

## Research Article

# Radial Basis Function-Sparse Partial Least Squares for Application to Brain Imaging Data

Hisako Yoshida,<sup>1,2</sup> Atsushi Kawaguchi,<sup>3</sup> and Kazuhiko Tsuruya<sup>2</sup>

<sup>1</sup> Department of Biostatistics, Graduate School of Medicine, Kurume University, Kurume 8300011, Japan

<sup>2</sup> Department of Integrated Therapy for Chronic Kidney Disease, Graduate School of Medical Sciences, Kyushu University, Fukuoka 8118582, Japan

<sup>3</sup> Biostatistics Center, Kurume University, Kurume 8300011, Japan

Correspondence should be addressed to Atsushi Kawaguchi; [kawaguchi.atsushi@med.kurume-u.ac.jp](mailto:kawaguchi.atsushi@med.kurume-u.ac.jp)

Received 11 January 2013; Revised 27 March 2013; Accepted 29 March 2013

Academic Editor: Shigeyuki Matsui

Copyright © 2013 Hisako Yoshida et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Magnetic resonance imaging (MRI) data is an invaluable tool in brain morphology research. Here, we propose a novel statistical method for investigating the relationship between clinical characteristics and brain morphology based on three-dimensional MRI data via radial basis function-sparse partial least squares (RBF-sPLS). Our data consisted of MRI image intensities for multimillion voxels in a 3D array along with 73 clinical variables. This dataset represents a suitable application of RBF-sPLS because of a potential correlation among voxels as well as among clinical characteristics. Additionally, this method can simultaneously select both effective brain regions and clinical characteristics based on sparse modeling. This is in contrast to existing methods, which consider prespecified brain regions because of the computational difficulties involved in processing high-dimensional data. RBF-sPLS employs dimensionality reduction in order to overcome this obstacle. We have applied RBF-sPLS to a real dataset composed of 102 chronic kidney disease patients, while a comparison study used a simulated dataset. RBF-sPLS identified two brain regions of interest from our patient data: the temporal lobe and the occipital lobe, which are associated with aging and anemia, respectively. Our simulation study suggested that such brain regions are extracted with excellent accuracy using our method.

## 1. Introduction

Recently, brain morphometry research has gained considerable attention for its proposed utility in the early detection of dementia and assessment of regional cerebral atrophy. Furthermore, several authors have reported an association between brain morphology and clinical characteristics such as age, chronic disease, and genetics [1–3] using magnetic resonance imaging (MRI) data. Voxel-based morphometry (VBM) is a commonly used technique for such analyses [4]. This method is based on general linear models with the values of each MRI voxel (in units of pixels, preprocessed for standardization) as a dependent variable and clinical characteristics (including the group indicator variable and covariates) as explanatory variables. However, this approach has some drawbacks which have been discussed by Davatzikos [5]. For example, a multiple-comparison correction requires several assumptions that are difficult to verify. An alternative

to this approach is to use prespecified assemblies of voxels based on anatomical knowledge, which is known as a region of interest (ROI) approach. Therefore, the ROI approach requires the investigator to have precise and accurate knowledge of true anatomical boundaries. Moreover, variables need to be selected carefully in order to minimize the influence of irrelevant variables in the statistical model. We have taken a data-mining approach using an entire brain region and have used voxel intensity levels for the dependent variables and clinical characteristics (including patient background and blood test results) as explanatory variables.

There are two important statistical problems in the regression model that concern our use of large, complex data. The first is the selection of a set of relevant variables among a large number of both dependent and explanatory variables that are highly correlated. The partial least squares (PLS) regression, which was introduced by Wold [6], is a latent factor approach that is suitable for data with correlated variables. It has been

used as an alternative approach to ordinary least squares (OLS) regression in ill-conditioned linear regression models that arise in several disciplines such as chemistry, economics, and medicine [7, 8]. Tibshirani has used PLS in neuroimaging [9]. The second problem is a problem in variable selection that often arises when the sample size  $n$  is much smaller than total number of variables ( $p$ ; the so-called “large  $p$  small  $n$  problem”) for both dependent and explanatory variables. Utilizing the sparsity principle with  $L_1$ -penalty has been promoted as an effective solution [9, 10]. This version of sparse PLS (sPLS) combines the  $L_1$ -penalty and has been proposed by Lê Cao et al. [11] and Chun and Keleş [12]. The number of applications for this approach is steadily increasing in not only neuroimaging fields but also bioinformatics and chemometrics. This technique produces sparse, linear combinations of the explanatory variables and achieves both dimension reduction and variable selection simultaneously. The pioneering application of this method to brain imaging data has been used to investigate genetic polymorphisms and functional imaging data [3]. However, it is based on PLS regression in its symmetric (also called canonical) mode. In this paper, we consider the PLS in its regression mode based on the singular value decomposition (PLS-SVD). The difference lies in the fact that factors are orthogonal in the canonical mode, contrary to PLS-SVD, in which the loadings are orthogonal. The main concern in this approach is the restriction of analysis to prespecified brain regions. Using brain regions that have not been specified *a priori* would be more data-driven approach that may yield new and unexpected results, but such approaches typically introduce computational difficulties because of the large number of voxels to be analyzed. For this reason, we decided to combine this approach with a first step of dimension reduction on brain images using basis expansion.

In this paper, we propose a sparse PLS approach with basis expansion (RBF-sPLS; radial basis function-sparse partial least squares) and provide an application for real data using three-dimensional MRI brain scans with about a million voxels and 73 clinical characteristics from chronic kidney disease (CKD) patients. In addition, we conducted a simulation study to compare our proposed method with the original method. Our proposed, RBF-sPLS, prediction model with dimension reduction devices offers discriminant functions with excellent prediction performance in terms of sensitivity and specificity.

This paper is organized as follows. Section 2 provided a discussion of three-dimensional MRI data and their preprocessing. Section 3 states the proposed statistical methods. In Section 4, we report a simulation study for the characteristics of sPLS with basis expansion (RBF-sPLS) or without it (sPLS).

## 2. Data

**2.1. Subjects.** Between 2009 and 2012, we recruited 102 patients (mean age:  $61 \pm 11$  years, 52% male, 48% female) with chronic kidney disease (CKD) to participate in our study. We examined brain volume using MRI scanning, and clinical parameters were measured on the same day. Patients were

TABLE 1: The clinical characteristics about CKD patients’ dataset.

	Mean $\pm$ SD
Number (male/female)	102 (49/53)
Age (years old)	$61 \pm 11$
Diabetes (%)	27 (27)
BMI <sup>a</sup> (kg/m <sup>2</sup> )	$24.0 \pm 3.9$
SBP <sup>b</sup> (mmHg)	$124 \pm 16$
DBP <sup>c</sup> (mmHg)	$70 \pm 12$
eGFR <sup>d</sup> (mL/min/1.73 m <sup>2</sup> )	$39.8 \pm 13.6$
Smoker ( $n$ [%])	56 (56)

<sup>a</sup>Body mass index. <sup>b</sup>Systolic blood pressure.

<sup>c</sup>Diastolic blood pressure. <sup>d</sup>Estimate glomerular filtration rate.

eligible if they were between 20 and 80 years old and had no prior history of brain injury such as stroke, traumatic brain injury, or brain tumor. The participant characteristics are shown in Table 1. Fifty-five percent of participants had a history of smoking (47 former and 9 current smokers). Blood pressure in the brachial artery was measured with the subjects in a sitting position after a 10 min rest. All patients provided informed consent. Kyushu University Institutional Review Board approved all procedures.

**2.2. Image Data.** Brain MRI was acquired from each subject using a 3.0 tesla MRI scanner of the same model. No major hardware upgrades occurred during the period. All subjects were scanned with identical pulse sequences: 124 contiguous, 3.0 mm thick axial planes of three-dimensional T1-weighted images (spoiled gradient recalled acquisition in steady state: echo time, 7 ms; flip angle, 30; voxel size,  $1.02 \times 1.02 \times 1.5$  mm).

We used the Statistical Parametric Mapping 8 software (SPM8, Wellcome Department of Cognitive Neurology, London, UK) to preprocess brain images. The segmentation algorithm from SPM8 was applied to every T1-weighted MRI scan to extract tissue maps corresponding to gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). The temporary common space of rigidly registered tissues is necessary as a starting point for the DARTEL algorithm. Next, the segmented tissues maps were used to create a custom template and associated warping fields were generated using the DARTEL template creation tool [4]. This tool estimates a best set of smooth, nonlinear deformations from each subject’s tissues to their common average, applies the deformations to create a new average, and then reiterates until convergence.

## 3. Methods

**3.1. Basis Expansion-Based Dimension Reduction.** Suppose that we have  $n$  independent subjects  $\{(\mathbf{x}_a, \mathbf{s}_a); a = 1, \dots, n\}$ , where  $\mathbf{x}_a \in \mathbb{R}^p$  are  $p$ -dimensional exploratory variable vectors (clinical characteristics) and  $\mathbf{s}_a = (s_a(\mathbf{w}_1), \dots, s_a(\mathbf{w}_N))'$  are  $N$ -dimensional vectors of brain images for the  $a$ th subject defined at points  $\mathbf{w}_i \in \mathbb{Z}^3$  ( $i = 1, \dots, N$ ). We used the radial

B-spline function  $\phi(\cdot)$  [13] to reduce the dimension, which is represented as follows. For given  $h \geq 0$ ,

$$\phi(u) = \frac{1}{4h^2} \begin{cases} h^3 + 3h^2(h-u) \\ \quad + 3h(h-u)^2 - 3(h-u)^3, & (u \leq h), \\ (2h-u)^3, & (h < u \leq 2h), \\ 0, & (u > 2h), \end{cases} \quad (1)$$

where  $u \geq 0$ . We used the distance of these knots to define  $h$  as  $h = \sqrt{3 \times h_0^2}$ , where  $h_0$  is the distance between adjacent knots. Then, the  $N \times q$  dimension reduction matrix,  $\mathbf{B}$ , is defined with the  $(l, m)$ -component being  $b_{lm} = \phi(\|\mathbf{w}_l - \mathbf{k}_m\|)$ , where  $\mathbf{k}_m \in \mathbb{Z}^3$  ( $m = 1, \dots, q$ ) are equally spaced knots. Note that the value of  $b_{lm}$  is proportional to distance of  $\mathbf{w}_l$  from  $\mathbf{k}_m$ . Therefore,  $\mathbf{S} = (\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n)'$  and the dependent variable matrix,  $\mathbf{Y}$ , is constructed as

$$\mathbf{Y} = \mathbf{S}\mathbf{B}. \quad (2)$$

Thus, for PLS regression, our response and predictor matrices are  $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n)' \in \mathbb{R}^{n \times q}$  and  $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)' \in \mathbb{R}^{n \times p}$ , respectively.

**3.2. Sparse Partial Least Squares.** Let  $\mathbf{Y}$  denote an  $n \times q$  dependent variable matrix and let  $\mathbf{X}$  denote an  $n \times p$  explanatory variable matrix. The core assumption of PLS regression is a latent decomposition of  $\mathbf{Y}$  and  $\mathbf{X}$  as follows:

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F}, \quad \mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}, \quad (3)$$

where  $\mathbf{T}$  is an  $n \times k$  score matrix,  $k$  is the number of components,  $\mathbf{P}$  and  $\mathbf{Q}$  are  $p \times k$  and  $q \times k$  loading matrices, and  $\mathbf{E}$  and  $\mathbf{F}$  are  $n \times p$  and  $n \times q$  matrices of random errors.

The version of sparse PLS (sPLS) regression proposed by Lê Cao et al. [11] invokes singular value decomposition (SVD) of  $\mathbf{M} = \mathbf{X}'\mathbf{Y}$  to yield the  $\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}'$ , where  $\mathbf{U}$  is a  $p \times k$  orthogonal matrix,  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_k)$  with  $d_1 \geq d_2 \geq \dots \geq d_k$ , and  $\mathbf{V}$  is an  $M \times V$  orthogonal matrix. Among variations of PLS regression, this is called PLS-SVD. From these, we can obtain the regression form  $\mathbf{Y} = \mathbf{X}\mathbf{C} + \mathbf{G}$  where  $\mathbf{C}$  is a  $p \times q$  regression coefficient matrix given by  $\mathbf{C} = \mathbf{U}(\mathbf{P}'\mathbf{U})^{-1}\mathbf{Q}'$  and  $\mathbf{G}$  is a residual matrix.

For ease of explanation for estimation, suppose that  $k = 1$ , then the objective function with a  $L_1$  penalization on  $\mathbf{u}$  and  $\mathbf{v}$ , which are column vectors of  $\mathbf{U}$  and  $\mathbf{V}$ , respectively, is given as follows:

$$L(\mathbf{u}, \mathbf{v}) = -\mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{v} + \lambda_X \|\mathbf{u}\|_1 + \lambda_Y \|\mathbf{v}\|_1, \quad (4)$$

where  $\lambda_X$  and  $\lambda_Y$  are  $L_1$  penalization parameters for the weight vectors of matrices  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively. This function is a minimized subject to  $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$ . The amplitudes of  $\lambda_X$  and  $\lambda_Y$  correspond to the increases and decreases of the number of  $\mathbf{X}$  and  $\mathbf{Y}$  variables, which contribute to the regression. For example, in the case of  $\mathbf{X}$ , if the value of  $\lambda_X$  is large, then a large number of variables  $\mathbf{X}$  would be selected. The same is the case for  $\mathbf{Y}$ . Therefore,

the sPLS concerns selection and modeling in a one-step procedure. This optimization problem is performed by the soft-thresholding function  $g_\lambda(y) = \text{sign}(y)(|y| - \lambda)_+$ , where  $(a)_+ = \max(0, a)$  at each iteration of the NIPALS inner loop. Weight vectors  $\mathbf{u}$  and  $\mathbf{v}$  are computed using the following algorithm.

- (1) Initialize  $\mathbf{u}$  and  $\mathbf{v}$  using, for instance, the first pair of singular vectors of the matrix  $\mathbf{X}'\mathbf{Y}$  and normalize  $\mathbf{u} \leftarrow \mathbf{u}/\|\mathbf{u}\|_2$  and  $\mathbf{v} \leftarrow \mathbf{v}/\|\mathbf{v}\|_2$ .
- (2) Until convergence of  $\mathbf{u}$  and  $\mathbf{v}$ :
  - (a) for fixed  $\mathbf{v}$ ,  $\hat{\mathbf{u}} = g_{\lambda_X}(\mathbf{X}'\mathbf{Y}\mathbf{v})$  and normalize  $\hat{\mathbf{u}}$  as in step 1;
  - (b) for fixed  $\mathbf{u}$ ,  $\hat{\mathbf{v}} = g_{\lambda_Y}(\mathbf{Y}'\mathbf{X}\mathbf{u})$  and normalize  $\hat{\mathbf{v}}$  as in step 1;
  - (c)  $\mathbf{u} = \hat{\mathbf{u}}$ ,  $\mathbf{v} = \hat{\mathbf{v}}$ .
- (3)  $\mathbf{t} = \mathbf{X}\mathbf{u}$ ,  $\mathbf{p} = \mathbf{X}'\mathbf{t}/\mathbf{t}'\mathbf{t}$ , and  $\mathbf{q} = \mathbf{Y}'\mathbf{t}/\mathbf{t}'\mathbf{t}$ , where  $\mathbf{t}$ ,  $\mathbf{p}$ , and  $\mathbf{q}$  correspond to column vector of  $\mathbf{T}$ ,  $\mathbf{P}$ , and  $\mathbf{Q}$ , respectively.

For the general case of  $k > 1$ , the above algorithm is repeated for  $k$  times with the deflation step  $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}\mathbf{p}'$  and  $\mathbf{Y} \leftarrow \mathbf{Y} - \mathbf{t}\mathbf{q}'$  as the fourth step. The final solution can be obtained as  $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_k)$ ,  $\mathbf{Q} = (\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_k)$ , and  $\mathbf{P} = (\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_k)$ , where the elements are obtained at each step among  $k$  steps.

**3.3. Choice of Tuning Parameters.** The choice of penalization parameters  $\lambda_X$ ,  $\lambda_Y$  and number of components  $k$  is important in model construction. We use a criteria called  $Q^2$  proposed by Tenenhaus [14], which were used to select the number of components in the sPLS model in Lê Cao et al. [11] by performing cross-validation. We used 10-fold cross validation. Thus, our  $Q^2$  has a functional form of  $\lambda_X$ ,  $\lambda_Y$ , and  $k$  and is defined as

$$Q^2(\lambda_X, \lambda_Y, k) = 1 - \frac{\sum_{j=1}^q \text{PRESS}_{jk}}{\sum_{j=1}^q \text{RSS}_{j(k-1)}}, \quad (5)$$

where  $\text{PRESS}_{jk}$  is the prediction error sum of squares and  $\text{RSS}_{jk}$  is the residual sum of squares for the  $j$ th-dependent variable and the PLS model with  $k$  components defined as follows. Let  $\kappa: \{1, 2, \dots, n\} \rightarrow \{1, 2, \dots, 10\}$  be an indexing function that indicates the partition to which observation  $i$  is allocated to  $\kappa(i)$ th part of the data by the randomization:

$$\text{PRESS}_{jk} = \sum_{i=1}^n (y_{ij} - \hat{y}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k))^2, \quad (6)$$

$$\text{RSS}_{jk} = \sum_{i=1}^n (y_{ij} - \hat{y}_{ij}(\lambda_X, \lambda_Y, k))^2,$$

$\hat{y}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k)$  is the predicted value for the  $j$ th-dependent variable from the sPLS model with penalization parameters  $\lambda_X$  and  $\lambda_Y$  and number of components  $k$  and estimated weight vectors from  $\kappa(i)$ th part of the data removed.

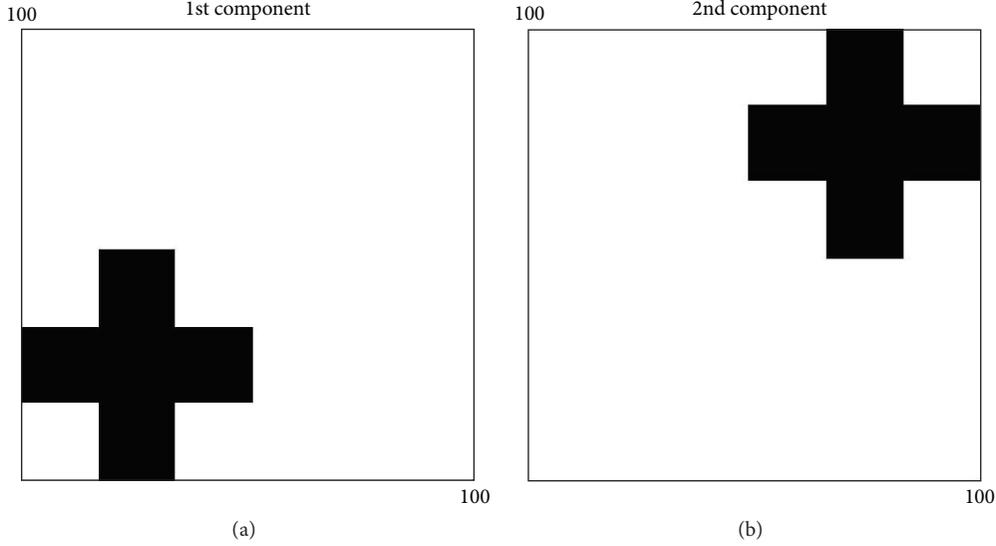


FIGURE 1: True grayscale images.

That is, for any  $i$  subject, we predict that  $\hat{y}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k) = \mathbf{x}_i' \hat{\mathbf{b}}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k)$ , where  $\hat{\mathbf{b}}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k)$  is the  $j$ th column of estimated regression coefficient matrix  $\hat{\mathbf{B}}$  from the sPLS model with penalization parameters  $\lambda_X$  and  $\lambda_Y$  and number of components  $k$  and  $\kappa(i)$ th part of the data removed.  $\hat{y}_{\kappa(i)j}(\lambda_X, \lambda_Y, k)$  is the predicted value with the same definition as  $\hat{y}_{(-\kappa(i))j}(\lambda_X, \lambda_Y, k)$  except for the estimated weight vector from all available  $n$  subjects. We select the optimal set  $(\lambda_X, \lambda_Y, k)$  based on the maximization of  $Q^2(\lambda_X, \lambda_Y, k)$  among given candidates. This is implemented by the grid search.

#### 4. Simulation Studies

In this section, we will illustrate the proposed methods in a simulation study. We demonstrate the impact of knot distance in affecting the representation of the results and clarify the advantage of dimension reduction by RBF by comparison to the method without basis expansion.

**4.1. Data Sets.** Consider  $n$  patients and  $p$  explanatory variables. We generated 100 datasets according to the following sPLS model with two components

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)', \quad \mathbf{x}_j \sim \text{MVN}(\mathbf{0}, \Sigma), \quad (7)$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}' + \mathbf{F} \quad \text{with } \mathbf{T} = \mathbf{X}\mathbf{P}^-, \quad \mathbf{F} \sim \text{MVN}(\mathbf{0}, \mathbf{I}),$$

where  $\text{MVN}(\mathbf{0}, \Sigma)$  denoted  $p$ -dimensional multidimensional normal distribution with zero mean and variance covariance matrix  $\Sigma$ .  $\mathbf{P} = (\mathbf{P}_1, \mathbf{P}_2)$  is the  $p \times 2$  matrix with  $\mathbf{P}_1 = \mathbf{b} \otimes (\mathbf{1}'_{p/20}, \mathbf{0}'_{3p/20})'$ ,  $\mathbf{P}_2 = \mathbf{b} \otimes (\mathbf{0}'_{p/20}, \mathbf{1}'_{p/20}, \mathbf{0}'_{p/10})'$ , and  $\mathbf{b} = (5, 2, 1, -2, -5)'$ , where  $\otimes$  is the Kronecker product,  $\mathbf{0}_c$  is a  $c$ -dimensional vector with all elements 0, and  $\mathbf{1}_c$  is a  $c$ -dimensional vector with all elements 1.  $\mathbf{Q}$  is the  $q \times 2$  matrix whose columns are vectorized true images displayed in

Figure 1. The images can be thought of as 2D grayscale images with pixel intensities on the  $[0, 1]$  scale. The black pixels are set to 1 and the white ones are set to zero.

We performed this step in order to assess how much the performance of sPLS is influenced by the basis expansion and by the number of clinical parameters kept by the filter and to select the best pair of parameters. We provided a comparison with the original method (sPLS without the basis expansion) and also analyzed the impact of the distance between adjacent knots in our method for  $h = 2, 4$ , and  $8$ . We tested our pattern of data set;  $n = 50/p = 40$ ,  $n = 50/p = 80$ ,  $n = 100/p = 40$ , and  $n = 100/p = 80$  to replicate the sample size  $n$  of the CKD patients data set and the number of covariates  $p$ . The images  $\mathbf{Y}$ 's were unfolded to obtain vectors of size  $q = 100 \times 100 = 10,000$ .

**4.2. Results.** We estimated  $\mathbf{P}$  and  $\mathbf{Q}$  from simulated data by the method described in Section 3. All results yielded the correct number of components. We computed the probability images by averaging up the estimated  $\mathbf{Q}$ 's from 100 datasets. The middle and bottom panels of Figure 2 display binary images converted from probability images with threshold 0.95 for the first and second components, respectively, in the case of  $n = 50/p = 40$ . The top of Figure 2 shows the combined true image. The result for sPLS without the basis expansion showed nothing at all because the maximum probability calculated was 0.7. On the other hand, the sPLS with the basis expansion with distance between knots  $h = 2$  had a good shape, while for  $h = 4$  and  $8$ , the true image could not be reconstructed.

To assess how effectively the estimated model predicts each variable, sensitivity, specificity, and c-index = sensitivity - (1 - specificity) were computed and averaged over 100 sets. As shown in Table 2, the mean values of the c-index for the proposed method with  $h = 2$  were relatively smaller than those for the method without the basis expansion and

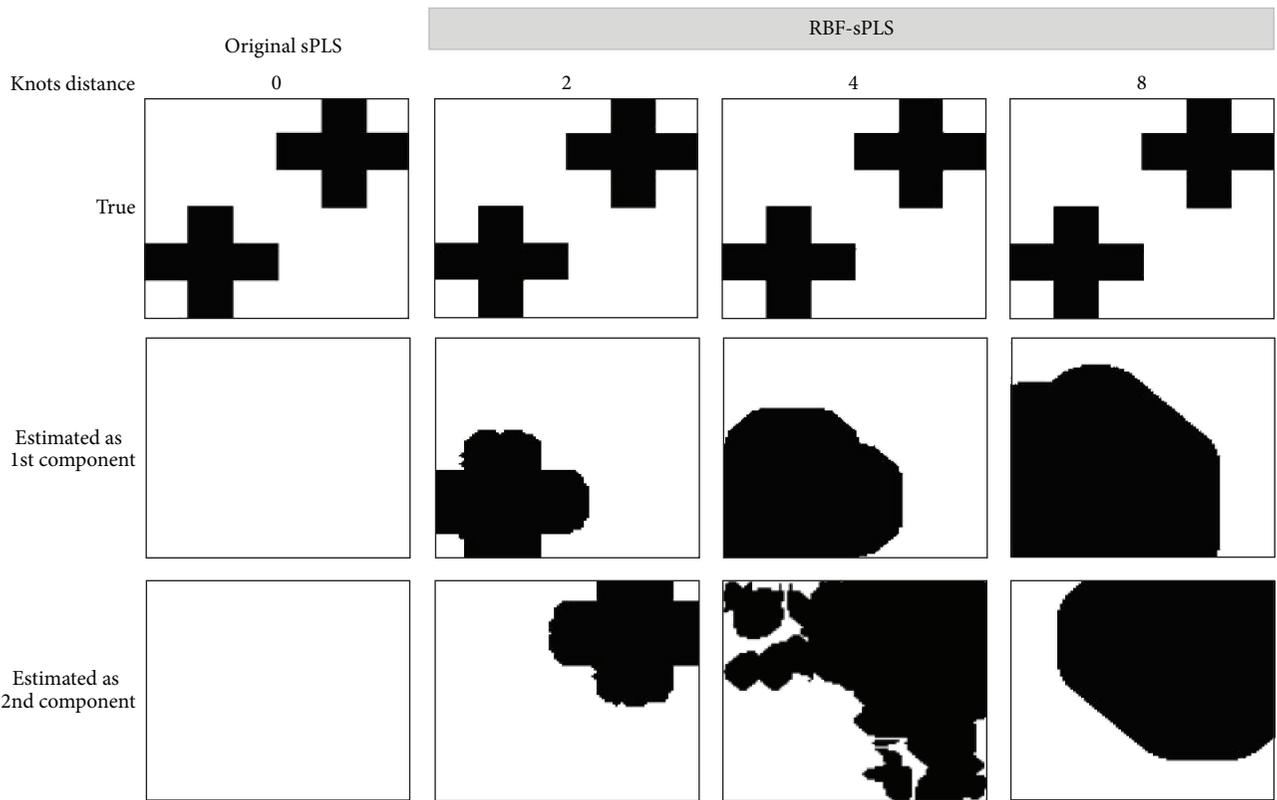


FIGURE 2: Binary images with threshold 0.95 for probability images from the simulation result of sPLS models with (knots distance = 2, 4, 8) or without basis expansion (knots distance = 0) for  $n = 50$  and  $p = 40$ .

TABLE 2: The result for sPLS without basis expansion and with, respectively, for 100 simulated data sets.

Knots distance	$P$	$n$	1st component			2nd component		
			Sensitivity	Specificity	C-index	Sensitivity	Specificity	C-index
Original sPLS: without basis expansion								
0	40	50	0.26	0.99	0.25	0.30	0.99	0.29
		100	0.34	0.99	0.33	0.39	0.99	0.38
	80	50	0.37	0.99	0.36	0.43	1.00	0.43
		100	0.39	0.99	0.38	0.44	1.00	0.44
RBF-sPLS: with basis expansion								
2	40	50	1.00	0.60	0.60	1.00	0.68	0.68
		100	1.00	0.86	0.86	1.00	0.87	0.87
	80	50	1.00	0.73	0.73	1.00	0.75	0.75
		100	1.00	0.84	0.84	1.00	0.88	0.88
4	40	50	1.00	0.29	0.29	1.00	0.13	0.13
		100	1.00	0.29	0.29	1.00	0.04	0.04
	80	50	1.00	0.27	0.27	1.00	0.13	0.13
		100	1.00	0.21	0.21	1.00	0.08	0.08
8	40	50	1.00	0.08	0.08	1.00	0.06	0.06
		100	1.00	0.05	0.05	1.00	0.00	0.00
	80	50	1.00	0.05	0.05	1.00	0.01	0.01
		100	1.00	0.02	0.02	1.00	0.00	0.00

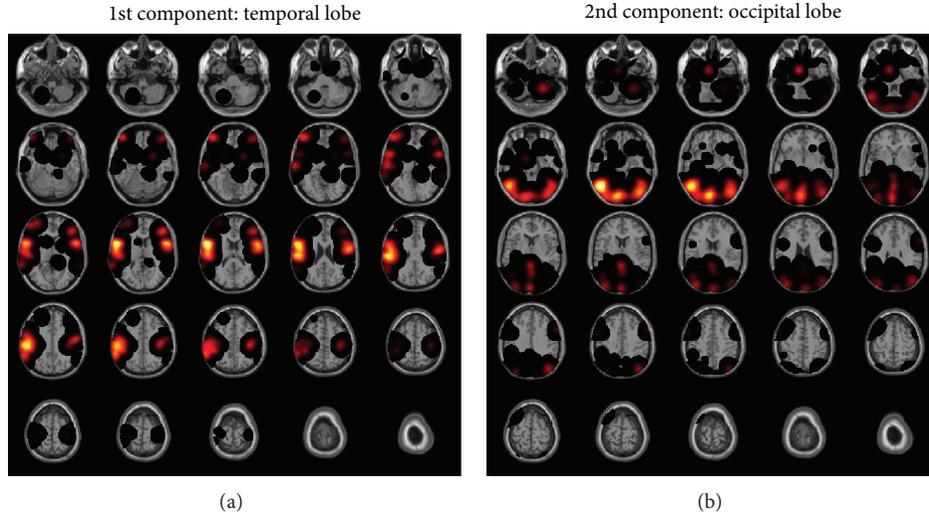


FIGURE 3: The brain region linked groups of each component.

$h = 4$  and 8 in any cases of  $n$  and  $p$ . This indicated that the proposed method performed better than the original, and the distance between knots took on the smallest possible value.

## 5. Real Data Application

We applied sPLS with basis expansion to our MRI dataset of the CKD patients described in Section 2. We assessed additional demographic and health-related variables, as well as laboratory data obtained on the same day. These data were used as covariates in our statistical analyses. The number of covariates is  $p = 73$ . Among the 2,122,945 ( $121 \times 145 \times 121$ ) voxels for one subject, the voxels that represent brain regions are extracted, resulting in 839,089 voxels. The dimension of the basis function is  $q = 13,047$  because of the 4-voxel ( $h_0 = 4$ ; therefore,  $h = \sqrt{3 \times 4^2} = 6.93$ ) equal spacing knots. The number of components was selected as  $k = 2$ . The number of selected variables in the first component of  $\mathbf{X}$  was 17, and 14 variables were in the second component. For  $\mathbf{Y}$ , 785 and 947 variables were selected in the same manner. Figure 3 shows the results by the axial view of brain. The left image shows the coefficient image estimated as the first component. Similarly, the right one shows the second component.

Our model revealed a relatively strong association between the bilateral temporal lobes and clinical markers of chronic kidney disease. The temporal lobes are one of the four main regions of the cerebral cortex. Structures of the limbic system, including the olfactory cortex, amygdala, and the hippocampus are located within the temporal lobes. The temporal lobes play an important role in organizing sensory input, auditory perception, language and speech production, and memory association and formation. These regions linked the 17 factors, in particular, age, sex, underlying disease (diabetes mellitus), smoking status, weight, serum albumin level, serum creatinine, total cholesterol, glucose, HDL-cholesterol, LDL-cholesterol, glycoalbumin, cholinesterase,

number of red blood cell, whole parathyroid hormone, pulse wave velocity, and coronary artery calcification score.

The occipital lobes were selected by our analysis as the second component. The occipital lobes are positioned at the back region of the cerebral cortex and are the main centers for visual processing, involved in several functions of the body including visual perception and color recognition. This region linked the following factors: sex, body height, body weight, diastolic blood pressure, ratio of toe to brachial systolic blood pressure, total bilirubin, glucose, chloride, serum iron levels, number of red blood cells, hemoglobin, hematocrit, plasminogen activator inhibitor-1, and transferrin saturation.

The variables selected as the first component are considered to be the factors most closely related to aging and arterial stiffness, while those associated with the second region are more closely related to markers of anemia. The extent of atherosclerosis, calcification, and renal anemia are important complications in CKD patients. Recently, these factors have been suggested to be involved in brain atrophy and depressed cerebral oxygen metabolism [15, 16], but its mechanism remains to be elucidated. We also found a significant correlation between regional gray matter volume and hemoglobin level after adjusting for age, gender, residual renal function, underlying kidney disease, history of smoking, diastolic blood pressure, and LDL cholesterol level using multiple-linear regression methods [17]. In this analysis, we used only the whole gray matter volume as an objective variable, because multiple variables cannot be applied to conventional linear regression models, whereas the sPLS could select variables and modeling in a one-step procedure and use many objective variables.

## 6. Discussion

This paper describes that the radial basis function-sparse partial least squares (RBF-sPLS) technique was proposed

and was applied to high dimensional brain imaging data. The original sPLS is a useful regression model to analyze data in which both dependent and explanatory variables are multivariate and correlated with one another. The most difficult problem in analyzing real brain data is the high dimensionality of these datasets. While prespecified regions were used in previous neuroimaging analyses, our method successfully handled a whole brain region following the basis expansion. The basis function has a spherical shape, but it was able to approximate the cross shape used in the simulation study. This would be expected because of the narrow spanned knots location. Thus, we set as close knots each other as possible in the real-data application, using 4-voxel equal spacing knots because computation using 2-voxel spacing was not possible. This method may be applicable to not only real brain data, but also general imaging datasets, because actual lesions would cause aggregates in adjacent voxels. Although the relative advantage of our proposed method was shown through the comparison between simulations run with and without the basis function conducted in the fair setting, further simulation studies with more realistic constraints are necessary. However, these simulations lie beyond the scope of the present paper and will be dealt with in the future. The significance of this study is to clarify the characteristics of RBF-sPLS presented visually for the analysis of imaging data.

We obtained clinically relevant findings about the relationship between aging, anemia, and brain morphology from the real-data application in our study. We are currently in the process of collecting longitudinal data and normal controls to expand this confirmatory evidence for future work. In summary, RBF-sPLS can help revealing the relationships between complex, large datasets, including brain imaging data.

## Conflict of Interests

The authors declare no financial/commercial conflict of interests regarding the study.

## Acknowledgments

The authors are deeply grateful to the referees for fruitful suggestions, especially for suggesting better terms and sentences. This work was supported by the center of medical checkup in Kyushu University. This research was supported in part by Intramural Research Grant (24-10) for Neurological and Psychiatric Disorders of NCNP and Grants-in-Aid from the Ministry of Education, Culture, Sport, Science and Technology of Japan (24700286 to A. Kawaguchi).

## References

- [1] Y. Taki, S. Kinomura, K. Sato et al., "Relationship between body mass index and gray matter volume in 1,428 healthy individuals," *Obesity*, vol. 16, no. 1, pp. 119–124, 2008.
- [2] C. M. Falvey, C. Rosano, E. M. Simonsick et al., "Macro- and microstructural magnetic resonance imaging indices associated with diabetes among community-dwelling older adults," *Diabetes Care*, vol. 36, no. 3, pp. 677–682, 2013.
- [3] E. Le Floch, V. Guillemot, V. Frouin et al., "Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares," *NeuroImage*, vol. 15, pp. 11–24, 2012.
- [4] J. Ashburner and K. J. Friston, "Voxel-based morphometry—the methods," *NeuroImage*, vol. 11, no. 6, part 1, pp. 805–821, 2000.
- [5] C. Davatzikos, "Why voxel-based morphometric analysis should be used with great caution when characterizing group differences," *NeuroImage*, vol. 23, no. 1, pp. 17–20, 2004.
- [6] H. Wold, "Estimation of principal components and related models by iterative least squares," *Journal of Multivariate Analysis*, pp. 391–420, 1966.
- [7] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, and C. L. Grady, "Spatial pattern analysis of functional brain images using partial least squares," *NeuroImage*, vol. 3, no. 3, pp. 143–157, 1996.
- [8] S. de Jong and C. T. Braak, "Comments on the PLS kernel algorithm," *Journal of Chemotherapy*, vol. 8, no. 2, pp. 169–174, 1994.
- [9] R. J. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, pp. 267–288, 1996.
- [10] B. Efron, T. Hastie, I. Johnstone, and R. J. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, pp. 407–451, 2004.
- [11] K. A. Lê Cao, D. Rossouw, and C. Robert-Granié, "A sparse PLS: variable selection when integrating omics data," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, article 35, 2008.
- [12] H. Chun and S. Keleş, "Sparse partial least squares regression for simultaneous dimension reduction and variable selection," *Journal of the Royal Statistical Society Series B*, vol. 72, no. 1, pp. 3–25, 2010.
- [13] A. Saranli and B. Baykal, "Complexity reduction in radial basis function (RBF) networks by using radial B-spline functions," *Neurocomputing*, vol. 18, no. 1–3, pp. 183–194, 1998.
- [14] M. Tenenhaus, *La Régression PLS. théorie et Pratique*, Technip, Paris, France, 1998.
- [15] Y. Yakushiji, Y. Nanri, T. Hirotsu et al., "Marked cerebral atrophy is correlated with kidney dysfunction in nondisabled adults," *Hypertension Research*, vol. 33, no. 12, pp. 1232–1237, 2010.
- [16] H. Hirakata, H. Kanai, and H. Nakane, "Depressed cerebral lxygen metabolism in patients with chronic renal failure: a positron emission tomography study," *Journal of the Japanese Society for Dialysis Therapy*, vol. 34, no. 7, pp. 1149–1155, 2001.
- [17] K. Tsuruya, H. Yoshida, and T. Kitazono, "Possible contribution of anemia to brain atrophy in predialysis patients with chronic kidney disease," *Journal of the American Society of Nephrology Supplement A*, vol. 667, 2012, SA-PO141.