we study in this paper composes 3,000 picture frames). Thus it is important and non-trivial to develop an automatic-yet-flexible pipeline for analyzing such type of data sets.

Our second example is the brain functional connectivity analysis. Functional connectivity
refers to the coherence of the activities among distinct brain regions (Horwitz, 2003), and it provides novel insights on how distributed brain regions are functional integrated (Biswal et al., 1995, 2010; Fox et al., 2005). Generally, studies on the functional connectivity are based on the temporal correlation between spatial remote neurophysiological events (Friston, 1994) with an implicit assumption that the functional connectivity is constant during the observation period. Recently, functional connectivity has been shown to fluctuate over time (Chang, Liu, Chen, Liu, and Duyn, 2013), implying that measures assuming stationarity over a full scan may be too simplistic to capture the full brain activity. Since the initial findings, researchers have investigated the so-called dynamic functional connectivity, see Calhoun et al. (2014); Calhoun and Adali (2016); Preti et al. (2017) for reviews to date. It then makes sense to represent the connectivity as a covariance matrix and model its change over the time. In our second motivating example, we analyze an EEG data set where the goal is to study the dynamic functional connectivity between alcoholic and non-alcoholic individuals. As shown in Figure 5, our developed methodology is capable of revealing a significant difference in terms of image pattern and temporal correlation between two groups of participants.

In this paper, we aim at developing a novel regression approach to quantify the association between 2D image outcomes and the predictors such as time, patient demographics, and other disease predictors. In particular, by including time as a predictor allows us to study the dynamic change of the images. The proposed regression model can also help detect the difference in image outcomes between study groups by including group indicator as a predictor. In contrast to the dominating use of linear models in the literature, we adopt a flexible nonparametric regression model to capture the commonly-seen nonlinear relationship in the data. For example, we performed a preliminary analysis on the calcium imaging data collected by Ilana Witten’s lab at the Princeton Neuroscience Institute (Petersen et al., 2018). Figure 2(b) shows a scatter plot of the changes of fluorescent intensities across time from a randomly selected pixel of the 2D-image.
The scatter plot shows a clear nonlinear pattern, which will be neglected by linear models. Note that classical nonlinear regression approaches such as Nadaraya-Watson method can not be directly used in our case to model the matrix-valued image responses, since doing so is equivalent to vectorizing the 2D-image data, which destroys the underlying spatial information of the image. Instead, we maintain the matrix structure of the image data, and propose a novel low rank nonparametric estimator by solving a nuclear norm regularization problem. By the singular value thresholding algorithm (Cai et al., 2010), we show that our estimator has a closed-form solution for each fixed bandwidth and regularization parameter. To select these tuning parameters, we derive a Bayesian information criterion (BIC) based on our model and estimation procedure. For theoretical justification, we derive the risk bound for our nonparametric estimator. We show that the rank of the true function can be consistently estimated as well.

Compared with the proposed methods in the literature, here we highlight our contributions. First, we propose a novel nonparametric matrix response regression model. There are some related works on (generalized) linear models for matrix-valued data. For example, Zhou and Li (2014) proposed a class of regularized matrix linear regression model by treating matrix data as covariates; Wang and Zhu (2017) developed a generalized scalar-on-image regression model via total variation; Ding and Cook (2018) studied the matrix response linear regression model using envelope methods; Kong et al. (2018) proposed a low-rank linear regression model with high-dimensional matrix response and high dimensional scalar covariates. To the best of our knowledge, no work has been done on using nonparametric models for matrix data analysis. Second, our nonparametric estimator is easy to derive and has a closed-form solution, which makes it computationally more efficient than the state-of-art multivariate varying coefficient model (Zhu et al., 2011, 2012). Third, we derive an analytic form of BIC, which is not straightforward in our nonparametric matrix response model. Finally, we develop the asymptotic theories including the risk bound and rank consistency for the proposed nonparametric estimator, which directly
connects to the existing work on nonparametric statistics theory.

The rest of the article proceeds as follows. In Section 2, we introduce a novel nonparametric matrix response regression model and propose a fast algorithm for our low-rank regularized estimation procedure. We further derive a BIC for our model to choose the tuning parameters. Section 3 investigates the theoretical properties of our method. We evaluate finite performance of our method in Section 4. Section 5 illustrates applications of our method to two real datasets from a calcium imaging study and an electroencephalography study.

2 Method

2.1 Model

Suppose we observe a set of 2D-images and some scalar predictors from \(n\) independent study subjects. Let \(Y_i\) be a \(p \times q\) matrix representing the 2D-image from the \(i\)th subject, and \(X_i = (x_{i1}, \ldots, x_{is})^T\) be an \(s \times 1\) vector denoting the scalar covariates of interest (e.g., time and disease predictors). We propose the following nonparametric matrix response model,

\[
E(Y_i|X_i) = g(X_i),
\]

where \(g(\cdot) : \mathbb{R}^s \rightarrow \mathbb{R}^{p \times q}\) is a nonparametric matrix-valued function that quantifies the nonlinear relationship between (each pixel of) \(Y_i\) and \(X_i\). Since \(g(x)\) is a \(p \times q\) matrix for all values of \(x\), we will also impose a structure constraint on \(g\) for scientific interpretability and regularization purpose.

Our goal is to estimate the nonparametric function \(g\). A commonly used estimator is the Nadaraya-Watson estimator for the matrix data, which can be written as

\[
\hat{g}_{\text{NW}}(x) = \frac{\sum_{i=1}^{n} K_H(x - X_i)Y_i}{\sum_{i=1}^{n} K_H(x - X_i)},
\]
where $K_H(\cdot) = \frac{1}{|H|}K(H^{-1} \cdot)$, $K(\cdot)$ is a kernel function, and $H = \text{diag}(h_1, h_2, \cdots, h_s)$ is a bandwidth matrix. It is often assumed that $h_1 = \cdots = h_s = h$ for computational convenience.

However, the Nadaraya-Watson estimator is a “naive” estimator in our case because it does not utilize the underlying structure of the matrix response $Y_i$. In particular, the estimator in (2) can also be obtained by vectorizing $Y_i$, applying the Nadaraya-Watson estimator for the vectorized data, and transforming the estimator back to a matrix. To account for the matrix structure, we take another look at the estimator in (2), which can be obtained by solving the following optimization problem

$$
\hat{g}_{\text{nw}}(x) = \arg\min_Y \sum_{i=1}^{n} K_H(x - X_i) \|Y_i - Y\|_F^2,
$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

To further exploit the underlying structure of the 2D response, we introduce a penalty on $Y$ and propose to solve

$$
\hat{g}(x) = \arg\min_Y \left\{ \frac{1}{2n} \sum_{i=1}^{n} K_H(x - X_i) \|Y_i - Y\|_F^2 + \lambda_n \|Y\| \right\},
$$

where $\lambda_n$ is the tuning parameter and $\|\cdot\|$ is some norm of a matrix. Possible choices are nuclear norms, total variation norms, and their combination; and each of those norms will have different regularization effects on the image outcomes. For this paper, we mainly focus on the nuclear norm regularization for illustration, i.e., writing $\|Y\|$ as $\|Y\|_*$, which is defined as the sum of all singular values of the matrix $Y$. The nuclear norm is very popular in 2D-image denoising (Gu et al., 2014). The underlying true 2D-image is often of low rank or approximately low rank, and the nuclear norm regularization can help recover the low rank structure given a noisy image (Chen et al., 2013). In our case, $\hat{g}(x)$ can be regarded as an image estimate at the point $x$, and therefore the penalty $\|Y\|_*$ can push for a low rank representation of the image estimate.
It can be shown that solving (4) is equivalent to solving
\[
\hat{g}(x) = \arg\min_Y \left\{ \frac{1}{2} \| \hat{g}_{nw}(x) - Y \|_F^2 + \frac{n \lambda_n}{\sum_{i=1}^n K_H(x - X_i)} \| Y \|_* \right\}.
\] (5)

The optimization problem (5) can be solved using the following proposition restated from Cai et al. (2010).

**Proposition 1.** Consider the singular value decomposition of a matrix \( Y \in \mathbb{R}^{p \times q} \) with rank \( r \),
\[
Y = U \Sigma V^*, \quad \Sigma = \text{diag}(\{\sigma_j\}_{1 \leq j \leq r}),
\]
where \( U \) and \( V \) are \( p \times r \) and \( q \times r \) matrices respectively with orthonormal columns, and singular values \( \sigma_j \) are positive. The soft-thresholding operator \( D_\tau \) is defined as
\[
D_\tau(Y) = UD_\tau(\Sigma)V^*, \quad D_\tau(\Sigma) = \text{diag}(\{(\sigma_j - \tau)^+\}_{1 \leq j \leq r}),
\]
where \((\cdot)^+\) is the positive part of \((\cdot)\). Then \( D_\tau(Y) \) satisfies
\[
D_\tau(Y) = \arg\min_X \left\{ \frac{1}{2} \| Y - X \|_F^2 + \tau \| X \|_* \right\},
\]
where \( \| X \|_* \) is defined as the nuclear norm of the matrix \( X \).

By Proposition 1, our estimator in (4) can be obtained using the following algorithm.

**Algorithm 1** Algorithm to solve the optimization problem (4).

**Input:** \( \{(X_i, Y_i), 1 \leq i \leq n\}, x, H, \lambda_n \).

**Step 1:** Perform singular value decomposition of \( \hat{g}_{nw}(x) = \frac{\sum_{i=1}^n K_H(x - X_i)Y_i}{\sum_{i=1}^n K_H(x - X_i)} \), and denote it by \( U \Sigma V^* \). The diagonal matrix \( \Sigma = \text{diag}(\{\sigma_j\}_{1 \leq j \leq r}) \), where \( \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0 \) with \( r \) being the rank of \( \Sigma \).

**Step 2:** Set \( \tau = \frac{n \lambda_n}{\sum_{i=1}^n K_H(x - X_i)} \), and calculate the soft-thresholding operator \( D_\tau(\Sigma) = \text{diag}(\{(\sigma_j - \tau)^+\}_{1 \leq j \leq r}) \).

**Step 3:** Calculate \( \hat{g}(x) = UD_\tau(\Sigma)V^* \).

**Output:** \( \hat{g}(x) \).
2.2 Bayesian information criterion

The optimization problem (4) involves two tuning parameters, the bandwidth $h$ and the regularization parameter $\lambda$. The choices of these two parameters are critical as they control the temporal smoothing level and the spatial low-rank level, respectively. In this paper, we derive a Bayesian information criterion (BIC) to select them. Define $\tilde{\lambda} = \frac{n\lambda}{\sum_{i=1}^{n} K_H(x - X_i)}$ and $\hat{Y}_i(\tilde{\lambda}) = \hat{g}(X_i)$. Without loss of generality, we assume $p \geq q$ and denote the singular values of $\hat{Y}_i(\tilde{\lambda})$ by $b_{i1}(\tilde{\lambda}) \geq \cdots \geq b_{iq}(\tilde{\lambda}) \geq 0$. From Algorithm 2.1, it can be seen that the singular values of $\hat{Y}_i(\tilde{\lambda})$ are corresponding truncated singular values of $\hat{g}_{nw}(X_i)$.

Since we are considering a least squared error loss in (4), the BIC can be defined as

$$\text{BIC}(\tilde{\lambda}) = npq \log\left( \frac{1}{npq} \sum_{i=1}^{n} \|Y_i - \hat{Y}_i(\tilde{\lambda})\|_F^2 \right) + \log(npq)df(\tilde{\lambda}),$$

where $df(\tilde{\lambda})$ is given in the following proposition.

**Proposition 2.** Denote $\hat{g}_{nw}(X_i)$’s singular values by $\sigma_{i1} \geq \sigma_{i2} \geq \cdots \geq \sigma_{ir_i} > 0$ and $\sigma_{ik} = 0$ for $k > r_i$. An unbiased estimator of the degree of freedom $df(\tilde{\lambda})$ is

$$\hat{d}f(\tilde{\lambda}) = K_H(0) \sum_{i=1}^{n} \frac{df_i(\tilde{\lambda})}{\sum_{j=1}^{n} K_H(X_i - X_j)},$$

where

$$df_i(\tilde{\lambda}) = \sum_{k=1}^{q} 1\{b_{ik}(\tilde{\lambda}) > 0\} \left\{ 1 + \sum_{1 \leq j \leq p, j \neq k, k \leq r_i} \frac{\sigma_{ik} - \sigma_{ij}}{\sigma_{ik}^2 - \sigma_{ij}^2} + \sum_{1 \leq j \leq q, j \neq k, k \leq r_i} \frac{\sigma_{ik} - \sigma_{ij}}{\sigma_{ik}^2 - \sigma_{ij}^2} \right\}. \quad (8)$$

3 Theory

In this section, we present theoretical results of the estimation procedure in Eq. (4), including a risk bound of the regularized estimator and a rank consistency result. Denote the strength of regularization as $\lambda_n$ and the true response as $g(X)$ given covariates $X$. Assume $g(X)$ has
unknown rank \( r \) and denote the global minimizer of (4) by \( \hat{g}(X) \). For any two sequences of real numbers \( a_n \) and \( b_n \), we write \( a_n \preceq b_n \) if there exists universal positive constants \( C_1 \) and \( C_2 \) such that \( C_1 b_n \leq a_n \leq C_2 b_n \). We define \( a \vee b = \max(a, b) \) and \( a \wedge b = \min(a, b) \) for any \( a, b \in \mathbb{R} \).

With a little abuse of the notation, we use \( C \) to denote a universal constant whose value may change in different context but does not affect the results. For a matrix \( A \) and a sequence of real numbers \( a_n \), we write \( A = \mathcal{O}(a_n) \) or \( A = \mathcal{O}(a_n) \) if every element of \( A \) is \( \mathcal{O}(a_n) \) or \( \mathcal{O}(a_n) \).

Let \( g_{jk}(x) \) be the \((j, k)\)-th component of \( g(x) \). We make the following assumptions:

**Assumption 1.** We assume that \( |g_{jk}(x) - g_{jk}(y)| < C \|x - y\|^\alpha_2 \) with \( \alpha_2 > 0 \), \( 1 \leq j \leq p \), \( 1 \leq k \leq q \) for any \( \|x - y\| < \delta \) and some \( C > 0 \), when \( \delta > 0 \) is sufficiently small.

**Assumption 2.** Assume that \( npq h^{2\alpha_2 + s} \to \infty \), \( nh^{2s} \to \infty \), and \( pqh^{2\alpha_2} \to 0 \) as \( n \to \infty \).

**Assumption 3.** We assume that the kernel function \( K(\cdot) \) is bounded on \( \mathbb{R}^s \). In other words, there exists a constant \( k_{\text{max}} > 0 \) such that \( K(x) \leq k_{\text{max}} \) and \( K_H(x) \leq h^{-s} k_{\text{max}} \) for any \( x \in \mathbb{R}^s \). Moreover, we assume that there exist constants \( C_f, c_f > 0 \) such that the density function of the covariate \( x \) satisfies \( c_f \leq f(x) \leq C_f \).

**Assumption 4.** Assume that \( n^{-1/2} (p \wedge q) \to 0 \), \( (pq)^{1/2} h^{\alpha_2} (p \wedge q)^{1/2} \lambda_n^{-1} \to 0 \), \( \lambda_n (p \wedge q)^2 \to 0 \), and \( \frac{pq (p \wedge q)}{n \lambda_n^2} \to 0 \) as \( n \to \infty \).

Assumption 1 assumes the \( \alpha_2 \)-smoothness for each element of the nonparametric function \( g(x) \). This assumption is commonly used in multivariate function estimation literature such as [Scott (2015)]#Scott2015. It is possible to extend the results for anisotropic case in the future work based on the techniques developed in this paper. Assumptions 2 and 4 are required for estimation consistency and rank consistency. Assumption 3 is satisfied for most kernel density functions. The boundedness condition for the density function of \( x \) will hold if \( x \) is defined on a compact support. With these assumptions, we can state the two main theorems of this paper. Their proofs are given in the Appendix.
Theorem 3. Suppose that Assumptions 1–3 hold. We consider two cases for $p$ and $q$ (e.g., whether they diverge or not).

1. If both $p$ and $q$ are fixed, let $h \approx \left(\frac{\log n}{n}\right)^{\frac{1}{2\alpha_2+s}}$ and $\lambda_n \approx h^{\alpha_2}$, then

$$\|\hat{g}(x) - g(x)\|_F^2 \leq C r \left(\frac{\log n}{n}\right)^{\frac{2\alpha_2}{2\alpha_2+s}}.$$ 

2. If $p \vee q \to \infty$ and $p \vee q = o\left(\frac{\log n}{n^{\alpha_2}}\right)$, then by letting $h \approx \left(n^{-\frac{1}{2\alpha_2+s}} \land \left(\frac{\log n}{n}\right)^{\frac{\alpha_2}{2s}}\right)$ and $\lambda_n = h^{\alpha_2} (pq)^{1/2}$, we have

$$\|\hat{g}(x) - g(x)\|_F^2 \leq C pqr \left(n^{-\frac{2\alpha_2}{2\alpha_2+s}} \lor \left(\frac{\log n}{n}\right)^{\frac{\alpha_2}{2s}}\right).$$

Note that the risk bound involves two quantities $n^{-\frac{2\alpha_2}{2\alpha_2+s}}$ and $(\log n/n)^{-\frac{\alpha_2}{2s}}$. As the number of predictors $s$ increases, it becomes more difficult to estimate $g(x)$. Meanwhile, $h^s$ is involved when proving strongly restricted convexity of the loss function. A larger value of $s$ indicates smaller probability of the loss function being strong restricted convex. In contrast, $\alpha_2$ describes the smoothness of $g(x)$. A larger $\alpha_2$ leads to a smaller risk bound and a faster convergence rate.

Remark 4. If $p$ and $q$ are fixed and $s \leq 2\alpha_2$, the optimal bandwidth $h$ can be chosen arbitrarily close to $n^{-\frac{1}{2\alpha_2+s}}$, which leads to the same convergence rate (with additional logarithmic factor) for estimating an $\alpha_2$-smooth, $s$-dimensional function without regularization.

Remark 5. When $\max(p, q) \to \infty$, if we further assume $s \leq \alpha_2$ and choose $nh^{2\alpha_2+s} \approx \frac{(\sqrt{p} + \sqrt{q})^2}{pq}$ and $\lambda_n = (pq)^{1/2} \left(\frac{(\sqrt{p} + \sqrt{q})^2}{npq}\right)^{\frac{\alpha_2}{2\alpha_2+s}}$, as we let $\lambda_n \to 0$ and $nh^{2s} \to 0$, we obtain $\max(p, q) = o(n^{\frac{\alpha_2}{2+\frac{s}{2}}})$. This is the necessary condition for $\hat{g}(x)$ being consistent. The assumption $s \leq \alpha_2$ rules out the case where there are too many covariates in the model.
Next we present the rank consistency result. We consider three general cases for different values of \( p \) and \( q \) (e.g., whether they diverge or not) and discuss the corresponding choices of \( \lambda_n \) and \( h \) as follows,

(C1) If both \( p, q \) are fixed, we can choose \( h \asymp \left( \frac{\log n}{n} \right)^{\frac{1}{2\alpha} + \frac{1}{2s}} \) and \( \lambda_n \asymp h^{\alpha_2} \log n \).

(C2) If \( p \land q \) is finite, and \( p \lor q \to \infty \) satisfying \( (\log n)^2 (p \lor q) = o\left( \left(\frac{n}{\log n}\right)^{\frac{\alpha_2}{2}} \land n^{\frac{2\alpha_2}{2s}} \right) \), then we choose \( h = \left( \frac{\log n}{n} \right)^{\frac{1}{2s}} \lor n^{-\frac{1}{2\alpha_2+s}} \) and \( \lambda_n \asymp (p \lor q)^{1/2} h^{\alpha_2} \log n \).

(C3) If \( p \asymp q \), and \( p \to \infty \), then we let \( h \asymp \left( \frac{\log n}{n} \right)^{\frac{1}{2s}} \lor n^{-\frac{1}{2\alpha_2+s}} \) and \( \lambda_n \asymp p^{\frac{3}{2}} h^{\alpha_2} (\log n) \). In addition, we assume \( (\log n)^{\frac{2}{3}} p = o(h^{-2\alpha_2/7}) \).

**Theorem 6.** Suppose that Assumptions [T-4] hold, and one of the cases in (C1)–(C3) holds, then \( \hat{g}(x) \) is consistent and rank consistent, i.e.,

\[
P \{ \text{rank}(\hat{g}(x)) = \text{rank}(g(x)) \} \to 1, \quad \text{as } n \to \infty.
\]

It can be seen that rank consistency requires stronger assumptions on \( p, q \) compared with those in Theorem [T-4]. For instance, the desired \( \lambda_n \) is much larger than the one from the previous theorem. Meanwhile, \( pq \) is not allowed to be greater than \( n \) for rank consistency.

4 Simulation

In this section, we evaluate the performance of our method and other competing methods. We consider both univariate and multivariate \( X \), different nonparametric functions and different correlation structures of the random error \( E_i \), where \( E_i = Y_i - g(x_i) \).
4.1 Univariate predictor

Setting I: We set the dimensions of the image $p = q = 64$, and set the $(j, k)$-th element of the nonparametric function $g(x)_{jk} = \{\sin(10\pi x) + \cos(10\pi x) + 0.1(j+k)\} \ast B_{jk}$, $1 \leq j, k \leq 64$, where $0 \leq x \leq 1$ and $B_{jk}$ is the $(j,k)$-th element of the true signal $B$. The true signal $B$ is generated from a 64-by-64 image, where we consider three shapes: a cross, a square and a T-shape. We have plotted the true shapes in Figure 1(a)(b)(c), where we assign $B$ a value of 5 for black regions and 0 for white regions. The sample size is set at $n = 200, 500$. The covariates $\{x_i\}$, $i = 1, 2, \ldots, n$ are equally spaced on $[0,1]$. The response $Y_i$ is generated from $Y_i = g(x_i) + E_i$, where $\text{vec}(E_i)$’s are i.i.d $N(0, I_{pq})$. The optimal bandwidth $h$ and $\lambda$ are selected by BIC. For the kernel function, we use the standard gaussian kernel defined as $K(x) = \exp\left(-\frac{x^2}{2}\right)/\sqrt{2\pi}$.

We compare our method with the naive Nadaraya-Watson estimator and the Lasso estimator, where the Lasso estimator is obtained by solving the following optimization problem

$$\hat{g}_{\text{lasso}}(x) = \arg\min_Y \left\{ \frac{1}{2n} \sum_{i=1}^{n} K_H(x - X_i) \| Y_i - Y \|_F^2 + \lambda_n \| Y \|_1 \right\}, \quad (9)$$

where $\| Y \|_1$ is defined as the sum of the absolute values of all the elements of the matrix $Y$. We also use the BIC as defined in (7) to choose the tuning parameter for Lasso. Here the degree of freedom can be obtained by the chain rule as

$$\hat{df} = \sum_{i=1}^{n} \text{tr} \left( \frac{\partial \text{vec}(\hat{Y}_i)}{\partial \text{vec}(Y_i)} \right) = \sum_{i=1}^{n} \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{\partial \hat{Y}_{ijk}}{\partial \hat{g}_{ijk(nw)}} \frac{\partial \hat{g}_{ijk(nw)}}{\partial Y_{ijk}}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} K_H(0) \| \text{sign}(\text{vec}(\hat{Y}_i)) \|_1,$$

where we have used the fact that $\frac{\partial \hat{Y}_{ijk}}{\partial \hat{g}_{ijk(nw)}} = |\text{sign}(\hat{Y}_{ijk})|$.

In each Monte Carlo simulation, we generate $n$ samples as the training set and another 500 samples as the test set. We report the integrated error $\int_x \| \hat{Y}(x) - Y(x) \|_F^2 dx$, which can be
Figure 1: (a)-(c): true signals, (d)-(f): recovered signals

approximated by $\frac{1}{500} \sum_{i=1}^{500} \| \hat{Y}(x_{test}^i) - Y(x_{test}^i) \|^2_F$. Table 1 shows the average integrated test error by our method, naive Nadaraya-Watson estimator and Lasso estimator based on 100 Monte Carlo replicates. We also report the average selected rank by our method using BIC, defined as $\frac{1}{n} \sum_{i=1}^{n} \operatorname{rank}(\hat{Y}(x_i))$.

From the results, we can see that our method performs better than Nadaraya-Watson estimator and Lasso estimator in all cases. In addition, our method can estimate the true rank of the image accurately. We have plotted the recovered signals from one randomly selected Monte Carlo study in Figure 1(d)(e)(f), and our method manages to recover the true signals very well.

**Setting II**: In this setting, we consider the case where the errors $\text{vec}(E_i)$ are correlated across different subjects $i$’s and the pixels within the same random error matrix $E_i$ are also correlated. Define $e = (\text{vec}(E_1)^T, \ldots, \text{vec}(E_n)^T)^T \in \mathbb{R}^{pqn}$. We assume $e \sim N(0, \Sigma)$, where $\Sigma = \Sigma_1 \otimes \Sigma_2 \in \mathbb{R}^{pqn \times pqn}$. Here $\Sigma_1$ is a $n \times n$ matrix representing the correlation within different subjects.
Table 1: Simulation results for Setting I: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes $B$. The results are based on 100 Monte Carlo replications.

| Shape | Our method | NW        | Lasso     | Selected rank | True rank |
|-------|------------|-----------|-----------|---------------|-----------|
| Cross | 4458 (0.70) | 6024 (0.96) | 5148 (0.81) | 3.53 (0.004) | 4         |
| n = 200 | 4288 (0.82) | 6107 (0.99) | 4875 (0.83) | 2.00 (0.000) | 2         |
| Tshape | 4472 (0.76) | 6094 (0.98) | 5193 (0.92) | 3.22 (0.006) | 4         |

| Cross | 4306 (0.41) | 5009 (0.52) | 4560 (0.48) | 3.99 (0.000) | 4         |
| n = 500 | 4186 (0.41) | 4803 (0.48) | 4440 (0.45) | 2.01 (0.000) | 2         |
| Tshape | 4255 (0.60) | 5042 (0.51) | 4579 (0.49) | 3.52 (0.007) | 4         |

$1 \leq i \leq n$, $\Sigma_2$ is a $pq \times pq$ matrix representing the correlation among different pixels of the 2D image, and $\otimes$ is the Kronecker product. This decomposition of $\Sigma$ is often referred to as the separability of the covariance matrix, which was studied in various literatures such as De Munck et al. (2002); Dawid (1981). For $\Sigma_1$, we assume it has a subject-wise 1D autoregressive structure. In particular, we set the $(i_1, i_2)$-th element of $\Sigma_1$ as $0.5|i_1 - i_2|$ for $1 \leq i_1, i_2 \leq n$. For $\Sigma_2$, we assume it is incorporated with a pixel-wise 2D autoregressive structure. Specifically, we set the $(j_1 + (k_1 - 1)q, j_2 + (k_2 - 1)q)$-th element of $\Sigma_2$ as $0.5|j_1 - j_2| + |k_1 - k_2|$ for $1 \leq j_1, j_2 \leq p$ and $1 \leq k_1, k_2 \leq q$. The average integrated test errors by three methods and the average selected rank of our method are summarized in Table 2. From the results, we can see that our methods still outperforms than Nadaraya-Watson estimator and Lasso estimator in all cases. Compared with independent error case, one notice that we may over select the rank a bit, possibly due to the error correlations, however, the average integrated errors are still similar for both cases.
Table 2: Simulation results for Setting II: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes $B$. The results are based on 100 Monte Carlo replications.

| Shape | Our method | NW     | Lasso  | Selected rank | True rank |
|-------|------------|--------|--------|---------------|-----------|
|       | $n = 200$ |        |        |               |           |
| Cross | 4656 (2.23)| 6120 (3.47) | 5528 (3.49) | 6.64 (0.009) | 4         |
| Square| 4463 (2.31)| 5785 (3.27) | 5169 (4.72) | 4.42 (0.009) | 2         |
| Tshape| 4667 (2.49)| 6017 (3.65) | 5591 (3.74) | 6.52 (0.008) | 4         |
|       | $n = 500$ |        |        |               |           |
| Cross | 4403 (1.23)| 5240 (1.96) | 4769 (1.71) | 6.90 (0.008) | 4         |
| Square| 4296 (1.10)| 5036 (1.64) | 4692 (1.68) | 4.76 (0.000) | 2         |
| Tshape| 4404 (1.21)| 5057 (1.60) | 4797 (1.69) | 6.75 (0.008) | 4         |

4.2 Multivariate predictors

Setting III: We consider shapes of the image with the same pixel value as setting I. We set the $(j, k)$-th element of the nonparametric function $g(x)_{jk} = \{\sin(2\pi \|x\|) + \cos(2\pi \|x\|) + 0.5(j + k)\} \ast B_{jk}$, $x \in [0, 1] \times [0, 1], 1 \leq j, k \leq 64$, where we consider the same three shapes of the true image $B$ and $\|x\|$ is the $l_2$-norm of $x$. The random error vec($E_i$)’s are i.i.d. $N(0, I_{pq})$. The covariates $x_i$ consist of a set of $\{x_{jk}\}, 1 \leq j \leq 20, 1 \leq k \leq 25$, that are equally spaced on $[0, 1] \times [0, 1]$. The sample sizes $n = 200, 500$ are considered, and the multivariate Gaussian kernel defined as $K(x) = \exp(-\|x\|^2/2)/2\pi$ is used. In each Monte Carlo simulation, we generate $n$ samples as the training set and another 500 samples as the test set. We report the average integrated test error obtained by our method, the naive Nadaraya-Watson estimator and Lasso estimator and the average selected rank of our method based on 100 Monte Carlo replicates in Table 3.
Table 3: Simulation results for Setting III: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes $B$. The results are based on 100 Monte Carlo replications.

| Shape    | Our method | NW       | Lasso    | Selected rank | True rank |
|----------|------------|----------|----------|---------------|-----------|
| n = 200  | Cross      | 4711 (0.86) | 7021 (1.17) | 5759 (1.01) | 4.03 (0.002) | 4         |
|          | Square     | 4518 (0.80) | 6723 (1.09) | 5326 (1.05) | 2.04 (0.002) | 2         |
|          | Tshape     | 4719 (0.83) | 7137 (1.11) | 5829 (1.15) | 4.34 (0.005) | 4         |
| n = 500  | Cross      | 4647 (0.49) | 5562 (0.62) | 4843 (0.48) | 4.84 (0.006) | 4         |
|          | Square     | 4281 (0.42) | 5506 (0.58) | 4649 (0.45) | 2.01 (0.001) | 2         |
|          | Tshape     | 4376 (0.45) | 5620 (0.59) | 4875 (0.49) | 4.12 (0.002) | 4         |

**Setting IV:** We consider the same setting as Setting III except that the random error vec($E_i$)'s are correlated across different $i$'s and the pixels within the same random error matrix $E_i$ are also correlated. The random error vec($E_i$)s are generated the same as Setting II. The average integrated test errors by three methods and the average selected rank of our method are summarized in Table 4.

The findings in the multivariate case (Settings III and IV) are consistent with the ones in the univariate case. The simulation results in this section confirm the excellent performance of the proposed nonparametric estimation procedure.
Table 4: Simulation results for Setting IV: mean of integrated test error and associated standard errors obtained from our method, NW estimator, and Lasso, the average selected rank and true rank are reported for three different shapes $B$. The results are based on 100 Monte Carlo replications.

| Shape  | Our method | NW       | Lasso   | Selected rank | True rank |
|--------|------------|----------|---------|---------------|-----------|
| Cross  | 4894 (2.58)| 7164 (3.89)| 5804 (4.13)| 5.40 (0.008)   | 4         |
|        | n = 200   |          |         |               |           |
| Square | 4642 (2.43)| 6858 (3.57)| 5360 (4.0)| 3.14 (0.009)   | 2         |
| Tshape | 4910 (2.66)| 7283 (3.91)| 5880 (4.36)| 5.36 (0.008)   | 4         |
|        | n = 500   |          |         |               |           |
| Cross  | 4779 (1.73)| 5687 (2.38)| 5067 (1.86)| 6.38 (0.008)   | 4         |
| Square | 4574 (1.51)| 5614 (2.22)| 4815 (1.62)| 4.12 (0.001)   | 2         |
| Tshape | 4797 (1.55)| 5745 (2.09)| 5080 (4.72)| 6.34 (0.008)   | 4         |

5 Real data application

5.1 Application to calcium imaging data study

In this section, we apply the proposed method to one-photon calcium imaging dataset collected by Ilana Witten’s lab at the Princeton Neuroscience Institute (Petersen et al., 2018), which can be downloaded from [https://ajpete.com/software](https://ajpete.com/software). Calcium imaging is an important fluorescent microscopy technique regulating a great variety of neuronal processes simultaneously (Berridge, 1998; Andilla and Hamprecht, 2014). Whenever a neuron fires, voltage-gated calcium channels in the axon terminal open and then calcium floods the cell. Such changes in concentration of calcium ions are detected by observing the fluorescence of calcium indicator molecules. Therefore, not surprisingly, intracellular calcium concentration becomes an important surrogate marker for the spiking activity of neurons in the absence of effective voltage imaging approach.
and is commonly used when analyzing local neuronal circuits in vivo and in vitro (Petersen et al., 2018; Grienberger and Konnerth, 2012).

The calcium imaging data can be viewed as a video clip (i.e., a collection of 2D-images recorded at the same frame over a period of time) that presents the location and time of neuron firing (Apthorpe et al., 2016; Petersen et al., 2018). Each pixel in a frame is continuous-valued and larger values indicate higher fluorescent intensities caused by greater calcium concentrations. The calcium imaging video we used consists of 3000 frames of size $205 \times 226$ pixels sampled at 10 Hz. An example frame randomly selected from the video is shown in Figure 4(a).

![Fig 2](image1.png)

(a) (b)

Figure 2: (a) A sequence of frames (b) scatterplot for a fixed voxel of coordinate (200,60) over frames

Figure 3 gives estimated fluorescent intensities versus true values of two randomly selected pixels across 3000 frames. The optimal bandwidth and regularization parameter are selected by the proposed BIC criteria. The average rank selected by our method is $22.34(\text{SE} = 1.34)$ across all frames. We also plot estimated images from a randomly selected frame by our method in Figure 4(b). From that figure, we can see that our method amplifies potential neuron signals, but
also weakens those unclear and smaller neurons.

We further evaluate the prediction performance of our method by cross-validation. We compare our method with two nonparametric regression methods: Nadaraya-Watson regression and Lasso defined in (3) and (9), respectively. We also compare our method with the low-rank matrix response linear regression (L2RM) method (Kong et al., 2018). As shown in Table 5, our method reaches the smallest leave-one-out cross-validation error among four methods. For Lasso estimator, the selected tuning parameter is always zero, hence making the Lasso estimator equivalent to the Nadaraya-Watson estimator. This confirms that the sparsity assumption does not seem plausible for the calcium imaging application, while low rankness is a more reasonable assumption. The linear model L2RM has the highest prediction error, which validates the use of a nonparametric model for the data set.
Figure 4: (a) Original 1500th frame (b) Estimated 1500th frame by our method

Table 5: Leave-one-out cross-validation errors (Err)($10^6$) and associated standard deviation by three methods for calcium imaging data

| Err - Our method | Err - NW | Err - Lasso | Err - L2RM |
|------------------|---------|------------|------------|
| 2.66(0.63)       | 2.91(1.04) | 2.91(1.04) | 5.45(3.63) |

5.2 Application to EEG data study

We also apply our method to an EEG dataset, which is available at [https://archive.ics.uci.edu/ml/datasets/EEG+Database](https://archive.ics.uci.edu/ml/datasets/EEG+Database). The data was collected from 122 subjects by the Neurodynamics Laboratory to examine the EEG correlates of genetic predisposition to alcoholism. More details about the study can be found in [Zhang et al. (1995)](https://archive.ics.uci.edu/ml/datasets/EEG+Database). Among the 122 subjects, 77 were alcoholic individuals and 45 were controls. The dataset included voltage values from 64 electrodes placed on each subject’s scalps sampled at 256 Hz (3.9- msec epoch) for 1 second. Each subject was exposed to three stimuli: a single stimulus, two matched stimuli, two unmatched stimuli. For each subject, we use the average of all trials for each subject under single-stimulus condition, which results in a $256 \times 64$ matrix. Among those 122 subjects, we randomly select one alcoholic individual and one control, and analyze the dynamic functional connectivity
among different electrodes across time. The simplest analytical strategy to investigate dynamic functional connectivity consists in segmenting the time courses from spatial locations into a set of temporal windows, inside which their pairwise connectivity is probed. By gathering functional connectivity descriptive measures over subsequent windows, fluctuations in connectivity can be captured. The basic sliding window framework has been applied by the neuroimaging community to understand how brain dynamics related to our cognitive abilities (Kucyi and Davis, 2014; Elton and Gao, 2015; Madhyastha and Grabowski, 2014), is affected by brain disorders (Sakoğlu et al., 2010; Jones et al., 2012), or compares to other functional or structural brain measures (Leonardi et al., 2013; Tagliazucchi et al., 2012; Liégeois et al., 2016). More specifically, we use a moving window of size 100 to calculate a series of covariance matrices along dimension of 256, resulting 157 covariance matrices of size $64 \times 64$ for each individual.

We apply the proposed method to analyze the dynamic change of covariance structures over the time in both alcoholic individual and control. The optimal bandwidth and regularization parameter are selected by BIC. Figure 5 shows estimated images of 10th frame by our method for alcoholic individual and control respectively. We observe a significant structural difference in their covariance matrices. Specifically, the alcoholic individual has a more complex covariance structure than that from the control. Moreover, the average selected rank of alcoholic individual is 22.44 (SE = 0.93) compared to 6.82 (SE = 0.87) of control. This can be explained by drastic fluctuation across time in EEG signals of alcoholic individuals compared to stable variation in control. We further evaluate our method by leave-one-out cross-validation error and compare it with Nadaraya-Watson regression, Lasso and L2RM. As shown in Table 6, our method achieves the smallest leave-one-out cross-validation error among three methods. For Lasso estimator, the selected tuning parameter is always zero. In other words, the Lasso estimator is the same as the Nadaraya-Watson estimator for this data application, which implies that the low rankness assumption is a more reasonable assumption than sparsity. We also notice that linear L2RM has
a much higher estimation error than the nonparametric methods. This indicates a strong nonlinear pattern in EEG signals for both alcoholic and control subjects.

Table 6: Leave-one-out cross-validation errors (SE) by three methods for EEG data

|                | Err - Our method | Err - NW  | Err - Lasso | Err - L2RM  |
|----------------|------------------|----------|-------------|-------------|
| Alcoholic      | 1.05(1.20)       | 2.20(2.84)| 2.20(2.84)  | 908.29(623.54)|
| Control        | 9.57(61.44)      | 21.87(118.06)| 21.87(0.75) | 22513.17(16763.92)|

Figure 5: (a) Estimated 10th frame for alcoholic (b) Estimated 10th frame for control

References

Andilla, F. D. and Hamprecht, F. A. (2014), “Sparse space-time deconvolution for calcium image analysis,” in Advances in Neural Information Processing Systems, pp. 64–72.

Apthorpe, N., Riordan, A., Aguilar, R., Homann, J., Gu, Y., Tank, D., and Seung, H. S. (2016), “Automatic neuron detection in calcium imaging data using convolutional networks,” in Advances in Neural Information Processing Systems, pp. 3270–3278.
Berridge, M. J. (1998), “Neuronal calcium signaling,” *Neuron*, 21, 13–26.

Biswal, B., Zerrin Yetkin, F., Haughton, V. M., and Hyde, J. S. (1995), “Functional connectivity in the motor cortex of resting human brain using echo-planar MRI,” *Magnetic Resonance in Medicine*, 34, 537–541.

Biswal, B. B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., Beckmann, C. F., Adelstein, J. S., Buckner, R. L., Colcombe, S., et al. (2010), “Toward discovery science of human brain function,” *Proceedings of the National Academy of Sciences*, 107, 4734–4739.

Cai, J.-F., Candès, E. J., and Shen, Z. (2010), “A singular value thresholding algorithm for matrix completion,” *SIAM Journal on Optimization*, 20, 1956–1982.

Calhoun, V. D. and Adali, T. (2016), “Time-Varying Brain Connectivity in fMRI Data: whole-brain data-driven approaches for capturing and characterizing dynamic states,” *IEEE Signal Processing Magazine*, 33, 52–66.

Calhoun, V. D., Miller, R., Pearlson, G., and Adali, T. (2014), “The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery,” *Neuron*, 84, 262–274.

Chang, C., Liu, Z., Chen, M. C., Liu, X., and Duyn, J. H. (2013), “EEG correlates of time-varying BOLD functional connectivity,” *NeuroImage*, 72, 227–236.

Chen, K., Dong, H., and Chan, K.-S. (2013), “Reduced rank regression via adaptive nuclear norm penalization,” *Biometrika*, 100, 901–920.

Dawid, A. P. (1981), “Some matrix-variate distribution theory: notational considerations and a Bayesian application,” *Biometrika*, 68, 265–274.
De Munck, J. C., Huizenga, H. M., Waldorp, L. J., and Heethaar, R. (2002), “Estimating stationary dipoles from MEG/EEG data contaminated with spatially and temporally correlated background noise,” *IEEE Transactions on Signal Processing*, 50, 1565–1572.

Ding, S. and Cook, D. (2018), “Matrix variate regressions and envelope models,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80, 387–408.

Elton, A. and Gao, W. (2015), “Task-related modulation of functional connectivity variability and its behavioral correlations,” *Human Brain Mapping*, 36, 3260–3272.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005), “The human brain is intrinsically organized into dynamic, anticorrelated functional networks,” *Proceedings of the National Academy of Sciences*, 102, 9673–9678.

Friston, K. J. (1994), “Functional and effective connectivity in neuroimaging: a synthesis,” *Human Brain Mapping*, 2, 56–78.

Grienberger, C. and Konnerth, A. (2012), “Imaging calcium in neurons,” *Neuron*, 73, 862–885.

Gu, S., Zhang, L., Zuo, W., and Feng, X. (2014), “Weighted nuclear norm minimization with application to image denoising,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2862–2869.

Helmchen, F. and Denk, W. (2005), “Deep tissue two-photon microscopy,” *Nature Methods*, 2, 932.

Horwitz, B. (2003), “The elusive concept of brain connectivity,” *NeuroImage*, 19, 466–470.

Jones, D. T., Vemuri, P., Murphy, M. C., Gunter, J. L., Senjem, M. L., Machulda, M. M., Przybelski, S. A., Gregg, B. E., Kantarci, K., Knopman, D. S., et al. (2012), “Non-stationarity in the “resting brain’s” modular architecture,” *PLOS ONE*, 7, e39731.
Kong, D., An, B., Zhang, J., and Zhu, H. (2018), “L2RM: Low-rank Linear Regression Models for High-dimensional Matrix Responses,” *Journal of the American Statistical Association*, 1–55.

Kucyi, A. and Davis, K. D. (2014), “Dynamic functional connectivity of the default mode network tracks daydreaming,” *NeuroImage*, 100, 471–480.

Leonardi, N., Richiardi, J., Gschwind, M., Simioni, S., Annoni, J.-M., Schluep, M., Vuilleumier, P., and Van De Ville, D. (2013), “Principal components of functional connectivity: a new approach to study dynamic brain connectivity during rest,” *NeuroImage*, 83, 937–950.

Liégeois, R., Ziegler, E., Phillips, C., Geurts, P., Gómez, F., Bahri, M. A., Yeo, B. T., Soddu, A., Vanhaudenhuyse, A., Laureys, S., et al. (2016), “Cerebral functional connectivity periodically (de) synchronizes with anatomical constraints,” *Brain Structure and Function*, 221, 2985–2997.

Madhyastha, T. M. and Grabowski, T. J. (2014), “Age-related differences in the dynamic architecture of intrinsic networks,” *Brain Connectivity*, 4, 231–241.

Petersen, A., Simon, N., and Witten, D. (2018), “Scalpel: Extracting neurons from calcium imaging data,” *The Annals of Applied Statistics*, 12, 2430.

Preti, M. G., Bolton, T. A., and Van De Ville, D. (2017), “The dynamic functional connectome: State-of-the-art and perspectives,” *Neuroimage*, 160, 41–54.

Sakoğlu, Ü., Pearlson, G. D., Kiehl, K. A., Wang, Y. M., Michael, A. M., and Calhoun, V. D. (2010), “A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia,” *Magnetic Resonance Materials in Physics, Biology and Medicine*, 23, 351–366.
Scott, D. W. (2015), *Multivariate density estimation: theory, practice, and visualization*, John Wiley & Sons.

Tagliazucchi, E., Von Wegner, F., Morzelewski, A., Brodbeck, V., and Laufs, H. (2012), “Dynamic BOLD functional connectivity in humans and its electrophysiological correlates,” *Frontiers in Human Neuroscience*, 6.

Wang, X. and Zhu, H. (2017), “Generalized scalar-on-image regression models via total variation,” *Journal of the American Statistical Association*, 112, 1156–1168.

Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., and Litke, A. (1995), “Event related potentials during object recognition tasks,” *Brain Research Bulletin*, 38, 531–538.

Zhou, H. and Li, L. (2014), “Regularized matrix regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76, 463–483.

Zhu, H., Kong, L., Li, R., Styner, M., Gerig, G., Lin, W., and Gilmore, J. H. (2011), “FADTTS: functional analysis of diffusion tensor tract statistics,” *NeuroImage*, 56, 1412–1425.

Zhu, H., Li, R., and Kong, L. (2012), “Multivariate varying coefficient model for functional responses,” *Annals of Statistics*, 40, 2634.