Higher Order Partial Least Squares (HOPLS): A Generalized Multilinear Regression Method

Qibin Zhao, *Member*, *IEEE*, Cesar F. Caiafa, *Member*, *IEEE*, Danilo P. Mandic, *Fellow*, *IEEE*, Zenas C. Chao, Yasuo Nagasaka, Naotaka Fujii, Liqing Zhang, *Member*, *IEEE*, and Andrzej Cichocki, *Fellow*, *IEEE*

Abstract—A new generalized multilinear regression model, termed the higher order partial least squares (HOPLS), is introduced with the aim to predict a tensor (multiway array) $\underline{\mathbf{Y}}$ from a tensor $\underline{\mathbf{X}}$ through projecting the data onto the latent space and performing regression on the corresponding latent variables. HOPLS differs substantially from other regression models in that it explains the data by a sum of orthogonal Tucker tensors, while the number of orthogonal loadings serves as a parameter to control model complexity and prevent overfitting. The low-dimensional latent space is optimized sequentially via a deflation operation, yielding the best joint subspace approximation for both $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$. Instead of decomposing $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ individually, higher order singular value decomposition on a newly defined generalized cross-covariance tensor is employed to optimize the orthogonal loadings. A systematic comparison on both synthetic data and real-world decoding of 3D movement trajectories from electrocorticogram signals demonstrate the advantages of HOPLS over the existing methods in terms of better predictive ability, suitability to handle small sample sizes, and robustness to noise.

Index Terms—Multilinear regression, partial least squares, higher order singular value decomposition, constrained block Tucker decomposition, electrocorticogram, fusion of behavioral and neural data

1 Introduction

The partial least squares (PLS) is a well-established framework for estimation, regression, and classification whose objective is to predict a set of dependent variables (responses) from a set of independent variables (predictors) through the extraction of a small number of latent variables. One member of the PLS family is partial least squares regression (PLSR)—a multivariate method which, in contrast to multiple linear regression and principal component regression (PCR), is proven to be particularly suited to highly collinear data [1], [2]. To predict response variables $\mathbf{Y}$ from independent variables $\mathbf{X}$, PLS finds a set of latent variables (also called latent vectors, score vectors, or components) by projecting both $\mathbf{X}$ and $\mathbf{Y}$ onto a new subspace while at the same time maximizing the pairwise covariance between the latent variables of $\mathbf{X}$ and $\mathbf{Y}$. A standard way to optimize the model parameters is the nonlinear iterative partial least squares (NIPALS) [3]; for an overview of PLS and its applications in neuroimaging, see [4], [5], [6]. There are many variations of the PLS model, including orthogonal projection on latent structures [7], biorthogonal PLS (BPLS) [8], recursive PLS [9], nonlinear PLS [10], [11]. The PLSR is known to exhibit high sensitivity to noise, a problem that can be attributed to redundant latent variables [12], whose selection still remains an open problem [13]. Penalized regression methods are also popular for simultaneous variable selection and coefficient estimation which impose, for example, L2 or L1 constraints on the regression coefficients. Algorithms of this kind are Ridge regression and Lasso [14]. The recent progress in sensor technology, biomedicine, and biochemistry has highlighted the necessity of considering multiple data streams as multiway data structures [15] for which the corresponding analysis methods are very naturally based on tensor decompositions [16], [17], [18]. Although matricization of a tensor is an alternative way to express such data, this would result in the "Large p Small n" problem and also make it difficult to interpret the results as the physical meaning and multiway data structures would be lost due to the unfolding operation.

The N-way PLS (N-PLS) decomposes the independent and dependent data into rank-one tensors, subject to maximum pairwise covariance of the latent vectors. This promises enhanced stability, resilience to noise, and intuitive interpretation of the results [19], [20]. Due to these desirable properties, N-PLS has found applications in areas ranging from chemometrics [21], [22], [23] to neuroscience [24], [25]. A modification of the N-PLS and the multiway

- *Q. Zhao is with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Saitama, Japan, and the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. E-mail: qbzhao@brain.riken.jp.*
- *C.F. Caiafa is with the Instituto Argentino de Radioastronomía (IAR), CCT La Plata-CONICET, Buenos Aires, Argentina. E-mail: ccaiafa@gmail.com.*
- *D.P. Mandic is with the Communication and Signal Processing Research Group, Department of Electrical and Electronic Engineering, Imperial College, London, United Kingdom. E-mail: d.mandic@imperial.ac.uk.*
- *Z.C. Chao, Y. Nagasaka, and N. Fujii are with the Laboratory for Adaptive Intelligence, Brain Science Institute, RIKEN, Saitama, Japan. E-mail: zenas.c.chao@gmail.com, {nyasuo, na}@brain.riken.jp.*
- *L. Zhang is with the MOE-Microsoft Key Laboratory for Intelligent Computing and Intelligent Systems, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. E-mail: zhang-lq@cs.sjtu.edu.cn.*
- *A. Cichocki is with the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Saitama, Japan and Systems Research Institute in Polish Academy of Science, Warsaw, Poland. E-mail: a.cichocki@riken.jp.*

Manuscript received 11 Apr. 2012; revised 29 Aug. 2012; accepted 12 Nov. 2012; published online 28 Nov. 2012.
Recommended for acceptance by A. Gretton.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2012-04-0274.
Digital Object Identifier no. 10.1109/TPAMI.2012.254.

covariates regression was studied in [26], [27], [28], where the weight vectors yielding the latent variables are optimized by the same strategy as in N-PLS, resulting in better fitness to independent data $\underline{\mathbf{X}}$ while maintaining no difference in predictive performance. The tensor decomposition used within N-PLS is canonical decomposition/parallel factor analysis (CANDECOMP/PARAFAC or CP) [29], which makes N-PLS inherit both the advantages and limitations of CP [30]. These limitations are related to poor fitness ability, computational complexity, and slow convergence when handling multivariate dependent data and higher order ($N > 3$) independent data, causing N-PLS to not be guaranteed to outperform standard PLS [23], [31].

In this paper, we propose a new generalized mutilinear regression model, called higher order partial least squares (HOPLS), which makes it possible to predict an Mth-order tensor $\underline{\mathbf{Y}}$ ($M \geq 3$) (or a particular case of two-way matrix $\mathbf{Y}$) from an Nth-order tensor $\underline{\mathbf{X}}$ ($N \geq 3$) by projecting tensor $\underline{\mathbf{X}}$ onto a low-dimensional common latent subspace. The latent subspaces are optimized sequentially through simultaneous rank-$(1, L_2, \ldots, L_N)$ approximation of $\underline{\mathbf{X}}$ and rank-$(1, K_2, \ldots, K_M)$ approximation of $\underline{\mathbf{Y}}$ (or rank-one approximation in particular case of two-way matrix $\mathbf{Y}$). Due to the better fitness ability of the orthogonal Tucker model as compared to CP [16] and the flexibility of the block Tucker model [32], the analysis and simulations show that HOPLS proves to be a promising multilinear subspace regression framework that provides not only an optimal tradeoff between fitness and model complexity but also enhanced predictive ability in general. In addition, we develop a new strategy to find a closed-form solution by employing higher order singular value decomposition (HOSVD) [33], which makes the computation more efficient than the classical iterative procedure.

The paper is structured as follows: In Section 2, an overview of two-way PLS is presented, and the notation and notions related to multiway data analysis are introduced. In Section 3, the new multilinear regression model is proposed, together with the corresponding solutions and algorithms. Extensive simulations on synthetic data and a real-world case study on the fusion of behavioral and neural data are presented in Section 4, followed by conclusions in Section 5.

2 Background and Notation

2.1 Notation and Definitions

Nth-order tensors (*multiway arrays*) are denoted by underlined boldface capital letters, matrices (*two-way arrays)* by boldface capital letters, and vectors by boldface lower case letters. The ith entry of a vector $\mathbf{x}$ is denoted by x_i, element (i, j) of a matrix $\mathbf{X}$ is denoted by x_{ij}, and element $(i_1, i_2, \ldots, i_N)$ of an Nth-order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ by $x_{i_1 i_2 \ldots i_N}$ or $(\underline{\mathbf{X}})_{i_1 i_2 \ldots i_N}$. Indices typically range from 1 to their capital version, for example, $i_N = 1, \ldots, I_N$. The mode-n matricization of a tensor is denoted by $\mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times I_1 \cdots I_{n-1} I_{n+1} \cdots I_N}$. The nth factor matrix in a sequence is denoted by $\mathbf{A}^{(n)}$.

The *n-mode product* of a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times \cdots \times I_N}$ and matrix $\mathbf{A} \in \mathbb{R}^{J_n \times I_n}$ is denoted by $\underline{\mathbf{Y}} = \mathbf{X} \times_n \mathbf{A} \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times J_n \times I_{n+1} \times \cdots \times I_N}$ and is defined as

$$y_{i_1 i_2 \ldots i_{n-1} j_n i_{n+1} \ldots i_N} = \sum_{i_n} x_{i_1 i_2 \ldots i_n \ldots i_N} a_{j_n i_n}. \tag{1}$$

The *rank-*$(R_1, R_2, \ldots, R_N)$ *Tucker model* [34] is a tensor decomposition defined and denoted as follows:

$$\begin{aligned} \underline{\mathbf{Y}} &\approx \underline{\mathbf{G}} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \times_3 \cdots \times_N \mathbf{A}^{(N)} \\ &= [\![\underline{\mathbf{G}}; \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!], \end{aligned} \tag{2}$$

where $\underline{\mathbf{G}} \in \mathbb{R}^{R_1 \times R_2 \times \ldots \times R_N}$ ($R_n \leq I_n$) is the *core tensor* and $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R_n}$ are the *factor matrices*. The last term is the simplified notation, introduced in [35] for the Tucker operator. When the factor matrices are orthonormal and the core tensor is all orthogonal, this model is called HOSVD [33], [35].

The *CP model* [16], [29], [36], [37], [38] became prominent in chemistry [28] and is defined as a sum of rank-one tensors:

$$\underline{\mathbf{Y}} \approx \sum_{r=1}^{R} \lambda_r \, \mathbf{a}_r^{(1)} \circ \mathbf{a}_r^{(2)} \circ \cdots \circ \mathbf{a}_r^{(N)}, \tag{3}$$

where the symbol "$\circ$" denotes the outer product of vectors, $\mathbf{a}_r^{(n)}$ is the column-r vector of matrix $\mathbf{A}^{(n)}$, and λ_r are scalars. The CP model can also be represented by (2), under the condition that the core tensor is superdiagonal, i.e., $R_1 = \cdots = R_N$ and $g_{i_1 i_2, \ldots, i_N} = 0$ if $i_n \neq i_m$ for all $n \neq m$.

The 1-mode product between $\underline{\mathbf{G}} \in \mathbb{R}^{1 \times I_2 \times \cdots \times I_N}$ and $\mathbf{t} \in \mathbb{R}^{I_1 \times 1}$ is of size $I_1 \times I_2 \times \cdots \times I_N$, and is defined as

$$(\underline{\mathbf{G}} \times_1 \mathbf{t})_{i_1 i_2 \ldots i_N} = g_{1 i_2 \ldots i_N} t_{i_1}. \tag{4}$$

The *inner product of two tensors* $\underline{\mathbf{A}}, \underline{\mathbf{B}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ is defined by $\langle \underline{\mathbf{A}}, \underline{\mathbf{B}} \rangle = \sum_{i_1 i_2 \ldots i_N} a_{i_1 i_2 \ldots i_N} b_{i_1 i_2 \ldots i_N}$, and the squared Frobenius norm by $\|\underline{\mathbf{A}}\|_F^2 = \langle \underline{\mathbf{A}}, \underline{\mathbf{A}} \rangle$.

The *n-mode cross covariance* between an Nth-order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_n \times \cdots \times I_N}$ and an Mth-order tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{J_1 \times \cdots \times I_n \times \cdots \times J_M}$ with the same size I_n on the nth mode, denoted by

$$\mathrm{COV}_{\{n;n\}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) \in \mathbb{R}^{I_1 \times \cdots \times I_{n-1} \times I_{n+1} \times \cdots \times I_N \times J_1 \times \cdots \times J_{n-1} \times J_{n+1} \times \cdots \times J_M},$$

is defined as

$$\underline{\mathbf{C}} = \mathrm{COV}_{\{n;n\}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) = <\underline{\mathbf{X}}, \underline{\mathbf{Y}}>_{\{n;n\}}, \tag{5}$$

where the symbol $< \bullet, \bullet >_{\{n;n\}}$ represents an n-mode multiplication between two tensors, and is defined as

$$\begin{aligned} &c_{i_1, \ldots, i_{n-1}, i_{n+1} \ldots i_N, j_1, \ldots, j_{n-1} j_{n+1} \ldots j_M} \\ &\quad = \sum_{i_n=1}^{I_n} x_{i_1, \ldots, i_n, \ldots, i_N} y_{j_1, \ldots, i_n, \ldots, j_M}. \end{aligned}$$

As a special case, for a matrix $\mathbf{Y} \in \mathbb{R}^{I_n \times M}$, the *n-mode cross covariance* between $\underline{\mathbf{X}}$ and $\mathbf{Y}$ simplifies as

$$\mathrm{COV}_{\{n;1\}}(\underline{\mathbf{X}}, \mathbf{Y}) = \underline{\mathbf{X}} \times_n \mathbf{Y}^{\mathrm{T}}, \tag{7}$$

under the assumption that n-mode column vectors of $\underline{\mathbf{X}}$ and columns of $\mathbf{Y}$ are mean centered.

2.2 Standard PLS (Two-Way PLS)

The PLSR was originally developed for econometrics by Wold [3], [39] in order to deal with collinear predictor variables. The usefulness of PLS in chemical applications was

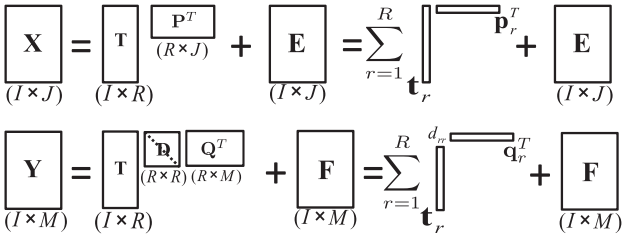


Fig. 1. The PLS model: Data decomposition as a sum of rank-one matrices.

illuminated by the group of Wold et al. [40], [41], after some initial work by Kowalski et al. [42]. Currently, the PLSR is being widely applied in chemometrics, sensory evaluation, industrial process control, and, more recently, in the analysis of functional brain imaging data [43], [44], [45], [46], [47].

The principle behind PLS is to search for a set of latent vectors by performing a simultaneous decomposition of $\mathbf{X} \in \mathbb{R}^{I \times J}$ and $\mathbf{Y} \in \mathbb{R}^{I \times M}$ with the constraint that these components explain as much as possible of the covariance between $\mathbf{X}$ and $\mathbf{Y}$. It can be formulated as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} = \sum_{r=1}^{R} \mathbf{t}_r \mathbf{p}_r^T + \mathbf{E}, \tag{8}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} = \sum_{r=1}^{R} \mathbf{u}_r \mathbf{q}_r^T + \mathbf{F}, \tag{9}$$

where $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_R] \in \mathbb{R}^{I \times R}$ consists of R orthonormal latent variables from $\mathbf{X}$, i.e., $\mathbf{T}^T\mathbf{T} = \mathbf{I}$, and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_R] \in \mathbb{R}^{I \times R}$ are latent variables from $\mathbf{Y}$ having maximum covariance with $\mathbf{T}$ columnwise. The matrices $\mathbf{P}$ and $\mathbf{Q}$ represent loadings and $\mathbf{E}$, $\mathbf{F}$ are, respectively, the residuals for $\mathbf{X}$ and $\mathbf{Y}$. To find the first set of components, the classical PLS algorithm is to optimize the two sets of weights $\mathbf{w}, \mathbf{q}$ so as to satisfy

$$\max_{\{\mathbf{w},\mathbf{q}\}} [\mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{q}]^2, \quad \text{s. t.} \quad \mathbf{w}^T\mathbf{w} = 1, \mathbf{q}^T\mathbf{q} = 1. \tag{10}$$

The latent variables are then given by $\mathbf{t} = \mathbf{X}\mathbf{w}/\|\mathbf{X}\mathbf{w}\|$ and $\mathbf{u} = \mathbf{Y}\mathbf{q}$. Here, two assumptions are made: 1) The latent variables $\{\mathbf{t}_r\}_{r=1}^R$ are good predictors of $\mathbf{Y}$; 2) a linear inner relation between the latent variables $\mathbf{t}$ and $\mathbf{u}$ exists, i.e., $\mathbf{U} = \mathbf{T}\mathbf{D} + \mathbf{Z}$, where $\mathbf{D}$ is a linear relation diagonal matrix and $\mathbf{Z}$ denotes the matrix of Gaussian i.i.d. residuals. The maximum likelihood estimate of parameters $\mathbf{D}$ is $\mathbf{d}_{rr} = (\mathbf{t}_r^T\mathbf{t}_r)^{-1}\mathbf{t}_r^T\mathbf{u}_r$. Upon combining it with the decomposition of $\mathbf{Y}$, (9) can be written as

$$\mathbf{Y} = \mathbf{T}\mathbf{D}\mathbf{Q}^T + (\mathbf{Z}\mathbf{Q}^T + \mathbf{F}) = \mathbf{T}\mathbf{D}\mathbf{Q}^T + \mathbf{F}^*, \tag{11}$$

where $\mathbf{F}^*$ is the residual matrix. Thus, (11) indicates that the problem boils down to finding common latent variables $\mathbf{T}$ that explain the variance of both $\mathbf{X}$ and $\mathbf{Y}$, as illustrated in Fig. 1.

3 HOPLS

In the case of a two-way matrix, column rank and row rank are equal and correspond to the minimal number of rank-one

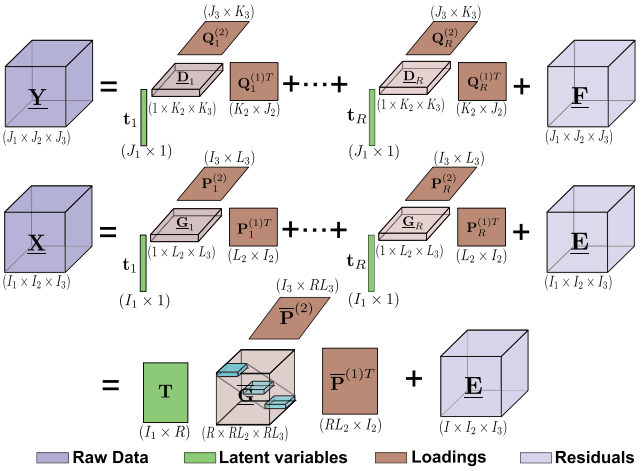


Fig. 2. Schematic diagram of the HOPLS model: approximating $\underline{\mathbf{X}}$ as a sum of rank-$(1, L_2, L_3)$ tensors. Approximation for $\underline{\mathbf{Y}}$ follows a similar principle with shared common latent components $\mathbf{T}$.

terms into which the matrix can be decomposed, and thus the low-rank approximation is equivalent to subspace approximation. However, for a higher order tensor, these criteria lead to two typical tensor decompositions (i.e., CP and Tucker model). The CP decomposes a tensor as a sum of rank-one tensors defined in (3), which can be considered as a low-rank approximation, while the Tucker model decomposes a tensor into a core tensor multiplied by matrices along each mode (2), which is a rank-(R_1, R_2, R_3) subspace approximation. Both the CP and Tucker decompositions can be regarded as higher order generalizations of the matrix singular value decomposition (SVD) and principal component analysis. In fact, CP can be viewed as a special case of Tucker where the core tensor is superdiagonal. Block term decompositions unify the Tucker and CP decompositions into one framework by decomposing a tensor into a sum of rank-(R_1, R_2, R_3) terms [32].

Consider an Nth-order independent tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and an Mth-order dependent tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{J_1 \times \cdots \times J_M}$, having the same size on the first mode, i.e., $I_1 = J_1$. Our objective is to find an optimal subspace approximation of $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, in which the latent vectors from $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ have maximum pairwise covariance. Considering a linear relation between the latent vectors, the problem boils down to finding the common latent subspace which can approximate both $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ simultaneously. We first address the general case of a tensor $\underline{\mathbf{X}}(N \geq 3)$ and a tensor $\underline{\mathbf{Y}}(M \geq 3)$. A particular case with a tensor $\underline{\mathbf{X}}(N \geq 3)$ and a matrix $\mathbf{Y}(M = 2)$ is presented separately in Section 3.3, using a slightly different approach.

3.1 Proposed Model

Applying Tucker decomposition within a PLS framework is not straightforward, and to that end we propose a novel block-wise orthogonal Tucker approach to model the data. More specifically, we assume $\underline{\mathbf{X}}$ is decomposed as a sum of rank-$(1, L_2, \ldots, L_N)$ Tucker blocks, while $\underline{\mathbf{Y}}$ is decomposed as a sum of rank-$(1, K_2, \ldots, K_M)$ Tucker blocks (see Fig. 2), which can be expressed as

$$\begin{aligned}\underline{\mathbf{X}} &= \sum_{r=1}^{R} \underline{\mathbf{G}}_r \times_1 \mathbf{t}_r \times_2 \mathbf{P}_r^{(1)} \times_3 \cdots \times_N \mathbf{P}_r^{(N-1)} + \underline{\mathbf{E}}_R, \\ \underline{\mathbf{Y}} &= \sum_{r=1}^{R} \underline{\mathbf{D}}_r \times_1 \mathbf{t}_r \times_2 \mathbf{Q}_r^{(1)} \times_3 \cdots \times_M \mathbf{Q}_r^{(M-1)} + \underline{\mathbf{F}}_R,\end{aligned} \tag{12}$$

where R is the number of latent vectors, $\mathbf{t}_r \in \mathbb{R}^{I_1}$ is the rth latent vector, $\{\mathbf{P}_r^{(n)}\}_{n=1}^{N-1} \in \mathbb{R}^{I_{n+1} \times L_{n+1}}$ and $\{\mathbf{Q}_r^{(m)}\}_{m=1}^{M-1} \in \mathbb{R}^{J_{m+1} \times K_{m+1}}$ are loading matrices on mode-n and mode-m respectively, and $\underline{\mathbf{G}}_r \in \mathbb{R}^{1 \times L_2 \times \cdots \times L_N}$ and $\underline{\mathbf{D}}_r \in \mathbb{R}^{1 \times K_2 \times \cdots \times K_M}$ are core tensors.

However, the Tucker decompositions in (12) are not unique [16] due to the permutation, rotation, and scaling issues. To alleviate this problem, additional constraints should be imposed such that the core tensors $\boldsymbol{G}_r$ and $\boldsymbol{D}_r$ are all-orthogonal, a sequence of loading matrices are column-wise orthonormal, i.e., $\mathbf{P}_r^{(n)T}\mathbf{P}_r^{(n)} = \mathbf{I}$ and $\mathbf{Q}_r^{(m)T}\mathbf{Q}_r^{(m)} = \mathbf{I}$, the latent vector is of length one, i.e., $\|\mathbf{t}_r\|_F = 1$. Thus, each term in (12) is represented as an orthogonal Tucker model, implying essentially uniqueness as it is subject only to trivial indeterminacies [32].

By defining a latent matrix $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_R]$, mode-$n$ loading matrix $\overline{\mathbf{P}}^{(n)} = [\mathbf{P}_1^{(n)}, \ldots, \mathbf{P}_R^{(n)}]$, mode-$m$ loading matrix $\overline{\mathbf{Q}}^{(m)} = [\mathbf{Q}_1^{(m)}, \ldots, \mathbf{Q}_R^{(m)}]$, and core tensor

$$\begin{aligned}\overline{\underline{\mathbf{G}}} &= \text{blockdiag}(\underline{\mathbf{G}}_1, \ldots, \underline{\mathbf{G}}_R) \in \mathbb{R}^{R \times RL_2 \times \cdots \times RL_N}, \\ \overline{\underline{\mathbf{D}}} &= \text{blockdiag}(\underline{\mathbf{D}}_1, \ldots, \underline{\mathbf{D}}_R) \in \mathbb{R}^{R \times RK_2 \times \cdots \times RK_M},\end{aligned}$$

the HOPLS model in (12) can be rewritten as

$$\begin{aligned}\underline{\mathbf{X}} &= \overline{\underline{\mathbf{G}}} \times_1 \mathbf{T} \times_2 \overline{\mathbf{P}}^{(1)} \times_3 \cdots \times_N \overline{\mathbf{P}}^{(N-1)} + \underline{\mathbf{E}}_R, \\ \underline{\mathbf{Y}} &= \overline{\underline{\mathbf{D}}} \times_1 \mathbf{T} \times_2 \overline{\mathbf{Q}}^{(1)} \times_3 \cdots \times_M \overline{\mathbf{Q}}^{(M-1)} + \underline{\mathbf{F}}_R,\end{aligned} \tag{13}$$

where $\underline{\mathbf{E}}_R$ and $\underline{\mathbf{F}}_R$ are residuals after extracting R components. The core tensors $\overline{\underline{\mathbf{G}}}$ and $\overline{\underline{\mathbf{D}}}$ have a special block-diagonal structure (see Fig. 2) and their elements indicate the level of local interactions between the corresponding latent vectors and loading matrices. Note that the tensor decomposition in (13) is similar to the block term decomposition discussed in [32], which aims at the decomposition of only one tensor. However, HOPLS attempts to find the block Tucker decompositions of two tensors with blockwise orthogonal constraints, which at the same time satisfies a certain criteria of sharing the common latent components on a specific mode.

Benefiting from the advantages of Tucker decomposition over the CP model [16], HOPLS promises to approximate data better than N-PLS. Specifically, HOPLS differs substantially from the N-PLS model in the sense that extraction of latent components in HOPLS is based on subspace approximation rather than on low-rank approximation and the size of loading matrices is controlled by a hyperparameter, providing a tradeoff between fitness and model complexity. Note that HOPLS simplifies into N-PLS if we define $\forall n : \{L_n\} = 1$ and $\forall m : \{K_m\} = 1$.

3.2 Optimization Criteria and Algorithm

There are two different approaches for extracting the latent components: sequential and simultaneous methods. A sequential method extracts one latent component at a time, deflates the proper tensors, and calculates the next component from the residuals. In a simultaneous method, all components are calculated simultaneously by minimizing a certain criterion. In the following, we employ a sequential method.

The tensor decompositions in (12) can be represented as an optimization problem of approximating $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ by orthogonal Tucker model while having a common latent component on a specific mode. If we apply HOSVD individually on $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$, the best rank-$(1, L_2, \ldots, L_N)$ approximation for $\underline{\mathbf{X}}$ and the best rank-$(1, K_2, \ldots, K_M)$ approximation for $\underline{\mathbf{Y}}$ can be obtained, while the common latent vector $\mathbf{t}_r$ cannot be ensured. Another way is to find the best approximation of $\underline{\mathbf{X}}$ by HOSVD first, subsequently, $\underline{\mathbf{Y}}$ can be approximated by a fixed $\mathbf{t}_r$. However, this procedure, which resembles multiway PCR [28], has the drawback that the common latent components are not necessarily predictive for $\underline{\mathbf{Y}}$.

The optimization of subspace transformation according to (12) will be formulated as a problem of determining a set of orthogonormal loadings $\mathbf{P}_r^{(n)}, \mathbf{Q}_r^{(m)}$, $r = 1, 2, \ldots, R$, and latent vectors $\mathbf{t}_r$ that satisfies a certain criterion. Since each term can be optimized sequentially with the same criteria based on deflation, in the following we shall simplify the problem to that of finding the first latent vector $\mathbf{t}$ and two sequences of loading matrices $\mathbf{P}^{(n)}$ and $\mathbf{Q}^{(m)}$.

To develop a strategy for the simultaneous minimization of the Frobenius norm of residuals $\mathbf{E}$ and $\mathbf{F}$ while keeping a common latent vector $\mathbf{t}$, we first need to introduce the following basic results.

Proposition 3.1. *Given a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ and column orthonormal matrices $\mathbf{P}^{(n)} \in \mathbb{R}^{I_{n+1} \times L_{n+1}}$, $n = 1, \ldots, N-1$, $\mathbf{t} \in \mathbb{R}^{I_1}$ with $\|\mathbf{t}\|_F = 1$, the least squares (LS) solution to $\min_{\underline{\mathbf{G}}} \|\underline{\mathbf{X}} - \underline{\mathbf{G}} \times_1 \mathbf{t} \times_2 \mathbf{P}^{(1)} \times_3 \cdots \times_N \mathbf{P}^{(N-1)}\|_F^2$ is given by $\underline{\mathbf{G}} = \mathbf{X} \times_1 \mathbf{t}^T \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T}$.*

Proof. This result is very well known and is widely used in the literature [16], [33]. A simple proof is based on writing the mode-1 matricization of tensor $\underline{\mathbf{X}}$ as

$$\mathbf{X}_{(1)} = \mathbf{t}\mathbf{G}_{(1)}(\mathbf{P}^{(N-1)} \otimes \cdots \otimes \mathbf{P}^{(1)})^T + \mathbf{E}_{(1)}, \tag{14}$$

where tensor $\underline{\mathbf{E}}_{(1)}$ is the residual and the symbol "$\otimes$" denotes the *Kronecker product*. Since $\mathbf{t}^T\mathbf{t} = 1$ and $(\mathbf{P}^{(N-1)} \otimes \cdots \otimes \mathbf{P}^{(1)})$ is column orthonormal, the LS solution of $\mathbf{G}_{(1)}$ with fixed matrices $\mathbf{t}$ and $\mathbf{P}^{(n)}$ is given by $\mathbf{G}_{(1)} = \mathbf{t}^T\mathbf{X}_{(1)}(\mathbf{P}^{(N-1)} \otimes \cdots \otimes \mathbf{P}^{(1)})$; writing it in a tensor form we obtain the desired result. □

Proposition 3.2. *Given a tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$, the following two constrained optimization problems are equivalent:*

1. $\min_{\{\mathbf{P}^{(n)}, \mathbf{t}, \underline{\mathbf{G}}\}} \|\underline{\mathbf{X}} - \underline{\mathbf{G}} \times_1 \mathbf{t} \times_2 \mathbf{P}^{(1)} \times_3 \cdots \times_N \mathbf{P}^{(N-1)}\|_F^2$, *s.t. matrices $\mathbf{P}^{(n)}$ are column orthonormal and $\|\mathbf{t}\|_F = 1$.*
2. $\max_{\{\mathbf{P}^{(n)}, \mathbf{t}\}} \|\underline{\mathbf{X}} \times_1 \mathbf{t}^T \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T}\|_F^2$, *s.t. matrices $\mathbf{P}^{(n)}$ are column orthonormal and $\|\mathbf{t}\|_F = 1$.*

The proof is available in [16, pp. 477-478].

Assume that the orthonormal matrices $\mathbf{P}^{(n)}, \mathbf{Q}^{(m)}, \mathbf{t}$ are given, then from Proposition 3.1, the core tensors in (12) can be computed as

$$\begin{aligned}\underline{\mathbf{G}} &= \mathbf{X} \times_1 \mathbf{t}^T \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T},\\ \underline{\mathbf{D}} &= \mathbf{Y} \times_1 \mathbf{t}^T \times_2 \mathbf{Q}^{(1)T} \times_3 \cdots \times_M \mathbf{Q}^{(M-1)T}.\end{aligned} \tag{15}$$

According to Proposition 3.2, minimization of $\|\underline{\mathbf{E}}\|_F$ and $\|\underline{\mathbf{F}}\|_F$ under the orthonormality constraint is equivalent to maximization of $\|\underline{\mathbf{G}}\|_F$ and $\|\underline{\mathbf{D}}\|_F$.

However, there is no straightforward tensor decomposition method to maximize $\|\underline{\mathbf{G}}\|_F$ and $\|\underline{\mathbf{D}}\|_F$ simultaneously over the factor matrices $\{\mathbf{P}^{(n)}\}_{n=1}^{N-1}$, $\{\mathbf{Q}^{(m)}\}_{m=1}^{M-1}$ and a common latent vector $\mathbf{t}$. To this end, we propose to maximize a product of norms of two core tensors, i.e., $\max\{\|\underline{\mathbf{G}}\|_F^2 \cdot \|\underline{\mathbf{D}}\|_F^2\}$. Since the latent vector $\mathbf{t}$ is determined by $\mathbf{P}^{(n)}, \mathbf{Q}^{(m)}$, the first step is to optimize the orthonormal loadings, then the common latent vectors can be computed by the fixed loadings.

Proposition 3.3. *Let* $\underline{\mathbf{G}} \in \mathbb{R}^{1\times L_2\times\cdots\times L_N}$ *and* $\underline{\mathbf{D}} \in \mathbb{R}^{1\times K_2\times\cdots\times K_M}$, *then* $\|<\underline{\mathbf{G}}, \underline{\mathbf{D}}>_{\{1;1\}}\|_F^2 = \|\underline{\mathbf{G}}\|_F^2 \cdot \|\underline{\mathbf{D}}\|_F^2$.

Proof.

$$\begin{aligned}\|<\underline{\mathbf{G}},\underline{\mathbf{D}}>_{\{1;1\}}\|_F^2 &= \|\mathrm{vec}(\underline{\mathbf{G}})\mathrm{vec}^T(\underline{\mathbf{D}})\|_F^2\\ &= \mathrm{trace}(\mathrm{vec}(\underline{\mathbf{D}})\mathrm{vec}^T(\underline{\mathbf{G}})\mathrm{vec}(\underline{\mathbf{G}})\mathrm{vec}^T(\underline{\mathbf{D}}))\\ &= \|\mathrm{vec}(\underline{\mathbf{G}})\|_F^2 \cdot \|\mathrm{vec}(\underline{\mathbf{D}})\|_F^2,\end{aligned} \tag{16}$$

where $\mathrm{vec}(\underline{\mathbf{G}}) \in \mathbb{R}^{L_2L_3\ldots L_N}$ is the vectorization of the tensor $\underline{\mathbf{G}}$. □

From Proposition 3.3, observe that to maximize $\|\underline{\mathbf{G}}\|_F^2 \cdot \|\underline{\mathbf{D}}\|_F^2$ is equivalent to maximizing $\|<\underline{\mathbf{G}}, \underline{\mathbf{D}}>_{\{1;1\}}\|_F^2$. According to (15) and $\mathbf{t}^T\mathbf{t} = 1$, $\|<\underline{\mathbf{G}}, \underline{\mathbf{D}}>_{\{1;1\}}\|_F^2$ can be expressed as

$$\|[<\underline{\mathbf{X}}, \underline{\mathbf{Y}}>_{\{1;1\}}; \mathbf{P}^{(1)T}, \ldots, \mathbf{P}^{(N-1)T}, \mathbf{Q}^{(1)T}, \ldots, \mathbf{Q}^{(M-1)T}]\|_F^2. \tag{17}$$

Note that this form is quite similar to the optimization problem for two-way PLS in (10), where the cross-covariance matrix $\mathbf{X}^T\mathbf{Y}$ is replaced by $<\underline{\mathbf{X}}, \underline{\mathbf{Y}}>_{\{1;1\}}$. In addition, the optimization item becomes the norm of a small tensor in contrast to a scalar in (10). Thus, if we define $<\underline{\mathbf{X}}, \underline{\mathbf{Y}}>_{\{1;1\}}$ as a mode-1 cross-covariance tensor $\underline{\mathbf{C}} = \mathrm{COV}_{\{1;1\}}(\underline{\mathbf{X}}, \underline{\mathbf{Y}}) \in \mathbb{R}^{I_2\times\cdots\times I_N\times J_2\times\cdots\times J_M}$, the optimization problem can be finally formulated as

$$\begin{aligned}&\max_{\{\mathbf{P}^{(n)},\mathbf{Q}^{(m)}\}} \|[\![\underline{\mathbf{C}}; \mathbf{P}^{(1)T}, \ldots, \mathbf{P}^{(N-1)T}, \mathbf{Q}^{(1)T}, \ldots, \mathbf{Q}^{(M-1)T}]\!]\|_F^2\\ &\text{s.t.} \quad \mathbf{P}^{(n)T}\mathbf{P}^{(n)} = \mathbf{I}_{L_{n+1}}, \mathbf{Q}^{(m)T}\mathbf{Q}^{(m)} = \mathbf{I}_{K_{m+1}},\end{aligned} \tag{18}$$

where $\mathbf{P}^{(n)}$, $n = 1, \ldots, N-1$, and $\mathbf{Q}^{(m)}$, $m = 1, \ldots, M-1$, are the parameters to optimize.

Based on Proposition 3.2 and orthogonality of $\mathbf{P}^{(n)}, \mathbf{Q}^{(m)}$, the optimization problem in (18) is equivalent to finding the best subspace approximation of $\underline{\mathbf{C}}$ as

$$\underline{\mathbf{C}} \approx [\![\underline{\mathbf{G}}^{(C)}; \mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(N-1)}, \mathbf{Q}^{(1)}, \ldots, \mathbf{Q}^{(M-1)}]\!], \tag{19}$$

which can be obtained by rank-$(L_2, \ldots, L_N, K_2, \ldots, K_M)$ HOSVD on tensor $\underline{\mathbf{C}}$. Based on Proposition 3.1, the optimization term in (18) is equivalent to the norm of core tensor $\underline{\mathbf{G}}^{(C)}$. To achieve this goal, the higher order orthogonal iteration (HOOI) algorithm [16], [37], which is known to converge fast, is employed to find the parameters $\mathbf{P}^{(n)}$ and $\mathbf{Q}^{(m)}$ by orthogonal Tucker decomposition of $\underline{\mathbf{C}}$.

Subsequently, based on the estimate of the loadings $\mathbf{P}^{(n)}$ and $\mathbf{Q}^{(m)}$, we can now compute the common latent vector $\mathbf{t}$. Since our goal is to predict $\underline{\mathbf{Y}}$ from $\underline{\mathbf{X}}$, similarly to the PLS method, we need to estimate $\mathbf{t}$ from predictors $\underline{\mathbf{X}}$ and also to estimate the regression coefficient $\underline{\mathbf{D}}$ in order to predict responses $\underline{\mathbf{Y}}$. For a given set of loading matrices $\{\mathbf{P}^{(n)}\}$, the latent vector $\mathbf{t}$ should explain variance of $\underline{\mathbf{X}}$ as much as possible, that is,

$$\mathbf{t} = \arg\min_{\mathbf{t}} \left\|\underline{\mathbf{X}} - [\![\underline{\mathbf{G}}; \mathbf{t}, \mathbf{P}^{(1)}, \ldots, \mathbf{P}^{(N-1)}]\!]\right\|_F^2, \tag{20}$$

which can be easily achieved by choosing $\mathbf{t}$ as the first leading left singular vector of the matrix $(\underline{\mathbf{X}} \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T})_{(1)}$ as used in the HOOI algorithm (see [16], [35]). Thus, the core tensors $\underline{\mathbf{G}}$ and $\underline{\mathbf{D}}$ are computed by (15).

The above procedure should be carried out repeatedly using the deflation operation until an appropriate number of components (i.e., R) are obtained or the norms of residuals are smaller than a certain threshold. The deflation[1] is performed by subtracting from $\underline{\mathbf{X}}$ and $\underline{\mathbf{Y}}$ the information explained by a rank-$(1, L_2, \ldots, L_N)$ tensor $\widehat{\underline{\mathbf{X}}}$ and a rank-$(1, K_2, \ldots, K_M)$ tensor $\widehat{\underline{\mathbf{Y}}}$, respectively. The HOPLS algorithm is outlined in Algorithm 1.

Algorithm 1. The Higher-order Partial Least Squares (HOPLS) Algorithm for a Tensor $\underline{\mathbf{X}}$ and a Tensor $\underline{\mathbf{Y}}$

Input: $\underline{\mathbf{X}} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, $\underline{\mathbf{Y}} \in \mathbb{R}^{J_1\times\cdots\times J_M}$, $N \geq 3, M \geq 3$ and $I_1 = J_1$.
Number of latent vectors is R and number of loading vectors are $\{L_n\}_{n=2}^N$ and $\{K_m\}_{m=2}^M$.

Output: $\{\mathbf{P}_r^{(n)}\}; \{\mathbf{Q}_r^{(m)}\}; \{\underline{\mathbf{G}}_r\}; \{\underline{\mathbf{D}}_r\}; \mathbf{T}$
$r = 1, \ldots, R;\ n = 1, \ldots, N-1;\ m = 1, \ldots, M-1.$
Initialization: $\underline{\mathbf{E}}_1 \leftarrow \underline{\mathbf{X}}, \quad \underline{\mathbf{F}}_1 \leftarrow \underline{\mathbf{Y}}.$
for $r = 1$ **to** R **do**
 if $\|\underline{\mathbf{E}}_r\|_F > \varepsilon$ **and** $\|\underline{\mathbf{F}}_r\|_F > \varepsilon$ **then**
  $\underline{\mathbf{C}}_r \leftarrow <\underline{\mathbf{E}}_r, \underline{\mathbf{F}}_r>_{\{1,1\}}$;
  Rank-$(L_2, \ldots, L_N, K_2, \ldots, K_M)$ orthogonal Tucker decomposition of $\underline{\mathbf{C}}_r$ by HOOI [16] as
  $\underline{\mathbf{C}}_r \approx [\![\underline{\mathbf{G}}_r^{(C_r)}; \mathbf{P}_r^{(1)}, \ldots, \mathbf{P}_r^{(N-1)}, \mathbf{Q}_r^{(1)}, \ldots, \mathbf{Q}_r^{(M-1)}]\!]$;
  $\mathbf{t}_r \leftarrow$ the first leading left singular vector by
   $\mathrm{SVD}\big[(\underline{\mathbf{E}}_r \times_2 \mathbf{P}_r^{(1)T} \times_3 \cdots \times_N \mathbf{P}_r^{(N-1)T})_{(1)}\big]$;
  $\underline{\mathbf{G}}_r \leftarrow [\![\underline{\mathbf{E}}_r; \mathbf{t}_r^T, \mathbf{P}_r^{(1)T}, \ldots, \mathbf{P}_r^{(N-1)T}]\!]$;
  $\underline{\mathbf{D}}_r \leftarrow [\![\underline{\mathbf{F}}_r; \mathbf{t}_r^T, \mathbf{Q}_r^{(1)T}, \ldots, \mathbf{Q}_r^{(M-1)T}]\!]$;
  Deflation:
  $\underline{\mathbf{E}}_{r+1} \leftarrow \underline{\mathbf{E}}_r - [\![\underline{\mathbf{G}}_r; \mathbf{t}_r, \mathbf{P}_r^{(1)}, \ldots, \mathbf{P}_r^{(N-1)}]\!]$;
  $\underline{\mathbf{F}}_{r+1} \leftarrow \underline{\mathbf{F}}_r - [\![\underline{\mathbf{D}}_r; \mathbf{t}_r, \mathbf{Q}_r^{(1)}, \ldots, \mathbf{Q}_r^{(M-1)}]\!]$;
 else
  Break;
 end if
end for

1. Note that the latent vectors are not orthogonal in the HOPLS algorithm, which is related to deflation. The theoretical explanation and proof are given in the supplemental material, which can be found in the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2012.254.

3.3 The Case of the Tensor $\underline{\mathbf{X}}$ and Matrix $\mathbf{Y}$

Consider an Nth-order independent tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times \cdots \times I_N}$ ($N \geq 3$) and a two-way dependent data $\mathbf{Y} \in \mathbb{R}^{I_1 \times M}$, with the same sample size I_1. The HOPLS models independent data $\underline{\mathbf{X}}$ as a sum of rank-$(1, L_2, \ldots, L_N)$ tensors while dependent data $\mathbf{Y}$ is represented by a sum of rank-one matrices as

$$\begin{aligned} \underline{\mathbf{X}} &= \sum_{r=1}^{R} \underline{\mathbf{G}}_r \times_1 \mathbf{t}_r \times_2 \mathbf{P}_r^{(1)} \times_3 \cdots \times_N \mathbf{P}_r^{(N-1)} + \underline{\mathbf{E}}_R, \\ \mathbf{Y} &= \sum_{r=1}^{R} \mathbf{u}_r \mathbf{q}_r^T + \mathbf{F}_R, \end{aligned} \tag{21}$$

where $\|\mathbf{q}_r\| = 1$ and $\|\mathbf{t}_r\| = 1$, $\mathbf{P}_r^{(n)T}\mathbf{P}_r^{(n)} = \mathbf{I}, \forall n$. Similarly to the standard PLS method, the assumption employed was that of a linear "inner" relationship between the latent variables $\mathbf{t}_r$ and $\mathbf{u}_r$, that is, $\mathbf{U} = \mathbf{TD} + \mathbf{Z}$, where $\mathbf{D}$ is the diagonal matrix, and $\mathbf{Z}$ denotes the Gaussian residuals. A combination with (21) yields the decomposition of $\mathbf{Y}$ in the form

$$\mathbf{Y} = \mathbf{TDQ}^T + \mathbf{F}_R^* = \sum_{r=1}^{R} d_r \, \mathbf{t}_r \mathbf{q}_r^T + \mathbf{F}_R^*. \tag{22}$$

The following proposition states the conditions for the optimal selection of $\mathbf{u}$ given $\mathbf{Y}$ and $\mathbf{q}$.

Proposition 3.4. *Let $\mathbf{Y} \in \mathbb{R}^{I \times M}$ and $\mathbf{q} \in \mathbb{R}^M$ is of length one, then $\min_{\mathbf{u}} \|\mathbf{Y} - \mathbf{u}\mathbf{q}^T\|_F^2$ is solved by $\mathbf{u} = \mathbf{Yq}$. In other words, a linear combination of the columns of $\mathbf{Y}$ using a weighting vector $\mathbf{q}$ of length one is second-order optimal for the approximation of $\mathbf{Y}$ in the LS sense.*

Proof. Since $\mathbf{q}$ is given and $\|\mathbf{q}\| = 1$, it is obvious that the ordinary LS solution to solve the problem is $\mathbf{u} = \mathbf{Yq}(\mathbf{q}^T\mathbf{q})^{-1}$, and hence, $\mathbf{u} = \mathbf{Yq}$ gives the best fit of $\mathbf{Y}$ for that $\mathbf{q}$. □

The sequential deflation type optimization is employed for the model parameters. Observe that according to Proposition 3.4, the optimal fit is achieved for $\mathbf{u} = \mathbf{Yq}$. Thus, combining the linear relation between $\mathbf{t}$ and $\mathbf{u}$ with the expression for the core tensor $\underline{\mathbf{G}}$ in (15), we can optimize the parameters of X-loading matrices $\mathbf{P}^{(n)}$ and Y-loading vector $\mathbf{q}$ via

$$\begin{aligned} &\max_{\{\mathbf{P}^{(n)}, \mathbf{q}\}} \|\underline{\mathbf{X}} \times_1 \mathbf{Y}^T \times_1 \mathbf{q}^T \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T}\|_F^2, \\ &\text{s.t.} \quad \mathbf{P}^{(n)T}\mathbf{P}^{(n)} = \mathbf{I}, \|\mathbf{q}\|_F = 1. \end{aligned} \tag{23}$$

This form is similar to (18), but has a different cross covariance tensor $\underline{\mathbf{C}} = \underline{\mathbf{X}} \times_1 \mathbf{Y}^{\mathrm{T}}$ defined between a tensor and a matrix, implying that the problem can be solved by performing a rank-$(1, L_2, \ldots, L_N)$ HOSVD on $\underline{\mathbf{C}}$, which also makes it possible to compute the core tensor $\underline{\mathbf{G}}^{(C)}$ corresponding to $\underline{\mathbf{C}}$.

Subsequently, the latent vector $\mathbf{t}$ is estimated from $\underline{\mathbf{X}}$ using the given loading matrices $\mathbf{P}^{(n)}$. The mode-1 matricization of the model for $\underline{\mathbf{X}}$ allows us to write

$$\mathbf{X}_{(1)} = \mathbf{t}\mathbf{G}_{(1)}(\mathbf{P}^{(N-1)T} \otimes \cdots \otimes \mathbf{P}^{(1)})^T + \mathbf{E}_{(1)}, \tag{24}$$

where $\mathbf{G}_{(1)} \in \mathbb{R}^{1 \times L_2 L_3 \ldots L_N}$ still remains unknown. However, owing to the linear relationship between the core tensor $\underline{\mathbf{G}}$ (i.e., $[\![\underline{\mathbf{X}}; \mathbf{t}^T, \mathbf{P}^{(1)T}, \ldots, \mathbf{P}^{(N-1)T}]\!]$) and the core tensor $\underline{\mathbf{G}}^{(C)}$ (i.e., $[\![\underline{\mathbf{C}}; \mathbf{q}^T, \mathbf{P}^{(1)T}, \ldots, \mathbf{P}^{(N-1)T}]\!]$), so that $\underline{\mathbf{G}}^{(C)} = d\underline{\mathbf{G}}$, an LS solution for the normalized $\mathbf{t}$, which minimizes the squared norm of the residual $\|\mathbf{E}_{(1)}\|_F^2$, can be obtained from

$$\mathbf{t} \leftarrow \left(\underline{\mathbf{X}} \times_2 \mathbf{P}^{(1)T} \times_3 \cdots \times_N \mathbf{P}^{(N-1)T}\right)_{(1)} \mathbf{G}_{(1)}^{(C)+}, \; \mathbf{t} \leftarrow \mathbf{t}/\|\mathbf{t}\|_F, \tag{25}$$

where we used the property that $\mathbf{P}^{(n)}$ are columnwise orthonormal and the symbol "+" denotes *Moore-Penrose pseudoinverse*. Based on the estimated latent vector $\mathbf{t}$ of length one and loadings $\mathbf{q}$, the regression coefficient used to predict $\mathbf{Y}$ is computed as

$$d = \mathbf{t}^T\mathbf{u} = \mathbf{t}^T\mathbf{Yq}. \tag{26}$$

The procedure for a two-way response matrix is summarized in Algorithm 2. In this case, the HOPLS model is shown to unify both the standard PLS and N-PLS, when the appropriate parameters L_n are selected.[2]

Algorithm 2. Higher-order Partial Least Squares (HOPLS2) for a Tensor $\underline{\mathbf{X}}$ and a Matrix $\mathbf{Y}$

Input: $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$, $N \geq 3$ and $\mathbf{Y} \in \mathbb{R}^{I_1 \times M}$
The Number of latent vectors is R and the number of loadings are $\{L_n\}_{n=2}^N$.

Output: $\{\mathbf{P}_r^{(n)}\}; \mathbf{Q}; \{\underline{\mathbf{G}}_r\}; \mathbf{D}; \mathbf{T}; r = 1, \ldots, R, n = 2, \ldots, N$.

Initialization: $\underline{\mathbf{E}}_1 \leftarrow \underline{\mathbf{X}}, \mathbf{F}_1 \leftarrow \mathbf{Y}$.
for $r = 1$ **to** R **do**
  if $\|\underline{\mathbf{E}}_r\|_F > \varepsilon$ **and** $\|\mathbf{F}_r\|_F > \varepsilon$ **then**
    $\underline{\mathbf{C}}_r \leftarrow \underline{\mathbf{E}}_r \times_1 \mathbf{F}_r^T$;
    Perform rank-$(1, L_2, \ldots, L_N)$ HOOI on $\underline{\mathbf{C}}_r$ as
    $\underline{\mathbf{C}}_r \approx \underline{\mathbf{G}}_r^{(C)} \times_1 \mathbf{q}_r \times_2 \mathbf{P}_r^{(1)} \times_3 \cdots \times_N \mathbf{P}_r^{(N-1)}$;
    $\mathbf{t}_r \leftarrow (\underline{\mathbf{E}}_r \times_2 \mathbf{P}_r^{(1)} \times_3 \cdots \times_N \mathbf{P}_r^{(N-1)})_{(1)} (\mathrm{vec}^T(\underline{\mathbf{G}}_r^{(C)}))^+$;
    $\mathbf{t}_r \leftarrow \mathbf{t}_r / \|\mathbf{t}_r\|_F$;
    $\underline{\mathbf{G}}_r \leftarrow [\![\underline{\mathbf{E}}_r; \mathbf{t}_r^T, \mathbf{P}_r^{(1)T}, \ldots, \mathbf{P}_r^{(N-1)T}]\!]$;
    $\mathbf{u}_r \leftarrow \mathbf{F}_r \mathbf{q}_r$;
    $d_r \leftarrow \mathbf{u}_r^T \mathbf{t}_r$;
    Deflation:
    $\underline{\mathbf{E}}_{r+1} \leftarrow \underline{\mathbf{E}}_r - [\![\underline{\mathbf{G}}_r; \mathbf{t}_r, \mathbf{P}_r^{(1)}, \ldots, \mathbf{P}_r^{(N-1)}]\!]$;
    $\mathbf{F}_{r+1} \leftarrow \mathbf{F}_r - d_r \mathbf{t}_r \mathbf{q}_r^T$;
  end if
end for

3.4 Prediction of the Response Variables

Predictions from the new observations $\underline{\mathbf{X}}^{new}$ are performed in two steps: projecting the data to the low-dimensional latent space based on model parameters $\underline{\mathbf{G}}_r$, $\mathbf{P}_r^{(n)}$, and predicting the response data based on latent vectors $\mathbf{T}^{new}$ and model parameters $\mathbf{Q}_r^{(m)}$, $\underline{\mathbf{D}}_r$. For simplicity, we use a matricized form to express the prediction procedure as

$$\hat{\underline{\mathbf{Y}}}_{(1)}^{new} \approx \mathbf{T}^{new}\mathbf{Q}^{*T} = \mathbf{X}_{(1)}^{new}\mathbf{W}\mathbf{Q}^{*T}, \tag{27}$$

2. Explanation and proof are given in the supplement material, available online.

where $\mathbf{W}$ and $\mathbf{Q}^*$ have R columns represented by

$$\begin{aligned}\mathbf{w}_r &= \left(\mathbf{P}_r^{(N-1)} \otimes \cdots \otimes \mathbf{P}_r^{(1)}\right)\underline{\mathbf{G}}_{r(1)}^{+},\\ \mathbf{q}_r^* &= \underline{\mathbf{D}}_{r(1)}\left(\mathbf{Q}_r^{(M-1)} \otimes \cdots \otimes \mathbf{Q}_r^{(1)}\right)^T.\end{aligned} \tag{28}$$

In the particular case of a two-way matrix $\mathbf{Y}$, the prediction is performed by

$$\hat{\mathbf{Y}}^{new} \approx \mathbf{X}_{(1)}^{new}\mathbf{W}\mathbf{D}\mathbf{Q}^T, \tag{29}$$

where $\mathbf{D}$ is a diagonal matrix whose entries are $\mathbf{d}_r$ and the rth column of $\mathbf{Q}$ is $\mathbf{q}_r$, $r = 1, \ldots, R$.

3.5 Properties of HOPLS

Robustness to noise. An additional constraint of keeping the largest $\{L_n\}_{n=2}^{N}$ loading vectors on each mode is imposed in HOPLS, resulting in a flexible model that balances the two objectives of fitness and the significance of associated latent variables. For instance, a larger L_n may fit $\underline{\mathbf{X}}$ better but introduces more noise to each latent vector. In contrast, N-PLS is more robust due to the strong constraint of rank-one tensor structure while lacking good fit to the data. The flexibility of HOPLS allows us to adapt the model complexity based on the dataset in hand, providing considerable prediction ability (see Figs. 4 and 6).

"Large p, Small n" problem. This is particularly important when the dimension of independent variables is high. In contrast to PLS, the relatively low dimension of model parameters that need to be optimized in HOPLS. For instance, assume that a third-order tensor $\underline{\mathbf{X}}$ has the dimension of $5 \times 10 \times 100$, i.e., there are 5 samples and 1,000 features. If we apply PLS on $\mathbf{X}_{(1)}$ with size of $5 \times 1{,}000$, there are only five samples available to optimize a 1,000-dimensional loading vector $\mathbf{p}$, resulting in an unreliable estimate of model parameters. In contrast, HOPLS allows us to optimize loading vectors having relatively low dimension on each mode alternately; thus the number of samples is significantly elevated. For instance, to optimize 10-dimensional loading vectors on the second mode, 500 samples are available, and to optimize the 100-dimensional loading vectors on the third mode there are 50 samples. Thus, a more robust estimate of low-dimensional loading vectors can be obtained, which is also less prone to overfitting and more suitable for "Large p, Small n" problem (see Fig. 4).

Ease of interpretation. The loading vectors in $\mathbf{P}^{(n)}$ reveal new subspace patterns corresponding to the n-mode features. However, the loadings from Unfold-PLS are difficult to interpret since the data structure is destroyed by the unfolding operation and the dimension of loadings is relatively high.

Computation. N-PLS is implemented by combining a NIPALS-like algorithm with the CP decomposition. Instead of using an iterative algorithm, HOPLS can find the model parameters using a closed-form solution, i.e., applying HOSVD on the cross-covariance tensor, resulting in enhanced computational efficiency.

Due to the flexibility of HOPLS, the tuning parameters of L_n and K_m controlling the model complexity need to be selected based on calibration data. Similarly to the parameter R, the tuning parameters can be chosen by cross validation. For simplicity, two alternative assumptions will been utilized: 1) $\forall n, \forall m, L_n = K_m = \lambda$, 2) $L_n = \eta R_n$, $K_m = \eta R_m$, $0 < \eta \leq 1$, i.e., explaining the same percentage of the n-mode variance.

4 Experimental Results

In the simulations, HOPLS and N-PLS were used to model the data in a tensor form, whereas PLS was performed on a mode-1 matricization of the same tensors. To quantify the predictability, the index Q^2 was defined as $Q^2 = 1 - \|\underline{\mathbf{Y}} - \hat{\underline{\mathbf{Y}}}\|_F^2 / \|\underline{\mathbf{Y}}\|_F^2$, where $\hat{\underline{\mathbf{Y}}}$ denotes the prediction of $\underline{\mathbf{Y}}$ using a model created from a calibration dataset. Root mean square errors of prediction (RMSEP) were also used for evaluation [48].

4.1 Synthetic Data

To quantitatively benchmark our algorithm against the state of the art, an extensive comparative exploration has been performed on synthetic datasets to evaluate the prediction performance under varying conditions with respect to data structure, noise levels, and ratio of variable dimension to sample size. For parameter selection, the number of latent vectors (R) and number of loadings ($L_n = K_m = \lambda$) were chosen based on five-fold cross validation on the calibration dataset. To reduce random fluctuations, evaluations were performed over 50 validation datasets generated repeatedly according to the same criteria.

4.1.1 Datasets with Matrix Structure

The independent data $\mathbf{X}$ and dependent data $\mathbf{Y}$ were generated as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \xi\,\mathbf{E}, \qquad \mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \xi\mathbf{F}, \tag{30}$$

where latent variables $\{\mathbf{t}, \mathbf{p}, \mathbf{q}\} \sim \mathcal{N}(0, 1)$, $\mathbf{E}$, $\mathbf{F}$ are Gaussian noises whose level is controlled by the parameter ξ. Both the calibration and the validation datasets were generated according to (30), with the same loadings $\mathbf{P}, \mathbf{Q}$, but a different latent $\mathbf{T}$ which follows the same distribution $\mathcal{N}(0, 1)$. Subsequently, the datasets were reorganized as Nth-order tensors.

To investigate how the prediction performance is affected by noise levels and small sample size, $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{20\times10\times10}$ (Case 1) and $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$ (Case 2) were generated under varying noise levels of 10, 5, 0, and -5 dB. In Case 3, $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$ were generated with the loadings $\mathbf{P}, \mathbf{Q}$ drawn from a uniform distribution $U(0, 1)$. The datasets were generated from five latent variables (i.e., $\mathbf{T}$ has five columns) for all the three cases.

There are two tuning parameters, i.e., number of latent variables R and number of loadings λ for HOPLS, and only one parameter R for PLS and N-PLS that need to be selected appropriately. The number of latent variables R is crucial to prediction performance, resulting in undermodeling when R was too small, while overfitting easily when R was too large. The cross validations were performed when R and λ were varying from 1 to 10 with the step length of 1. To alleviate the computation burden, the procedure was stopped when the performance starts to decrease with increasing λ. Fig. 3 shows the grid of cross-validation performance of HOPLS in Case 2 with the optimal parameters marked by green squares. Observe that the

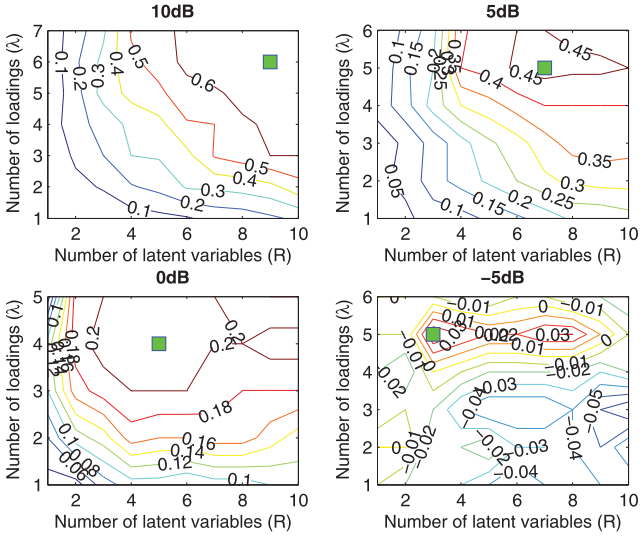


Fig. 3. Five-fold cross-validation performance of HOPLS at different noise levels versus the number of latent variables (R) and loadings (λ). The optimal values for these two parameters are marked by green squares.

optimal λ for HOPLS is related to the noise levels, and for increasing noise levels, the best performance is obtained by smaller λ, implying that only a few significant loadings on each mode are kept in the latent space. This is expected due to the fact that the model complexity is controlled by λ to suppress noise. The optimal R and λ for all three methods at different noise levels are shown in Table 1.

After the selection the parameters, HOPLS, N-PLS, and PLS were retrained on the whole calibration dataset using the optimal R and λ, and were applied to the validation datasets for evaluation. Fig. 4 illustrates the predictive performance over 50 validation datasets for the three cases at four different noise levels. In Case 1, a relatively larger sample size was available; when $\mathrm{SNR} = 10$ dB, HOPLS achieved a similar prediction performance to PLS while outperforming N-PLS. With increasing the noise level in both the calibration and validation datasets, HOPLS showed a relatively stable performance, whereas the performance of PLS decreased significantly. The superiority of HOPLS was shown clearly with increasing the noise level. In Case 2, where a smaller sample size was available, HOPLS exhibited better performance than the other two models and the superiority of HOPLS was more pronounced at high noise levels, especially for SNR ≤ 5 dB. These results demonstrated that HOPLS is more robust to noise in comparison with N-PLS and PLS. If we compare Case 1 with Case 2 at different noise levels, the results revealed that the superiority of HOPLS over the other two methods was enhanced in Case 2, illustrating the advantage

TABLE 1
The Selection of Parameters R and λ in Case 2

SNR	PLS	N-PLS	HOPLS		SNR	PLS	N-PLS	HOPLS	
			R	λ				R	λ
10dB	5	7	9	6	0dB	3	5	5	4
5dB	5	6	7	5	-5dB	3	1	3	5

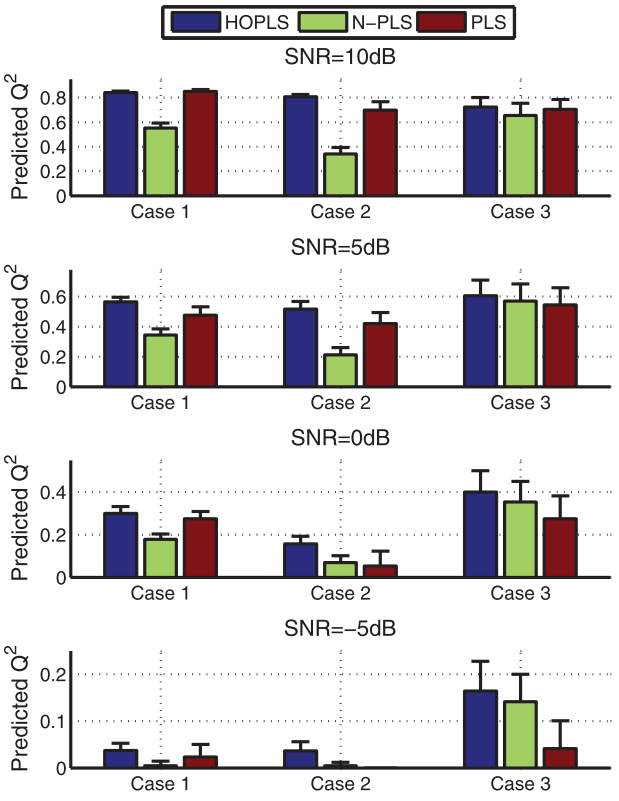


Fig. 4. The prediction performance comparison among HOPLS, N-PLS, and PLS at different noise levels for three cases. Case 1: $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{20\times10\times10}$ and $\{\mathbf{P}, \mathbf{Q}\} \sim \mathcal{N}(0,1)$; Case 2: $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$ and $\{\mathbf{P}, \mathbf{Q}\} \sim \mathcal{N}(0,1)$; Case 3: $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$ and $\{\mathbf{P}, \mathbf{Q}\} \sim U(0,1)$.

of HOPLS in modeling datasets with small sample size. Note that N-PLS also showed better performance than PLS when SNR ≤ 0 dB in Case 2, demonstrating the advantages of modeling the dataset in a tensor form for small sample sizes. In Case 3, N-PLS showed much better performance as compared to its performance in Cases 1 and 2, implying sensitivity of N-PLS to data distribution. With the increasing noise level, both HOPLS and N-PLS showed enhanced predictive abilities over PLS.

4.1.2 Datasets with Tensor Structure

Note that the datasets generated by (30) do not originally possess multiway data structures, although they were organized in a tensor form; thus the structure information of data was not important for prediction. We here assume that HOPLS is more suitable for the datasets which originally have multiway structure, i.e., information carried by interaction among each mode are useful for our regression problem. To verify our assumption, the independent data $\underline{\mathbf{X}}$ and dependent data $\underline{\mathbf{Y}}$ were generated according to the Tucker model that is regarded as a general model for tensors. The latent variables $\mathbf{t}$ were generated in the same way as described in Section 4.1.1. A sequence of loadings $\mathbf{P}^{(n)}, \mathbf{Q}^{(m)}$ and the core tensors were drawn from $\mathcal{N}(0,1)$. For the validation dataset, the latent matrix $\mathbf{T}$ was generated from the same distribution as the calibration dataset, while the core tensors and loadings were fixed. Similarly to the study in Section 4.1.1, to investigate how the prediction performance is affected by noise levels and sample size, $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{20\times10\times10}$ (Case 1) and $\{\underline{\mathbf{X}}, \underline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$ (Case 2) were generated under

TABLE 2
The Selection of Parameters R and λ in Case 2

SNR	PLS	N-PLS	HOPLS		SNR	PLS	N-PLS	HOPLS	
			R	λ				R	λ
10dB	5	7	9	4	0dB	4	4	4	2
5dB	4	6	8	2	-5dB	2	4	2	1

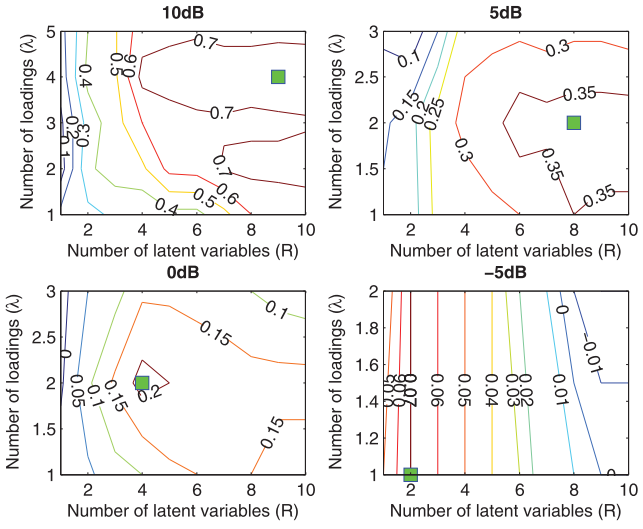


Fig. 5. Five-fold cross-validation performance of HOPLS at different noise levels versus the number of latent variables (R) and loadings (λ). The optimal values for these two parameters are marked by green squares.

noise levels of 10, 5, 0, and -5 dB. The datasets for both cases were generated from five latent variables.

The optimal parameters of R and λ were shown in Table 2. Observe that the optimal R is smaller with the increasing noise level for all three methods. The parameter λ in HOPLS was also shown to have a similar behavior. For more detail, Fig. 5 exhibits the cross-validation performance grid of HOPLS with respect to R and λ. When SNR was 10 dB, the optimal λ was 4, while it was 2, 2, and 1 for 5, 0, and -5 dB, respectively. This indicates that the model complexity can be adapted to provide a better model when a specific dataset was given, demonstrating the flexibility of the HOPLS model.

The prediction performance evaluated over 50 validation datasets using HOPLS, N-PLS, and PLS with individually selected parameters was compared for different noise levels and different sample sizes (i.e., two cases). As shown in Fig. 6, for both cases the prediction performance of HOPLS was better than both N-PLS and PLS at 10 dB, and the discrepancy among them was enhanced when SNR changed from 10 to -5 dB. The performance of PLS decreased significantly with the increasing noise levels while HOPLS and N-PLS showed relative robustness to noise and outperformed PLS when SNR ≤ 5 dB, illustrating the advantages of tensor-based methods with respect to noisy data. Regarding the small sample size problem, we found the performances of all three methods were decreased when comparing Case 1 with Case 2. Observe that the superiority of HOPLS over N-PLS and PLS was enhanced in Case 2 as compared to Case 1 at all noise

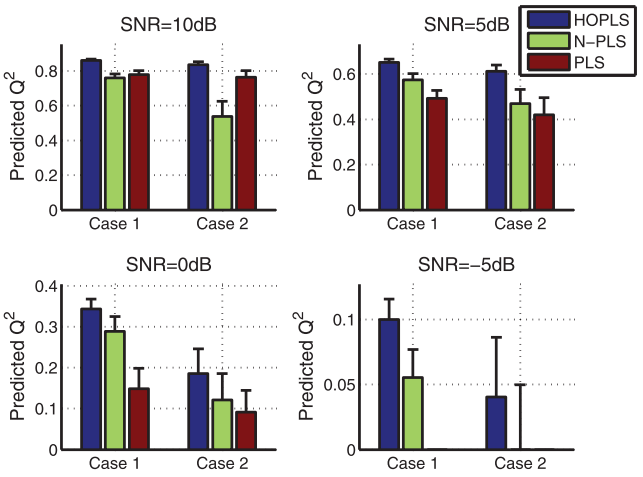


Fig. 6. The prediction performance comparison among HOPLS, N-PLS, and PLS at different noise levels for the two cases (i.e., Case1: $\{\overline{\mathbf{X}}, \overline{\mathbf{Y}}\} \in \mathbb{R}^{20\times10\times10}$ and Case 2: $\{\overline{\mathbf{X}}, \overline{\mathbf{Y}}\} \in \mathbb{R}^{10\times10\times10}$) with different sample size.

levels. A comparison of Figs. 6 and 4 shows that the performances are significantly improved when handling the datasets having tensor structure by tensor-based methods (e.g., HOPLS and N-PLS). As for N-PLS, it outperformed PLS when the datasets have tensor structure and in the presence of high noise, but it may not perform well when the datasets have no tensor structure. By contrast, HOPLS performed well in both cases, in particular, it outperformed both N-PLS and PLS in critical cases with high noise and small sample size.[3]

4.1.3 Comparison on Matrix Response Data

In this simulation, the response data were a two-way matrix; thus, the HOPLS2 algorithm was used to evaluate the performance. $\underline{\mathbf{X}} \in \mathbb{R}^{5\times5\times5\times5}$ and $\mathbf{Y} \in \mathbb{R}^{5\times2}$ were generated from a full-rank normal distribution $\mathcal{N}(0,1)$, which satisfies $\mathbf{Y} = \mathbf{X}_{(1)}\mathbf{W}$, where $\mathbf{W}$ was also generated from $\mathcal{N}(0,1)$. Fig. 7A visualizes the predicted and original data with the red line indicating the ideal prediction. Observe that HOPLS was able to predict the validation dataset with smaller error than PLS and N-PLS. The independent data and dependent data are visualized in the latent space as shown in Fig. 7B.

4.2 Decoding of ECoG Signals

In [46], ECoG-based decoding of 3D hand trajectories was demonstrated by means of classical PLSR[4] [49]. The movement of monkeys was captured by an optical motion capture system (Vicon Motion Systems, USA). In all experiments, each monkey wore a custom-made jacket with reflective markers for motion capture affixed to the left shoulder, elbows, wrists, and hand; thus the response data were naturally represented as a third-order tensor (i.e., time $\times$ 3D positions $\times$ markers). Although PLS can be applied to predict the trajectories corresponding to each marker individually, the structure information among four markers would be unused. For every point in time of the movement trajectory, a corresponding ECoG epoch

3. The Matlab code and one simulation dataset are available from http://www.bsp.brain.riken.jp/%7Eqibin/homepage/HOPLS.html.

4. The datasets and more detailed description are freely available from http://neurotycho.org.

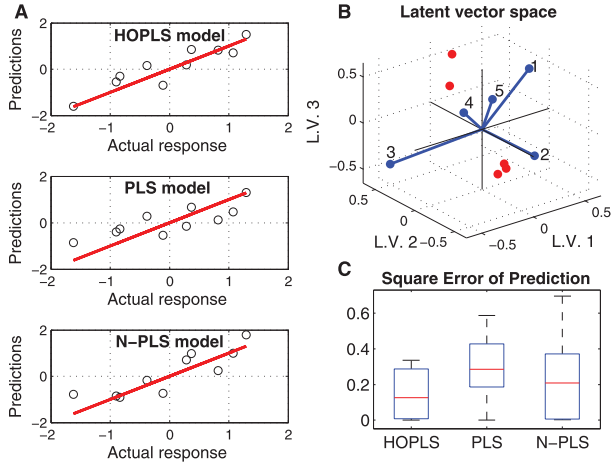


Fig. 7. (A) The scatter plot of predicted against actual data for each model. (B) Data distribution in the latent spaces. Each blue point denotes one sample of the independent variable, while the red points denote samples of response variables. (C) depicts the distribution of the square error of prediction on the validation dataset.

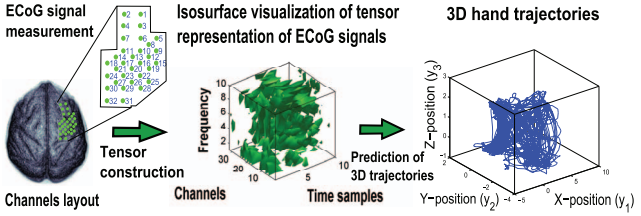


Fig. 8. The scheme for decoding of 3D hand movement trajectories from ECoG signals.

(1 second before the current time point) was extracted and transformed to the time-frequency domain in order to construct the predictors. Hence, the independent data are also naturally represented as a higher order tensor (i.e., epoch × channel × time × frequency). In this study, the proposed HOPLS regression model was applied for decoding movement trajectories based on ECoG signals to verify its effectiveness in real-world applications.

The overall scheme of ECoG decoding is illustrated in Fig. 8. Specifically, ECoG signals were preprocessed by a band-pass filter with cutoff frequencies at 0.1 and 600 Hz and a spatial filter with a common average reference. Motion marker positions were downsampled to 20 Hz. To represent features related to the movement trajectory from ECoG signals, the Morlet wavelet transformation at 10 different center frequencies (10-150 Hz, arranged in a logarithmic scale) was used to obtain the time-frequency representation. The epochs of 1-second ECoG signals before every time point of the movement data were extracted and downsampled to 10 Hz in order to construct the predictors, which can be represented by channel × time × frequency ($32 \times 10 \times 10$).

We first applied the HOPLS2 algorithm to predict only the hand movement trajectory, represented as a matrix $\mathbf{Y} \in \mathbb{R}^{I_1 \times 3}$ (time points × 3D positions) from a three-order tensor of ECoG predictors $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times 32 \times 100}$ (epoch × channel × time-frequency) that was constructed by combining the time and frequency modes into one mode. The ECoG data were divided into a calibration dataset (10 minutes) and a validation dataset (5 minutes). To select the optimal parameters of L_n and R, the cross validation was applied

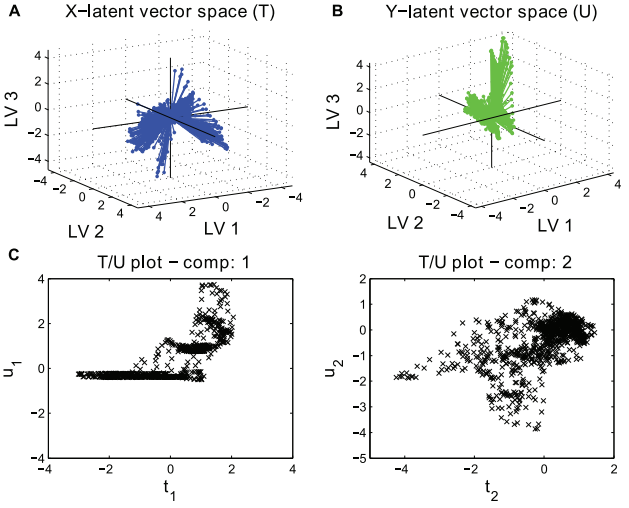


Fig. 9. Panels (A) and (B) depict data distributions in the $\underline{\mathbf{X}}$-latent space $\mathbf{T}$ and $\mathbf{Y}$-latent space $\mathbf{U}$, respectively. (C) presents a joint distribution between $\underline{\mathbf{X}}$- and $\mathbf{Y}$-latent vectors.

on the calibration dataset. Finally, $L_n = 10$ and $R = 23$ were selected for the HOPLS model. Likewise, the best values of R for PLS and N-PLS were 19 and 60, respectively. The X-latent space is visualized in Fig. 9A, where each point represents one sample of independent variables, while the Y-latent space is presented in Fig. 9B, with each point representing one-dependent sample. Observe that the distributions of these two latent variable spaces were quite similar, and the two dominant clusters are clearly distinguished. The joint distributions between each $\mathbf{t}_r$ and $\mathbf{u}_r$ are depicted in Fig. 9C. Two clusters can be observed from the first component which might be related to the "movement" and "nonmovement" behaviors.

Another advantage of HOPLS was better physical interpretation of the model. To investigate how the spatial, spectral, and temporal structure of ECoG data were used to create the regression model, loading vectors can be

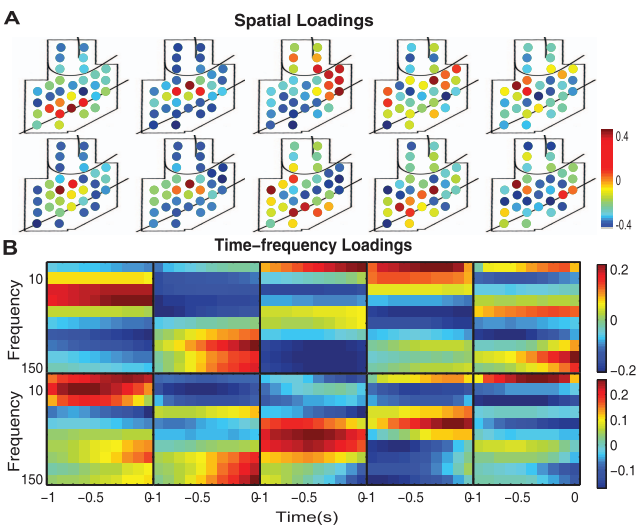


Fig. 10. (A) Spatial loadings $\mathbf{P}_r^{(1)}$ corresponding to the first two latent components. Each row shows five significant loading vectors. Likewise, (B) depicts time-frequency loadings $\mathbf{P}_r^{(2)}$, with the β- and γ-bands exhibiting significant contribution.

TABLE 3
Comprehensive Comparison of the HOPLS, N-PLS, and PLS for the Prediction of 3D Hand Trajectories on the Four Datasets Recorded from the Same Monkey Brain

Data Set	Model	Q^2(ECoG)	Q^2(3D hand positions)				RMSEP (3D hand positions)				Correlation		
			X	Y	Z	Mean	X	Y	Z	Mean	X	Y	Z
Dataset 1	HOPLS	0.25	**0.43**	**0.48**	**0.61**	**0.51**	**0.82**	**0.70**	**0.66**	**0.73**	**0.67**	**0.72**	**0.78**
	N-PLS	0.33	0.39	0.44	0.59	0.47	0.85	0.73	0.68	0.75	0.64	0.71	0.77
	Unfold-PLS	0.23	0.39	0.45	0.59	0.48	0.85	0.72	0.68	0.75	0.64	**0.72**	0.77
Dataset 2	HOPLS	0.25	**0.12**	**0.42**	0.50	**0.35**	**0.99**	**0.77**	0.72	**0.83**	**0.35**	**0.64**	0.71
	N-PLS	0.33	0.03	0.40	0.51	0.32	1.04	0.78	0.71	0.84	0.32	**0.64**	0.71
	Unfold-PLS	0.22	0.05	0.40	**0.53**	0.32	1.04	0.78	**0.70**	0.84	0.34	0.63	**0.73**
Dataset 3	HOPLS	0.22	**0.36**	**0.39**	**0.48**	**0.41**	**0.74**	**0.77**	**0.66**	**0.73**	**0.62**	**0.62**	**0.69**
	N-PLS	0.30	0.31	0.37	0.46	0.38	0.77	0.78	0.68	0.74	0.61	**0.62**	0.68
	Unfold-PLS	0.21	0.30	0.37	0.46	0.38	0.77	0.79	0.67	0.74	0.61	**0.62**	0.68
Dataset 4	HOPLS	0.16	**0.16**	**0.50**	**0.57**	**0.41**	**1.04**	**0.66**	**0.62**	**0.77**	**0.43**	**0.71**	**0.76**
	N-PLS	0.23	0.12	0.45	0.55	0.37	1.06	0.69	0.67	0.80	0.41	0.70	**0.76**
	Unfold-PLS	0.15	0.11	0.46	**0.57**	0.38	1.07	0.69	**0.62**	0.79	0.42	0.70	**0.76**

The number of latent vectors for HOPLS, N-PLS, and Unfold-PLS was 23, 60, and 19, respectively.

regarded as a subspace basis in spatial and time-frequency domains, as shown in Fig. 10. With regard to time-frequency loadings, the β- and γ-band activities were most significant implying the importance of β, γ-band activities for encoding of movements; the duration of the β-band was longer than that of the γ-band, which indicates that hand movements were related to long history oscillations of the β-band and short history oscillations of the γ-band. These findings also demonstrated that high gamma band activity in the premotor cortex is associated with movement preparation, initiation, and maintenance [50].

From Table 3, observe that the improved prediction performances were achieved by HOPLS, for all the performance metrics. In particular, the results from dataset 1 demonstrated that the improvements by HOPLS over N-PLS were 0.03 for the correlation coefficient of the X-position, 0.02 for averaged RMSEP, 0.04 for averaged Q^2, whereas the improvements by HOPLS over PLS were 0.03 for the correlation coefficient of X-position, 0.02 for averaged RMSEP, and 0.03 for averaged Q^2.

Since HOPLS enables us to create a regression model between two higher order tensors, all trajectories recorded from shoulder, elbow, wrist, and hand were constructed as a tensor $\underline{\mathbf{Y}} \in \mathbb{R}^{I_1 \times 3 \times 4}$ (samples $\times$ 3D positions $\times$ markers). To verify the superiority of HOPLS for small sample sizes, we used 100 second data for calibration and 100 second data for validation. The resolution of time-frequency representations was improved to provide more detailed features; thus, we have a fourth-order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I_1 \times 32 \times 20 \times 20}$ (samples $\times$ channels $\times$ time $\times$ frequency). The prediction performances from HOPLS, N-PLS, and PLS are shown in Fig. 11, illustrating the effectiveness of HOPLS when the response data originally have tensor structure.

Time-frequency features of ECoG epochs for each sample are extremely overlapped, resulting in a lot of information redundancy and high computational burden. In addition, it is generally not necessary to predict behaviors with a high time resolution. Hence, an additional analysis has been performed by downsampling motion marker positions at 1 Hz, to ensure that nonoverlapped features were used in any adjacent samples. The cross-validation performance

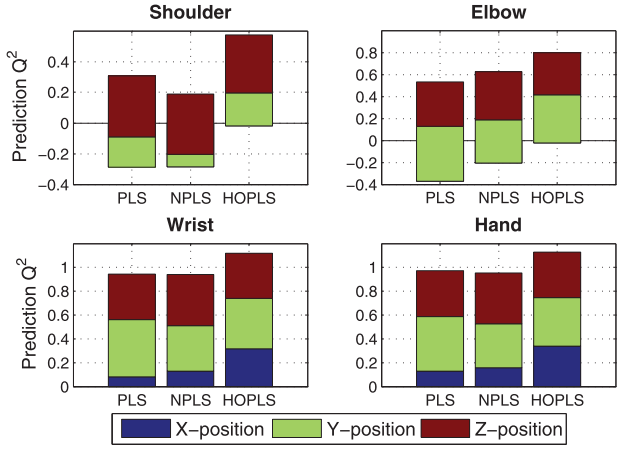


Fig. 11. The prediction performance of 3D trajectories recorded from shoulder, elbow, wrist, and hand. The optimal R is 16, 28, and 49 for PLS, N-PLS, and HOPLS, respectively, and $\lambda = 5$ for HOPLS.

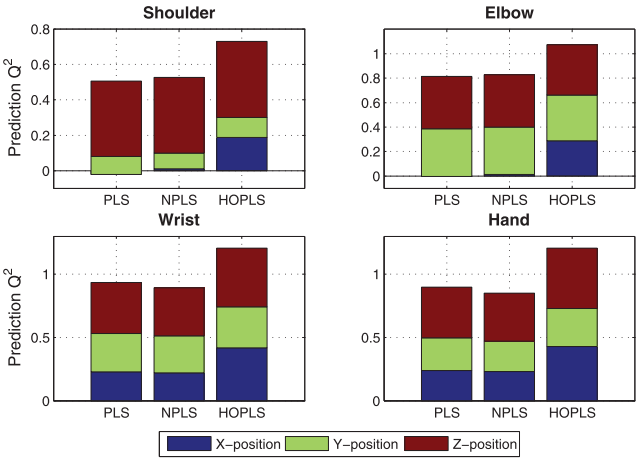


Fig. 12. The prediction performance of 3D trajectories for shoulder, elbow, wrist, and hand using nonoverlapped ECoG features.

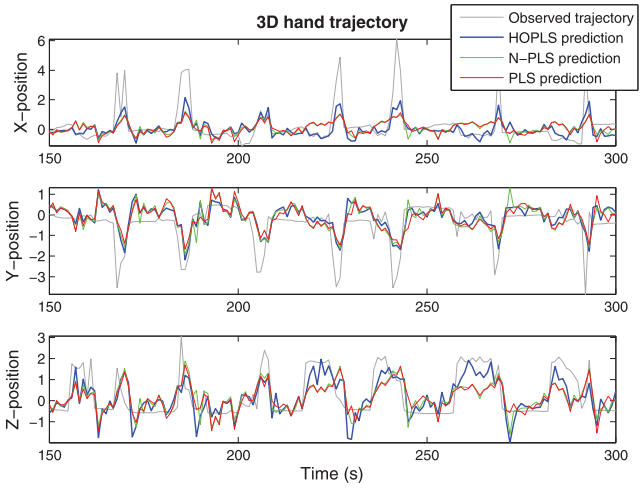


Fig. 13. Visualization of observed trajectories (150 s time window) and the trajectories predicted by HOPLS, N-PLS, and PLS.

was evaluated for all the markers from the 10 minute calibration dataset and the best performance for PLS of $Q^2 = 0.19$ was obtained using $R = 2$, for N-PLS it was $Q^2 = 0.22$ obtained by $R = 5$, and for HOPLS it was $Q^2 = 0.28$ obtained by $R = 24$, $\lambda = 5$. The prediction performances on the 5 minute validation dataset are shown in Fig. 12, implying the significant improvements obtained by HOPLS over N-PLS and PLS for all four markers. For visualization, Fig. 13 exhibits the observed and predicted 3D hand trajectories in the 150 s time window. The video is also available for visualization of the results.[5]

5 Conclusions

The HOPLS has been proposed as a generalized multilinear regression model. The analysis and simulations have shown that the advantages of the proposed model include its robustness to noise and enhanced performance for small sample sizes. In addition, HOPLS provides an optimal tradeoff between fitness and overfitting due to the fact that model complexity can be adapted by a hyperparameter. The proposed strategy to find a closed-form solution for HOPLS makes computation more efficient than the existing algorithms. The results for a real-world application in decoding 3D movement trajectories from ECoG signals have also demonstrated that HOPLS would be a promising multilinear subspace regression method.

Acknowledgments

This work was supported in part by JSPS Grants-in-Aid for Scientific Research (Grant No. 24700154), the National Natural Science Foundation of China (Grant Nos. 90920014, 91120305, 61202155), and the CONICET project PIP 2012-2014 (No. 11420110100021).

References

[1] D.C. Montgomery, E.A. Peck, and G.G. Vining, *Introduction to Linear Regression Analysis*, third ed. John Wiley & Sons, 2001.

[2] C. Dhanjal, S. Gunn, and J. Shawe-Taylor, "Efficient Sparse Kernel Feature Extraction Based on Partial Least Squares," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1347-1361, Aug. 2009.

[3] H. Wold, "Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach," *Perspectives in Probability and Statistics*, pp. 117-142, Academic Press, 1975.

[4] A. Krishnan, L. Williams, A. McIntosh, and H. Abdi, "Partial Least Squares (PLS) Methods for Neuroimaging: A Tutorial and Review," *NeuroImage*, vol. 56, no. 2, pp. 455-475, 2010.

[5] H. Abdi, "Partial Least Squares Regression and Projection on Latent Structure Regression (PLS Regression)," *Wiley Interdisciplinary Rev.: Computational Statistics*, vol. 2, no. 1, pp. 97-106, 2010.

[6] R. Rosipal and N. Krämer, "Overview and Recent Advances in Partial Least Squares," *Proc. Int'l Conf. Subspace, Latent Structure and Feature Selection*, pp. 34-51, 2006.

[7] J. Trygg and S. Wold, "Orthogonal Projections to Latent Structures (O-PLS)," *J. Chemometrics*, vol. 16, no. 3, pp. 119-128, 2002.

[8] R. Ergon, "PLS Score-Loading Correspondence and a Bi-Orthogonal Factorization," *J. Chemometrics*, vol. 16, no. 7, pp. 368-373, 2002.

[9] S. Vijayakumar and S. Schaal, "Locally Weighted Projection Regression: An O(n) Algorithm for Incremental Real Time Learning in High Dimensional Space," *Proc. 17th Int'l Conf. Machine Learning*, vol. 1, pp. 288-293, 2000.

[10] R. Rosipal and L. Trejo, "Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space," *J. Machine Learning Research*, vol. 2, pp. 97-123, 2002.

[11] J. Ham, D. Lee, S. Mika, and B. Schölkopf, "A Kernel View of the Dimensionality Reduction of Manifolds," *Proc. 21st Int'l Conf. Machine Learning*, pp. 47-54, 2004.

[12] R. Bro, A. Rinnan, and N. Faber, "Standard Error of Prediction for Multilinear PLS-2. Practical Implementation in Fluorescence Spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 75, no. 1, pp. 69-76, 2005.

[13] B. Li, J. Morris, and E. Martin, "Model Selection for Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 64, no. 1, pp. 79-89, 2002.

[14] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Royal Statistical Soc., Series B (Methodological)*, vol. 58, pp. 267-288, 1996.

[15] R. Bro, "Multi-Way Analysis in the Food Industry: Models, Algorithms, and Applications," Academish Proefschrift, Dinamarca, 1998.

[16] T. Kolda and B. Bader, "Tensor Decompositions and Applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455-500, 2009.

[17] A. Cichocki, R. Zdunek, A.H. Phan, and S.I. Amari, *Nonnegative Matrix and Tensor Factorizations*. John Wiley & Sons, 2009.

[18] E. Acar, D. Dunlavy, T. Kolda, and M. Mørup, "Scalable Tensor Factorizations for Incomplete Data," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, pp. 41-56, 2011.

[19] R. Bro, "Multiway Calibration. Multilinear PLS," *J. Chemometrics*, vol. 10, no. 1, pp. 47-61, 1996.

[20] R. Bro, "Review on Multiway Analysis in Chemistry—2000-2005," *Critical Rev. Analytical Chemistry*, vol. 36, no. 3, pp. 279-293, 2006.

[21] K. Hasegawa, M. Arakawa, and K. Funatsu, "Rational Choice of Bioactive Conformations through Use of Conformation Analysis and 3-Way Partial Least Squares Modeling," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, no. 2, pp. 253-261, 2000.

[22] J. Nilsson, S. de Jong, and A. Smilde, "Multiway Calibration in 3D QSAR," *J. Chemometrics*, vol. 11, no. 6, pp. 511-524, 1997.

[23] K. Zissis, R. Brereton, S. Dunkerley, and R. Escott, "Two-Way, Unfolded Three-Way and Three-Mode Partial Least Squares Calibration of Diode Array HPLC Chromatograms for the Quantitation of Low-Level Pharmaceutical Impurities," *Analytica Chimica Acta*, vol. 384, no. 1, pp. 71-81, 1999.

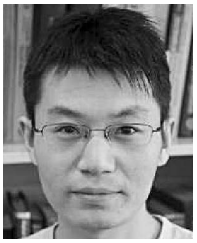
[24] E. Martinez-Montes, P. Valdés-Sosa, F. Miwakeichi, R. Goldman, and M. Cohen, "Concurrent EEG/fMRI Analysis by Multiway Partial Least Squares," *NeuroImage*, vol. 22, no. 3, pp. 1023-1034, 2004.

[25] E. Acar, C. Bingol, H. Bingol, R. Bro, and B. Yener, "Seizure Recognition on Epilepsy Feature Tensor," *Proc. 29th Ann. IEEE Int'l Conf. Eng. Medical Biomedical Soc.*, pp. 4273-4276, 2007.

[26] R. Bro, A. Smilde, and S. de Jong, "On the Difference between Low-Rank and Subspace Approximation: Improved Model for Multi-Linear PLS Regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 1, pp. 3-13, 2001.

5. http://www.bsp.brain.riken.jp/%7Eqibin/homepage/HOPLS.html.

[27] A. Smilde and H. Kiers, "Multiway Covariates Regression Models," *J. Chemometrics,* vol. 13, no. 1, pp. 31-48, 1999.
[28] A. Smilde, R. Bro, and P. Geladi, *Multi-Way Analysis with Applications in the Chemical Sciences.* Wiley, 2004.
[29] R.A. Harshman, "Foundations of the PARAFAC Procedure: Models and Conditions for an Explanatory Multimodal Factor Analysis," *UCLA Working Papers in Phonetics,* vol. 16, pp. 1-84, 1970.
[30] A. Smilde, "Comments on Multilinear PLS," *J. Chemometrics,* vol. 11, no. 5, pp. 367-377, 1997.
[31] M.D. Borraccetti, P.C. Damiani, and A.C. Olivieri, "When Unfolding Is Better: Unique Success of Unfolded Partial Least-Squares Regression with Residual Bilinearization for the Processing of Spectral-pH Data with Strong Spectral Overlapping. Analysis of Fluoroquinolones in Human Urine Based on Flow-Injection pH-Modulated Synchronous Fluorescence Data Matrices," *Analyst,* vol. 134, pp. 1682-1691, 2009.
[32] L. De Lathauwer, "Decompositions of a Higher-Order Tensor in Block Terms—Part II: Definitions and Uniqueness," *SIAM J. Matrix Analysis Applications,* vol. 30, no. 3, pp. 1033-1066, 2008.
[33] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A Multilinear Singular Value Decomposition," *SIAM J. Matrix Analysis and Applications,* vol. 21, no. 4, pp. 1253-1278, 2000.
[34] L.R. Tucker, "Implications of Factor Analysis of Three-Way Matrices for Measurement of Change," *Problems in Measuring Change,* C.W. Harris, ed., pp. 122-137, Univ. of Wisconsin Press, 1963.
[35] T. Kolda, "Multilinear Operators for Higher-Order Decompositions," Technical Report SAND2006-2081, Sandia Nat'l Laboratories, Albuquerque, N.M., Livermore, Calif., 2006.
[36] J.D. Carroll and J.J. Chang, "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of 'Eckart-Young' Decomposition," *Psychometrika,* vol. 35, pp. 283-319, 1970.
[37] L. De Lathauwer, B. De Moor, and J. Vandewalle, "On the Best Rank-1 and Rank-(R1, R2,..., RN) Approximation of Higher-Order Tensors," *SIAM J. Matrix Analysis and Applications,* vol. 21, no. 4, pp. 1324-1342, 2000.
[38] L.-H. Lim and V.D. Silva, "Tensor Rank and the Ill-Posedness of the Best Low-Rank Approximation Problem," *SIAM J. Matrix Analysis and Applications,* vol. 30, no. 3, pp. 1084-1127, 2008.
[39] H. Wold, "Soft Modeling: The Basic Design and Some Extensions," *Systems under Indirect Observation,* vol. 2, pp. 1-53, 1982.
[40] S. Wold, M. Sjostroma, and L. Erikssonb, "PLS-Regression: A Basic Tool of Chemometrics," *Chemometrics and Intelligent Laboratory Systems,* vol. 58, pp. 109-130, 2001.
[41] S. Wold, A. Ruhe, H. Wold, and W. Dunn III, "The Collinearity Problem in Linear Regression: The Partial Least Squares (PLS) Approach to Generalized Inverses," *SIAM J. Scientific and Statistical Computing,* vol. 5, pp. 735-743, 1984.
[42] B. Kowalski, R. Gerlach, and H. Wold, "Systems under Indirect Observation," *Chemical Systems under Indirect Observation,* pp. 191-209, 1982.
[43] A. McIntosh and N. Lobaugh, "Partial Least Squares Analysis of Neuroimaging Data: Applications and Advances," *Neuroimage,* vol. 23, pp. S250-S263, 2004.
[44] A. McIntosh, W. Chau, and A. Protzner, "Spatiotemporal Analysis of Event-Related fMRI Data Using Partial Least Squares," *Neuroimage,* vol. 23, no. 2, pp. 764-775, 2004.
[45] N. Kovacevic and A. McIntosh, "Groupwise Independent Component Decomposition of EEG Data and Partial Least Square Analysis," *NeuroImage,* vol. 35, no. 3, pp. 1103-1112, 2007.
[46] Z. Chao, Y. Nagasaka, and N. Fujii, "Long-Term Asynchronous Decoding of Arm Motion Using Electrocorticographic Signals in Monkeys," *Frontiers in Neuroengineering,* vol. 3, no. 3, 2010.
[47] L. Trejo, R. Rosipal, and B. Matthews, "Brain-Computer Interfaces for 1-D and 2-D Cursor Control: Designs Using Volitional Control of the EEG Spectrum or Steady-State Visual Evoked Potentials," *IEEE Trans. Neural Systems and Rehabilitation Eng.,* vol. 14, no. 2, pp. 225-229, June 2006.
[48] H. Kim, J. Zhou, H. Morse III, and H. Park, "A Three-Stage Framework for Gene Expression Data Analysis by L1-Norm Support Vector Regression," *Int'l J. Bioinformatics Research and Applications,* vol. 1, no. 1, pp. 51-62, 2005.
[49] Y. Nagasaka, K. Shimoda, and N. Fujii, "Multidimensional Recording (MDR) and Data Sharing: An Ecological Open Research and Educational Platform for Neuroscience," *PLoS ONE,* vol. 6, no. 7, p. e22561, 2011.
[50] J. Rickert, S. de Oliveira, E. Vaadia, A. Aertsen, S. Rotter, and C. Mehring, "Encoding of Movement Direction in Different Frequency Ranges of Motor Cortical Local Field Potentials," *J. Neuroscience,* vol. 25, no. 39, pp. 8815-8824, 2005.

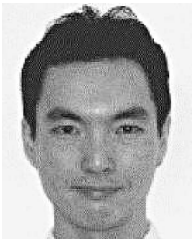


Qibin Zhao received the PhD degree in engineering from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a research scientist in the Laboratory for Advanced Brain Signal Processing at the RIKEN Brain Science Institute, Japan. His research interests include multiway data analysis, brain-computer interface, and machine learning. He is a member of the IEEE and the IEEE Computer Society.

Cesar F. Caiafa received the PhD degree in engineering from the Faculty of Engineering, University of Buenos Aires, in 2007. He is currently an adjunct researcher with the Argentinean Radioastronomy Institute (IAR)—CONICET and an assistant professor with the Faculty of Engineering, University of Buenos Aires. He is also a visiting scientist at the Laboratory for Advanced Brain Signal Processing, BSI, RIKEN, Japan. He is a member of the IEEE.

Danilo P. Mandic is a professor of signal processing at Imperial College London, London, United Kingdom, where he has been working in the area of nonlinear and multidimensional adaptive signal processing and time-frequency analysis. His publication record includes two research monographs titled *Recurrent Neural Networks for Prediction* (Wiley, August 2001) and *Complex Valued Nonlinear Adaptive Filters: Noncircularity, Widely Linear and Neural Models* (Wiley, April 2009), an edited book titled *Signal Processing for Information Fusion* (Springer, 2008), and more than 200 publications on signal and image processing. He has been a member of the IEEE Technical Committee on Machine Learning for Signal Processing, an associate editor for the *IEEE Transactions on Circuits and Systems II*, *IEEE Transactions on Signal Processing*, *IEEE Transactions on Neural Networks*, and the *International Journal of Mathematical Modelling and Algorithms*. He is a fellow of the IEEE.

Zenas C. Chao received the PhD degree in biomedical engineering from the Georgia Institute of Technology in 2007. He is currently a research scientist in the Laboratory for Adaptive Intelligence, RIKEN Brain Science Institute, Japan, where his research focus is on primate neurophysiological recording and analysis.



Yasuo Nagasaka received the PhD degree in psychology from Rikkyo University, Japan, in 2000. He was a postdoctoral fellow at the University of Iowa in 2004-2006. He is currently a postdoctoral fellow at the RIKEN Brain Science Institute, Japan. His research interests broadly span the areas of animal cognition, including illusory perception, motor control behavior, and social brain function.

Naotaka Fujii received the MD and PhD degrees from the Tohoku University School of Medicine, Sendai, Japan, in 1991 and 1997, respectively. After the PhD degree, he was with BCS, MIT, as a postdoctoral fellow. He is currently the head of the Laboratory for Adaptive Intelligence, RIKEN Brain Science Institute, Japan. His research interests span the wide areas of social neuroscience in primates and humans and development of brain-machine interfaces and substitute reality systems.

Liqing Zhang received the PhD degree from Zhongshan University, Guangzhou, China, in 1988. He was promoted to a full professor position in 1995 at the South China University of Technology. He was a research scientist at the RIKEN Brain Science Institute, Japan, from 1997 to 2002. He is now a professor in the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests cover computational theory for cortical networks, brain-computer interface, perception and cognition computing models, statistical learning, and inference. He has published more than 170 papers in international journals and conferences. He is a member of the IEEE and the IEEE Computer Society.

Andrzej Cichocki received the PhD and DrSc (Habilitation) degrees from the Warsaw University of Technology, Poland, both in electrical engineering. He is currently the senior team leader head of the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Japan. He is the coauthor of more than 250 technical papers and 4 monographs (two of them translated into Chinese). He is an associate editor of the *Journal of Neuroscience Methods* and the *IEEE Transactions on Signal Processing*. He is a fellow of the IEEE.