Interpretable machine learning approaches to high-dimensional date and their applications to biomedical engineering problems

Kosuke Yoshida
(Graduate School of Informatics at Kyoto University)

March 2018

Interpretable machine learning approaches to high-dimensional data and their applications to biomedical engineering problems

Kosuke Yoshida

Contents

Abstract

Machine learning techniques have led to developments of algorithms that can learn from datasets alone to perform various pattern recognition, classification, and prediction tasks under rapid increase in computational power. In the field of biomedical engineering, they are expected to be promising solutions to deal with data growth associated with high-throughput experiments and high-resolution imaging techniques. However, there are still difficulties in applying machine learning techniques to the large amount of high-dimensional data.

The first is about high-dimensionality, that is, the number of features may be much larger than that of samples while the most of features could be irrelevant to the phenotypes. In this situation, imposing sparsity-inducing regularization on machine learning models is able to remove irrelevant features automatically and effective in yielding insights into underlying biology to allow experimentalists to have a new working hypothesis. The next is about co-linearity, that is, a feature is linearly correlated with others, leading to unstable estimation of model parameters. One of the solutions is to apply dimension reduction that maps the original dataset onto a low dimensional space while keeping most of important information, under assuming that the entire dataset has been generated by a dominant system with a fewer dimensionality. The last is a nonlinearity inherent in biological phenomena. Since many biological phenomena could be nonlinear, it is worth introducing a nonlinear assumption into the models. Here, kernel-based methods are used for this purpose which deal with various types of nonlinear transformation within a unified framework.

In this thesis, I explored interpretable machine learning methods that can be applied to high-dimensional data with a particular interest in analyses of human brain activities, and moreover, investigated a novel method for multimodal analyses that can be seen in the biomedical engineering field.

In chapter 3, I focused on diagnosis of depression based on human functional magnetic resonance imaging (fMRI). In order to achieve both accurate diagnosis and identifying relevant anatomical regions for depression, I introduced region-wise sub-kernels that corresponded one-to-one to anatomical brain regions, and an associated learning method as an extension of multiple kernel learning. This method achieved reasonably good accuracy in terms of leave-one-out cross-validation and also identified a restricted number of anatomical regions which will be profitable for depression diagnosis in the future study.

In chapter 4, I predicted a number of clinical scores from resting-state functional connectivity, with an interest in knowing the relation between the psychiatry scores and patient ' s brain activities. To ease the effects from co-linearity in the functional connectivity, I applied partial least squares regression (PLSR) and its kernel variants. As a result, I successfully demonstrated that they provided significantly better prediction of psychiatry scores than that by ordinary linear regression after applying a low-dimensional feature extraction.

In chapter 5, I examined the problems that often occur when nonlinear correlation analyses are applied to high-dimensional data. I introduced a novel method called two-stage kernel canonical correlation analysis, in order to introduce an appropriate design of kernels within the framework of multiple kernel learning. Using synthetic datasets, I confirmed that this method enabled us to remove irrelevant features. An application to gene expression analysis for the mice metabolic system is also demonstrated.

As a conclusion, I explored machine learning algorithms that allowed us to interpret the obtained results and to simultaneously deal with co-linearity and nonlinearity in high-dimensional datasets. Specifically, I focused on the relationship between clinical diagnosis, scores, and fMRI data of depression patients with an attempt to reveal a physiological basis of traditional psychological

evaluations in chapters 3 and 4. Moreover, to obtain deeper understanding of biological phenotypes in a data driven manner, I introduced a novel method based on multiple kernel learning in chapter 5, which enabled us to obtain non-linear associations in the given biomedical data. Thus, the studies shown in this thesis could be new methodologies for dealing with high-dimensional and multimodal data in the field of biomedical engineering.

1 Introduction

1.1 Background

Machine learning techniques, which grew out of computational statistics, develop algorithms that can learn from datasets to make pattern recognitions, classifications, and predictions under rapid increase of computer power. In the field of biomedical engineering, they are expected to be promising solutions for biologist and medical scientists since the amount of data is rapidly increasing due to appearance of high-throughput experiments and high resolution imaging techniques. However, applying machine learning techniques to the large amount of data is not always straightforward. One of difficulties is high-dimensionality, where the number of features p is much larger than that of samples N. In this situation, conventional use of statistics including multivariate analysis doesn't work properly. Another difficulty is about interpretability. In this field, a result of machine learning techniques is required to yield some new insights about underlying biology to lead experimentalists to a new working hypothesis.

Let us consider a concrete example of machine learning applications to associate behavioral responses with brain activity data measured by functional magnetic resonance imaging (fMRI). In the framework of machine learning, this application is formulated as a prediction problem to build a model that can assign a newly measured brain activity data to behavioral responses automatically. Note that brain activity data could be high-dimensional data since fMRI techniques have high spatial resolution. There are a number of important topics and I mention three of them.

The first is that brain activity in most of anatomical regions could be irrelevant to the behavioral responses of interest due to theory of localization of brain function. In this case, it is natural to assume that only a small number of anatomical regions are contributing to the prediction and the rest are not.

Therefore, it is desirable to apply a method that can automatically identify contributing anatomical regions and remove the irrelevant regions in the learning process. In the field of machine learning, it is useful to impose sparsity-inducing penalty such as L1 regularization that focuses on a small number of features that contribute to the outputs. This procedure can provide interpretability of the model that could guide us to some new insights about underlying relationship between brain activity and the behavioral responses.

The next is that brain activity in many regions could be strongly correlated when they have strong coherence in their activity. In terms of machine learning, this causes a phenomenon called co-linearity that leads to unstable estimation of parameters. One of the solutions is to apply dimension reduction that transfers the original dataset into a low dimensional space while keeping most of important information. Note that the underlying assumption is that the entire dataset is generated by a few dominant systems and identifying the systems is beneficial for interpretation of the result.

The last is nonlinearity in biological phenomenons. Since many biological phenomenons could be nonlinear, it is worth incorporating a nonlinear assumption into the learning algorithms. In addition, a framework that can deal with the various types of data, such as graph data, string data, and so on, is expected in this field. Here, the method frequently used for this purpose is kernel method that deals with various types of nonlinear transformation in a unified framework.

In this thesis, I explore interpretable machine learning in high-dimensional data for analysis of brain activity to solve the above mentioned problems and investigate a novel method for multimodal analysis.

1.2 Structure

This thesis is organized as follows.

In chapter 2, I briefly review three topics in machine learning. First, I review sparsity-inducing regularizations with L1 norm and their previous applications. Second, I review modeling with latent variables to explain basis of canonical correlation analysis and partial least squares method, and overview their previous applications. Finally, I review the basis of kernel methods, such as reproducing kernel Hilbert space and representer theorem, and overview the previous studies.

In chapter 3, I focus on diagnosis of depression based on functional magnetic resonance imaging (fMRI). In order to achieve both accurate diagnosis and identifying relevant anatomical regions for depression, I introduce region-wise sparseness using multiple kernel learning.

In chapter 4, I predict a number of clinical scores from resting-state functional connectivity in order to investigate the relation between conventional psychiatry and brain activity. Using partial least squares regression (PLSR) and its kernel variants, I demonstrate that they provide significantly better prediction of clinical scores than ordinary linear regression and subsequent classification using predicted clinical scores distinguishes depression patients from healthy controls with 80% accuracy. Moreover, I evaluate loading vectors for latent variables to identify brain regions relevant to depression.

In chapter 5, I investigate a novel method to introduce sparseness and interpretability in terms of nonlinear correlation analysis. Using synthetic datasets and nutrigenomic datasets, I show that my introduced method can extract multiple, nonlinear associations among high-dimensional data and multiplicative interactions among variables.

In chapter 6, I give some conclusion remarks.

2 Machine Learning Methods

In this chapter, I review three topics in machine learning algorithms.

2.1 Sparsity-Inducing Regularizations

Regularization was originally introduced to solve an ill-posed problem and avoid overfitting in the field of machine learning. Among many variants of regularization, L1 regularization was invented to learn the model and perform feature selection simultaneously [1].

2.1.1 Problems

L1 norm regularization

By adding L1 norm $\|\mathbf{w}\|_1$, an optimization problem with respect to $\mathbf{w} \in \mathbb{R}^d$ is written as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{w}) + \lambda\|\mathbf{w}\|_1, \tag{1}$$

where $L(\mathbf{w})$ is a convex and differentiable loss function and λ is a regularization parameter that controls the amount of regularization: the larger the value λ is, the larger the amount of regularization is. This optimization problem is reformulated as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{w}) \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq C, \tag{2}$$

where C is an upper bound of L1 norm. There is a one to one correspondence between λ and C. If the regularization parameter is large enough, or the upper bound is small enough, some of the coefficient $\mathbf{w}$ go to exact zero.

It is worth mentioning the sparsity-inducing mechanism of L1 regularization. Since the L1 norm $\|\mathbf{w}\|_1 = \sum |w_l|$ is convex but non-differentiable at a point

$w_l = 0$, subgradient is introduced as below.

Definition 2.1 *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and non-smooth function. A subgradient of f at a point $\mathbf{w}_0$ denoted by $\partial f(\mathbf{w}_0)$ is defined as follows:*

$$\partial f(\mathbf{w}_0) = \{\mathbf{g} \in \mathbb{R}^d : \forall \mathbf{w}, f(\mathbf{w}) - f(\mathbf{w}_0) \geq \langle \mathbf{g}, \mathbf{w} - \mathbf{w}_0 \rangle\}. \tag{3}$$

If zero vector is included in a subgradient at a point $\mathbf{w}_0$, $f(\mathbf{w}_0)$ is a minimum since $f(\mathbf{w}) \geq f(\mathbf{w}_0)$ for any $\mathbf{w}$.

The subgradient of Eq. (1) with respect to l-th coefficient is obtained as follows:

$$\frac{\partial L}{\partial w_l} + \lambda \partial |w_l|. \tag{4}$$

If the gradient of loss function evaluated as a point $w_l = 0$ is included in an interval $[-\lambda, \lambda]$ as follows:

$$\frac{\partial L}{\partial w_l} \, |_{w_l=0} \in [-\lambda, \lambda], \tag{5}$$

optimal minimum is achieved when $w_l = 0$.

Fu, et al. (1998) introduced one sort of coordinate descent gradient algorithms called shooting algorithm for L1 regularization with least squares regression [2]. Their method optimizes the following:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad \frac{1}{2} \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1, \tag{6}$$

where $\mathbf{y} \in \mathbb{R}^N$ are response variables and $X \in \mathbb{R}^{Np}$ is a design matrix where x_{ij} represents i-th sample's j-th feature. A subgradient of Eq. (6) is obtained as follows:

$$\frac{\partial L}{\partial w_l} + \lambda \partial |w_l| = \begin{cases} a_l w_l - \rho_l - \lambda, & w_l < 0, \\ [-\rho_l - \lambda, -\rho_l + \lambda] & w_l = 0, \\ a_l w_l - \rho_l + \lambda, & w_l > 0, \end{cases} \tag{7}$$

where $a_l = \sum_{i=1}^{N} x_{il}^2$ and $\rho_l = \sum_{i=1}^{N} x_{il}(y_i - \sum_{j \neq l} w_l x_{ij})$. ρ_l is proportional to the correlation of l-th feature and residuals obtained without it, indicating the relevance of l-th feature. Due to the optimality condition, the solution is as follows:

$$\mathbf{w}^* = \begin{cases} \frac{\rho_l + \lambda}{a_l}, & \rho_l \leq -\lambda \\ 0, & -\lambda < \rho_l < \lambda \\ \frac{\rho_l - \lambda}{a_l}, & \rho_l \geq \lambda. \end{cases} \tag{8}$$

This indicates that if relevance of each feature is less than the given λ, the corresponding weight becomes exact zero.

Although least angle regression and shrinkage (LARS) [3] is limited to the case where the loss function is a least squared error, it is well studied as a method to efficiently obtain an optimal solution for all possible λ.

Group L1 norm regularization

Suppose that d predictors are divided into G groups without any overlaps. Group lasso was proposed to introduce sparseness at the group level [4], formulated as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \quad L(\mathbf{w}) + \lambda \sum_{g=1}^{G} \|\mathbf{w}_g\|_2, \tag{9}$$

where $\| \cdot \|_2$ represents Euclidean norm. This optimization problem is reformulated as follows:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \quad L(\mathbf{w}) \quad \text{s.t.} \quad \sum_{g=1}^{G} \|\mathbf{w}_g\|_2 \leq C_g, \tag{10}$$

where C_g is an upper bound of a regularization term. This term forces all predictors within a group drop out of the model.

2.1.2 Proximal Gradient Methods

In Eq. (1) and (9), while $L(\mathbf{w})$ is differentiable convex function, such as least squares loss and logistic loss, a regularization term is non-differentiable function, suggesting conventional smooth optimization algorithms are not applicable. Proximal gradient method is a useful iterative algorithm applicable to both L1 and group L1 regularization similarly.

Definition 2.2 *A proximity operator $prox_g$ is defined as follows:*

$$prox_g(\mathbf{y}) = \arg\min_{\mathbf{w}\in\mathbb{R}^d}(\frac{1}{2}\|\mathbf{y}-\mathbf{w}\|_2^2 + g(\mathbf{w})), \tag{11}$$

where g is a convex function.

A proximity operator with L1 norm is defined as follows:

$$prox_\lambda^{l1}(\mathbf{y}) = \arg\min_{\mathbf{w}\in\mathbb{R}^d}(\frac{1}{2}\|\mathbf{y}-\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1). \tag{12}$$

The solution is as follows:

$$[prox_\lambda^{l1}(\mathbf{y})]_i = \begin{cases} y_i - \lambda, & y_i \geq \lambda \\ 0, & -\lambda \leq y_i \leq \lambda \\ y_i + \lambda, & y_i \leq -\lambda, \end{cases} \tag{13}$$

for $i = 1, ..., d$. A proximity operator with L1 norm is called soft threshold function.

A proximity operator with group L1 norm is defined as follows:

$$prox_{\lambda}^{gr}(\mathbf{y}) = \arg\min_{\mathbf{w}\in\mathbb{R}^d}(\frac{1}{2}\|\mathbf{y}-\mathbf{w}\|_2^2 + \sum_{g=1}^{G}\|\mathbf{w}_g\|_2). \tag{14}$$

The solution is obtained in each group as follows:

$$prox_{\lambda}^{gr}(\mathbf{y_g}) = \begin{cases} (\|\mathbf{y}_g\|_2 - \lambda)\frac{\mathbf{y}_g}{\|\mathbf{y}_g\|_2}, & \|\mathbf{y}_g\|_2 > \lambda \\ \mathbf{0}, & \text{others.} \end{cases} \tag{15}$$

In the case of Eq. (1), given t-th iteration $\mathbf{w}_t$, next iteration is obtained as follows:

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \nabla L(\mathbf{w}_t)^T(\mathbf{w}-\mathbf{w}_t) + \lambda\|\mathbf{w}\|_1 + \frac{1}{2\eta_t}\|\mathbf{w}-\mathbf{w}_t\|_2^2, \tag{16}$$

where the first term is a linear approximation of $L(\mathbf{w})$ at $\mathbf{w}_t$ and the third term is called proximity term that penalizes the distance from the current value. Due to this term, the solution is obtained using proximity operator as follows:

$$\mathbf{w}_{t+1} = prox_{\lambda\eta_t}^{l1}(\mathbf{w}_t - \eta_t\nabla L(\mathbf{w}_t)). \tag{17}$$

When the loss function is H smooth, proximal gradient method is known to be converged with $O(1/k)$ with the number of steps k.

2.1.3 L1 Regularization in Biomedical Data

Multivariate regression and classification with L1 regularization are useful to identify a small subset of millions of genotypes linked to the given phenotype such as disease outcome, in a genome-wide association (GWA) study [5, 6]. Wu, et al. (2009) applied logistic regression with L1 regularization to SNPs data in case-controls study and successfully identified a subset of SNPs relevant to coeliac disease. This method was extended to multi-task regression that

performed a joint GWA study from multiple populations, rather than to analyze each population separately to identify a subset of relevant genotypes shared across different populations [7].

2.2 Modeling with Latent Variables

A wide class of machine learning methods for modeling with latent variables, including canonical correlation analysis (CCA) and partial least squares regression (PLSR), are obtaining a great amount of attention in the field of biomedical engineering. They were introduced to model latent variables shared across two datasets.

2.2.1 Canonical Correlation Analysis

Canonical correlation analysis (CCA) is an method that extracts common characteristics involved in two datasets by maximizing the correlation coefficient between linear projections of two datasets [8].

Let $D = \{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$ be N pairs of samples, where $\mathbf{x}_i$ and $\mathbf{z}_i$ are the i-th samples drawn from p- and q-dimensional Euclidian space, respectively. Let $\mathbf{w} \in \mathbb{R}^p$ and $\mathbf{v} \in \mathbb{R}^q$ be projection vectors for $\mathbf{x}$ and $\mathbf{z}$, respectively. The objective of linear CCA is to find projection vectors that maximize Pearson's correlation between $\{\mathbf{w}^T\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{v}^T\mathbf{z}_i\}_{i=1}^N$ and formulated as the following optimization problem:

$$\max_{\mathbf{w}\in\mathbb{R}^p, \mathbf{v}\in\mathbb{R}^q} \quad \mathbf{w}^T C_{xz}\mathbf{v} \tag{18a}$$

$$\text{subject to} \quad \mathbf{w}^T C_{xx}\mathbf{w} = 1 \tag{18b}$$

$$\mathbf{v}^T C_{zz}\mathbf{v} = 1, \tag{18c}$$

where C_{xz} denotes the empirical covariance matrix between $\mathbf{x}$ and $\mathbf{z}$, and C_{xx} and C_{zz} denote the empirical covariance matrices of $\mathbf{x}$ and $\mathbf{z}$, respectively. Note that both $\mathbf{x}$ and $\mathbf{z}$ are set to have zero mean. The optimal solution $(\mathbf{w}^*, \mathbf{v}^*)$ of Eq. (18) is obtained by solving the method of Lagrange multiplier as follows:

$$\max_{\mathbf{w}\in\mathbb{R}^p, \mathbf{v}\in\mathbb{R}^q} L(\mathbf{w}, \mathbf{v}, \lambda_w, \lambda_v) = \mathbf{w}^T C_{xz}\mathbf{v} + \lambda_w(1 - \mathbf{w}^T C_{xx}\mathbf{w}) + \lambda_v(1 - \mathbf{v}^T C_{zz}\mathbf{v}), \tag{19}$$

where λ_w, λ_v are Lagrange multipliers. Optimization of Eq.(19) with respect to $\mathbf{w}, \mathbf{v}, \lambda_w$, and λ_v is formulated as generalized eigenvalue problem:

$$\begin{bmatrix} O & C_{xz} \\ C_{xz}^T & O \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix} = \lambda \begin{bmatrix} C_{xx} & O \\ O & C_{zz} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \mathbf{v} \end{bmatrix}, \tag{20}$$

where $\lambda = 2\lambda_w = 2\lambda_v$ is the canonical coefficient.

In the generalized eigenvalue problems, the k-th eigenvectors $\mathbf{w}_k^*$ and $\mathbf{v}_k^*$, corresponding to the k-th largest eigenvalue are the k-th projection vectors, and $\mathbf{w}_k^{*T}\mathbf{x}$ and $\mathbf{v}_k^{*T}\mathbf{z}$ are said to be the k-th canonical variables for $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$, respectively.

The different formulation of CCA is obtained from the view point of a rank-K matrix decomposion [9]. Suppose that identity matrices can be substituted for covariance matrices C_{xx} and C_{zz}, the following formulation is obtained as follows:

$$\max_{\mathbf{w}\in\mathbb{R}^p,\mathbf{v}\in\mathbb{R}^q} \quad \mathbf{w}^T C_{xz}\mathbf{v} \tag{21a}$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{w} = 1 \tag{21b}$$

$$\mathbf{v}^T\mathbf{v} = 1. \tag{21c}$$

Let $\mathbf{w}_k$ and $\mathbf{v}_k$ be the k-th column of matrix W and V, and d_k be the k-th diagonal element of a diagonal matrix D, respectively. Eq. (21) can be considered as the singular value decomposition of covariance matrix C_{xz} due to the following equation:

$$\|C_{xz} - WDV^T\|_F^2 = \|C_{xz}\|_F^2 - 2\sum_{k=1}^{K} \mathbf{w}_k^T C_{xz}\mathbf{v}_k + \sum_{k=1}^{K} d_k^2, \tag{22}$$

where $\mathbf{w}_k$ and $\mathbf{v}_k$ represent the k-th singular vectors.

Formulating CCA as the singular value decomposition of a covariance matrix enables us to introduce sparseness of weight vectors efficiently [9]. When the number of features p and q are larger than the number of sample n, unique solutions can be obtained using sparseness.

2.2.2 Canonical Correlation Analysis in Biomedical Data

CCA was used for modeling relations between gene expression data and single nucleotide polymorphism (SNPs) copy number measurements [9, 10], gene expression data and concentration of chemical compound [11, 12], several brain imaging data, such as fMRI and EEG in schizophrenia patients [13, 14]. Sparse CCA was developed to perform feature selection by introducing L1 regularization on projection vectors [9, 15]. Moreover, group sparseness was introduced to CCA [16] and used to analyze the relation between fMRI data and SNPs data

in schizophrenia patients [17].

2.2.3 Partial Least Squares Methods

Partial least squares methods are used to analyze the relations between two datasets by transforming datasets into a low dimensional space [18]. There are two types of partial least squares methods: (1) partial least squares correlation (PLSC) is maximizing a covariance of projections of two datasets, indicating that it is similar with CCA, (2) partial least squares regression (PLSR) is a regression technique that predicts one dataset from another.

PLS models a linear relation between two blocks of variables $\{\mathbf{x}_i\}_{i=1}^n \in \mathbb{R}^p$ and $\{\mathbf{y}_i\}_{i=1}^n \in \mathbb{R}^q$. In the following parts, $X = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ represents the $(n \times p)$ predictor matrix and $Y = (\mathbf{y}_1, ..., \mathbf{y}_n)^T$ represents the $(n \times q)$ response matrix. This procedure obtains L latent components as $\{\mathbf{t}_i\}_{i=1}^L$ and $\{\mathbf{u}_i\}_{i=1}^L$ and assumes following decompositions:

$$X = TP^T + F_x \tag{23}$$

$$Y = UQ^T + F_y, \tag{24}$$

where both $T = (\mathbf{t}_1, ..., \mathbf{t}_L)$ and $U = (\mathbf{u}_1, ..., \mathbf{u}_L)$ are the $(n \times L)$ matrices of L latent components corresponding to X and Y, respectively. The $(p \times L)$ matrix P and the $(q \times L)$ matrix Q are loadings and the $(n \times p)$ matrix F_x and the $(n \times q)$ matrix F_y are matrices of residuals.

Our objective is to obtain weight vectors $\mathbf{w} \in \mathbb{R}^p$ and $\mathbf{c} \in \mathbb{R}^q$ such that

$$\max_{\mathbf{t},\mathbf{u}} \mathtt{cov}(\mathbf{t}, \mathbf{u}) = \max_{\mathbf{w},\mathbf{c}} \mathtt{cov}(X\mathbf{w}, Y\mathbf{c}). \tag{25}$$

The classical form of PLS is called nonlinear iterative partial least squares (NI-

PALS) algorithm which optimizes weight vectors $\mathbf{w}$ and $\mathbf{c}$ and latent vectors $\mathbf{t}$ and $\mathbf{u}$ iteratively with random initialization of latent vector $\mathbf{u}$ (see Algorithm 1).

Algorithm 1 NIPALS

Input: X, Y
randomly initialize $\mathbf{u}$
repeat
  $\mathbf{w} = X^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$
  $\|\mathbf{w}\| \to 1$
  $\mathbf{t} = X\mathbf{w}$
  $\mathbf{c} = Y^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$
  $\|\mathbf{c}\| \to 1$
  $\mathbf{u} = Y\mathbf{c}$
until Convergence
series Output: $\mathbf{w}, \mathbf{t}, \mathbf{c}, \mathbf{u}$

Note that $\|\mathbf{w}\|_2$ and $\|\mathbf{c}\|_2$ is one.

deflation

After extracting the first latent component, the observation matrices X and Y are deflated by subtracting their rank-1 approximation to obtain series of latent component. Different variants of deflation define different forms of analysis. One of the variants is a symmetric deflation for PLSC called PLS mode A that was originally introduced by Wold, et al. [19]. By repeating the above Algorithm 1 and deflation procedure, L times, the weight matrices $W = (\mathbf{w}_1, ..., \mathbf{w}_L)$ and $C = (\mathbf{c}_1, ..., \mathbf{c}_L)$ are obtained. Note that loading of X, $\mathbf{p}$ is obtained by solving following problem:

$$\min_{\mathbf{p}} \|X - \mathbf{t}\mathbf{p}^T\|_F^2. \tag{26}$$

Loading $\mathbf{p}$ is given as $X^T\mathbf{t}/(\mathbf{t}^T\mathbf{t})$. In the same way, the oading of Y is obtained, $\mathbf{q} = Y^T\mathbf{u}/(\mathbf{u}^T\mathbf{u})$.

Considering the objective Eq. (25) can be seen as singular value decomposition of X^TY, a series of latent components and weight vectors is obtained

Algorithm 2 PLS mode A

series Input: X and Y
repeat
  obtain $\mathbf{w}, \mathbf{t}, \mathbf{c}, \mathbf{u}$ using Algorithm 1
  $X \leftarrow X - \mathbf{t}\mathbf{p}^T$
  $Y \leftarrow Y - \mathbf{u}\mathbf{q}^T$
until obtain L components
series Output: W, T, C, U

at once without iterative deflation. This procedure is equivalent to CCA as singular value decomposition.

In contrast to the above symmetric deflation, asymmetric scheme that Y is deflated based on $\mathbf{t}$ is required for PLSR. The underlying assumption is that the latent component of Y, $\mathbf{u}_i$, is well predicted from $\mathbf{u}_i$ such that

$$U = TD + F, \tag{27}$$

where D is a regression coefficients matrix and F is a residual matrix.

Algorithm 3 PLSR

series Input: X and Y
repeat
  obtain $\mathbf{w}, \mathbf{t}, \mathbf{c}, \mathbf{u}$ using Algorithm 1
  $X \leftarrow X - \mathbf{t}\mathbf{p}^T$
  $Y \leftarrow Y - \mathbf{t}\mathbf{c}^T$
until obtain L components
series Output: W, T, C, U

Finally, the relation in the original data space can be expressed by

$$Y = XB + E, \tag{28}$$

where B is the $(p \times q)$ matrix of regression coefficients and E is the $(n \times q)$ matrix of residuals.

Plugging the relationship $B = W(P^TW)^{-1}C^T$ [20, 21] into Eq. (28), the

different expression of Y is obtained as follows:

$$\begin{aligned} \hat{Y} &= XB && (29) \\ &= XW(P^TW)^{-1}C^T && (30) \\ &= XX^TU(T^TXX^TU)^{-1}T^TY. && (31) \end{aligned}$$

The final transformation is derived by following equalities [22]:

$$\begin{aligned} W &= X^TU, && (32) \\ P &= X^TT(T^TT)^{-1}, && (33) \\ C &= Y^TT(T^TT)^{-1}. && (34) \end{aligned}$$

Note that $\mathbf{t}_i^T\mathbf{t}_j = \delta_{ij}$ (the Kronecher delta) takes the values 1 for $i = j$ and 0 for $i \neq j$ by the consequence of algorithm.

In general, B is obtained from a centered training data set. The response $\mathbf{y}_{new}$ for a new subject $\mathbf{x}_{new}$, referred to as test data set, is then estimated as follows:

$$\mathbf{y}_{new} = \bar{\mathbf{y}} + B^T(\mathbf{x}_{new} - \bar{\mathbf{x}}), \quad (35)$$

where $\bar{\mathbf{y}}$ and $\bar{\mathbf{x}}$ represent mean predictor and response in training data set, respectively.

2.2.4 Partial Least Squares Methods in Biomedical Data

After McIntosh, et al. firstly introduced PLSC to identify common information between functional neuroimaging data and the block design of task experiment or behavior pattern of subjects [23], it was actively studied in the field of neu-

roimaging. For example, it was used to model the relation between gray matter density in human brain measured by MRI and psychological assessments of cognitive ability [24], PET imaging with in-vivo amyloid marker and mental score for Alzheimer's patients [25]. Moreover, it was applicable for dimension reduction prior to a classification of gene expression data to types of tumor and confirmed to outperform existing dimension reduction method, such as principal component analysis (PCA) [26]. In the context of multivariate regression, PLSR was used to predict individuals' nicotine effect from fMRI data [27] and survival time from gene expression data [28].

2.3 Kernel Methods

In the field of machine learning, kernel methods using positive semi-definite matrices are used for complex pattern analysis. The idea of kernel methods is to transform data from an original space into a high-dimensional feature space where the standard linear algorithm is applied. While various nonlinear information is incorporated in a high-dimensional feature space, kernel method does not require to specify the transformation explicitly. In stead, by specifying a kernel, a similarity between data points in the original space, and a corresponding reproducing kernel Hilbert space (RKHS), nonlinear pattern recognition can be performed efficiently.

2.3.1 A Positive Semi-Definite Kernel

First, suppose a function called a kernel that represents similarity between data points. Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, y \in \mathcal{X}$,

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}. \tag{36}$$

A map transforms original data into a high-dimensional feature space $\mathcal{H}$. In the Eq. (36), $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ represents an inner product in the feature space $\mathcal{H}$.

The kernel trick is to define k in an original data space $\mathcal{X}$ without specifying a map and computing an inner product in the feature space $\mathcal{H}$. The following kernels are examples of kernels frequently used in kernel algorithms:

1. linear kernel

   $k(x, y) = x^T y,$

2. polynomial kernel

   $k(x, y) = (x^T y + a)^b \quad (a \geq 0, b \in \mathbb{N}),$

3. Gaussian kernel

   $k(x, y) = \exp(-\gamma \|x - y\|_2^2) \quad (\gamma > 0),$

4. exponential kernel

   $k(x, y) = \exp(-\beta x^T y) \quad (\beta > 0).$

Due to the kernel trick, kernels can be defined on sequence data, text data, graphs, as well as vector data.

A kernel, k is a positive semi-definite kernel if $k(x, y) = k(y, x)$, and for every $x_1, ..., x_n \in \mathcal{X}$ and $c_1, ..., c_n \in \mathbb{R}$,

$$\sum_{i,j}^{n} c_i c_j k(x_i, x_j) \geq 0. \tag{37}$$

Let $k_1, k_2, ...$ be positive semi-definite kernel kernels defined on $\mathcal{X}$, the following kernels are also positive semi-definite kernels:

1. the sum

   $k = \sum \lambda_i k_i, \quad (\lambda_1, \lambda_2, ... \geq 0),$

2. the product

   $k(x, y) = k_1(x, y) k_2(x, y),$

3. the limit

   $k = \lim_{i\to\infty} k_i$,

4. the normalize

   $\tilde{k}(x, y) = \frac{k(x,y)}{\sqrt{k(x,x)k(y,y)}}$.

In the field on machine learning, it is important to point out that every positive semi-definite kernel defines a unique reproducing kernel Hilbert space. A function $f : \mathcal{X} \to \mathbb{R}$ in a reproducing kernel Hilbert space is characterized by the reproducing property. Evaluation of f at a point x denoted by $f(x)$ is defined as an inner product in feature space as follows:

$$f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}}, \tag{38}$$

with $k(\cdot, x) \in \mathcal{H}$.

2.3.2 Representer Theorem

Given input data $\{x\}_{i=1}^n$, output data $\{y\}_{i=1}^n$, a kernel k over $\mathcal{X}$, and a corresponding reproducing kernel Hilbert space $\mathcal{H}$, the following optimization problem can be considered:

$$\min_{f\in\mathcal{H}} \quad \sum_{i=1}^{N} L(y_i, f(x_i)) + \Phi(\|f\|), \tag{39}$$

where L is a loss function and $\Phi : \mathbb{R} \to \mathbb{R}$ is a monotonically increasing function. Following theorem called representer theorem shows that kernel algorithm reduces to an N-dimensional optimization problem [29].

Theorem 2.1 *A optimum of Eq. (39) is written as follows:*

$$f^* = \sum_{i=1}^{N} \alpha_i k(\cdot, x_i). \tag{40}$$

Proof 2.1 *Let $\mathcal{H}_0$ be a finite dimensional subspace spanned by $\{k(\cdot, x_1),...,k(\cdot, x_N)\}$ and $\mathcal{H}_0^{\perp}$ be orthogonal compliment, orthogonal decomposition is given as follows:*

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_0^{\perp}. \tag{41}$$

A function $f \in \mathcal{H}$ is decomposed into $f_0 \in \mathcal{H}_0$ and $f_{\perp} \in \mathcal{H}_0^{\perp}$. Due to the reproducing property,

$$\begin{aligned} \langle f, k(\cdot, x_i)\rangle &= \langle f_0, k(\cdot, x_i)\rangle + \langle f_0^{\perp}, k(\cdot, x_i)\rangle & \text{(42a)} \\ &= \langle f_0, k(\cdot, x_i)\rangle & \text{(42b)} \\ &= f_0(x_i). & \text{(42c)} \end{aligned}$$

In addition, considering $\|f\| = \|f_0\| + \|f_0^{\perp}\|$ and monotonically increasing function Φ, the relation $\Phi(\|f_0\|) \leq \Phi(\|f\|)$ is obtained. Therefore, an optimum f^* is $f_0 \in \mathcal{H}_0$. □

2.3.3 Kernel Methods in Biomedical Data

Among variants of kernel methods, support vector machine (SVM) is well studied in the field of biomedical enginerring [30]. As microarray experiments which yield a huge amount of gene expression measurements prevail, SVM was used for functionally classification of gene expression [31, 32]. In diagnosis of multiple malignancies, SVM outperformed conventional histopathological method in accuracy using gene expression measurements from cancer tissue [33, 34]. It was also used in prediction of protein property using protein structure [35, 36], and the objective diagnosis of depressed patients [37]. It was also used to identify genes associated with cancer by applying SVM recursively to a smaller and smaller sets of features [38].

One of challenging topics in current biomedical enginerring is to character-

ize data represented by graphs (e.g. protein-protein interactions network and genetic interaction network) and strings (e.g. sequence of DNA or amino acid). In previous studies, diffusion kernels and string kernels were used to capture the similarity of arbitrary pair of samples [39, 40]. For an example, string kernels was used to yield sequence similarity based on shared occurrence of fixed length patterns in two sequences. They were applied to protein classification into functional and structural families [40], prediction of transcription start site location [41], feature extraction of microarray data [42], and prediction of protein interactions [43].

3 Sparse Multiple Kernel Learning for Diagnosis of Depression

3.1 Introduction

Major depressive disorder (MDD) is though to be caused by a malfunction of the neural circuit, but little is understood about its detailed etiology. Recent advance in brain imaging, such as functional magnetic resonance imaging (fMRI) allows us to monitor human brain activity and opens up a possibility to discover neural correlates of MDD in a data-driven manner.

Several recent studies have shown that machine learning algorithms, such as Gaussian process classifier and support vector machine are useful to classify patients with depression and healthy controls accurately. Especially, support vector machine (SVM) achieved significantly accurate diagnosis of depression and providee the neurobiological biomarker in depression [44, 45]. While SVM uses a single kernel, multiple kernel learning (MKL) uses a weighted summation of several sub-kernels and the weights are optimized [46, 47]. This technique was applied to fMRI data to combine different sources of data and improve the classification performance [48, 49].

In this study, I propose to apply a sparse multiple kernel learning algorithm called Spicy MKL [50] with sub-kernels defined in a region-wise manner. Here, each sub-kernel (called region-wise kernel) has it's own weight, enabling us to interpret the obtained model (i.e. I can interpret which anatomical regions are relevant to classification). In addition, using a sparse regularization term in MKL, the irrelevant anatomical regions are automatically removed.

Table 1: Patients and controls. Mean ±standard deviation

	Patients	Controls
age	33.5 ± 12	38.81 ± 9.76
gender (male)	15	16
mean reaction time (s)	1.28 ± 0.32	1.50 ± 0.33

3.2 Methods

This study was approved by the Human Subjects Research Review Committee at Okinawa Institute of Science of Technology as well as the Research Ethics Committee of Hiroshima University (permission nr. 172). Written and informed consent was obtained from all subjects participating in the study.

3.2.1 Data Acquisition and Preprocessing

31 drug naive patients diagnosed with depression were recruited by the Psychiatry Department of Hiroshima University and collaborating medical institutes. As a control group, 31 persons with no history of mental disorders were recruited (Table 1). All subjects performed a two-blocked cognitive experiment called phonological and semantic verbal fluency task. In the control condition, subjects were asked to repeat a word presented on display in their mind and press the button. In the task condition, they were asked to find some word, utter it in their mind, and press a button. The word found in each trial had to begin with a given syllable in phonological task and match a given category in semantic task.

The obtained images were processed with SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8), using slice timing correction, motion correction, normalization with standard brain template, and smoothing. The contrast between task and control conditions evaluated by z-scores was proceed to input of my method.

I denote the set of resulting data by $T = \{(x_i, y_i)\}_{i=1}^{N}$ $(N = 62)$, where

$x_i \in \mathbb{R}^D (D = 13972)$ is the fMRI data of the i-th subject and $y_i \in \{\pm 1\}$ is the target class indicating a patient ($y_i = +1$) or a healthy control ($y_i = -1$).

3.2.2 Classification Algorithm

To construct the binary classifier from T, I employed a sparse multiple kernel learning called "Spicy MKL" [50] with hinge loss. Given M sub-kernel functions $k_m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and corresponding reproducing kernel Hilbert spaces $\mathcal{H}_m$, sub-kernel matrices were defined as $[K_m]_{ij} = k_m(x_i, x_j)$ $(m = 1, 2, ..., M)$. I considered the learning problem with weighted sum of sub-kernels defined as $K = \sum_{m=1}^{M} d_m K_m$, where $\{d_m\}_{m=1}^{M}$ are non-negative values. Let $\| \cdot \|_{\mathcal{H}_m}$ and f_m be norms and functions in $\mathcal{H}_m$, respectively. The optimization problem is written as follows:

$$\underset{f_1 \in \mathcal{H}_1, ..., f_M \in \mathcal{H}_M, d_1 \geq 0, ..., d_M \geq 0}{\text{minimize}} E[f_1, ..., f_M] + \lambda \left(\sum_{m=1}^{M} \frac{\|f_m\|_{\mathcal{H}_m}^2}{d_m} + d_m \right). \tag{43}$$

In this equation, the first and second term are empirical hinge loss function and regularization term, respectively. The third term is L1 norm of weight vector introduced to penalize weight $\{d_m\}_{m=1}^{M}$. By optimizing weight with fixed $f_1, f_2, ..., f_M$, the optimal weights $d_m^* = \|f_m\|_{\mathcal{H}_m}$ are obtained. The equation (43) reduces to

$$\underset{f_1 \in \mathcal{H}_1, ..., f_M \in \mathcal{H}_M}{\text{minimize}} E[f_1, ..., f_M] + \lambda \sum_{m=1}^{M} \|f_m\|_{\mathcal{H}_m}. \tag{44}$$

Since the second term is block 1-norm regularization, corresponding to group lasso [50], norm of irrelevant function (i.e. irrelevant kernel weight) should be zero.

Due to the representer theorem [29], the solution is expressed in the form of

$f_{m,i} = \sum_{i=1}^{N} K_m(x_i, x_j)\alpha_j$, in which $\{\alpha_i\}_{i=1}^{N}$ can be obtained by an algorithm named dual augmented Lagrangian (DAL) [51, 50]. [1].

3.2.3 Region-wise Kernels

In the application of Spicy MKL to fMRI data, I defined M sub-kernels, $\{K_m\}_{m=1}^{M}$, based on the Automated Anatomical Labeling (AAL) [52] that divides whole-brain into $M = 116$ anatomical regions (see Fig. 1). Specifically, let $V(m)$ be a set of voxel indices belonging to the m-th anatomical region, and $\phi_m(x) \equiv (x_d)_{d \in V(m)}$ be a sub-vector of x which is constructed by extracting voxels included in the region. Then, I defined K_m as follows:

$$K_m(x_i, x_j) = \frac{\phi_m(x_i)^T \phi_m(x_j)}{\|\phi_m(x_i)\| \|\phi_m(x_j)\|}, \quad m = 1, ..., M,$$

where $\|\cdot\|$ denotes the Euclidean norm. Since the m-th sub-kernel has only information about the m-th anatomical region, $\{m | d_m \neq 0\}$ can be interpreted as a set of regions relevant to major depression.

3.2.4 Nested Leave One Out Cross Validation

Cross validation is employed to assure generalizability of a model or evaluate optimal parameters. Since I had a parameter λ that controls complexity, I hence made use of nested leave one out cross validation (LOOCV), which consisted of outer- and inner-LOOCV. The outer-LOOCV repeated iterations that splited whole set of samples into a single outer-validation sample used to evaluate the generalizability and an outer-training set for model estimation. The inner-loop of LOOCV was performed on the outer-training set to optimize the parameters. The parameter which achieved the highest classification accuracy based on inner-validation sample was adopted as an optimal parameter and used to

[1] The matlab toolbox is available at http://www.is.titech.ac.jp/~s-taiji/software/SpicyMKL

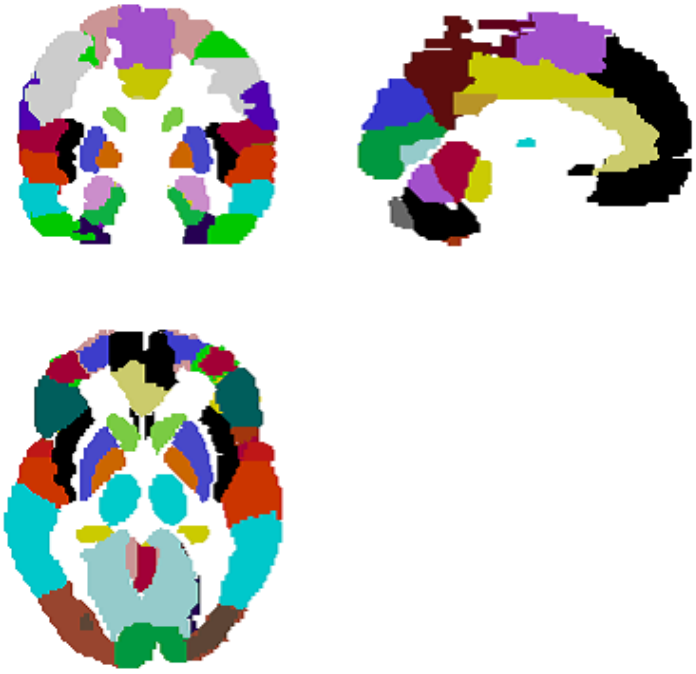

Figure 1: **Automatic anatomical labeling.** Each anatomical region defined by automatic anatomical labeling is color-coded. There are 116 anatomical regions in whole-brain.

evaluate the model using the outer-LOOCV. These steps were repeated until each sample has served as validation sample.

3.3 Results

To evaluate performance of my method based on Spicy MKL, I compared it with logistic regression with LASSO (denoted by Logit) [1] and linear SVM [30], in terms of the classification performance, such as accuracy, sensitivity (i.e. correct detection of patients), and specificity (i.e. correct detection of healthy controls). My method achieved 76.8 and 78.4% accuracy in phonological and semantic data, respectively (Fig. 2 and Table 2 , 3). It was significantly better than that of Logit (69.6 and 65.1% accuracy, respectively) and comparable with that of SVM (80.3 and 81.0% accuracy, respectively). All statistical comparisons were adjusted for multiplicity using the Bonferroni-Holm method with significance level, $\alpha = 0.05$.

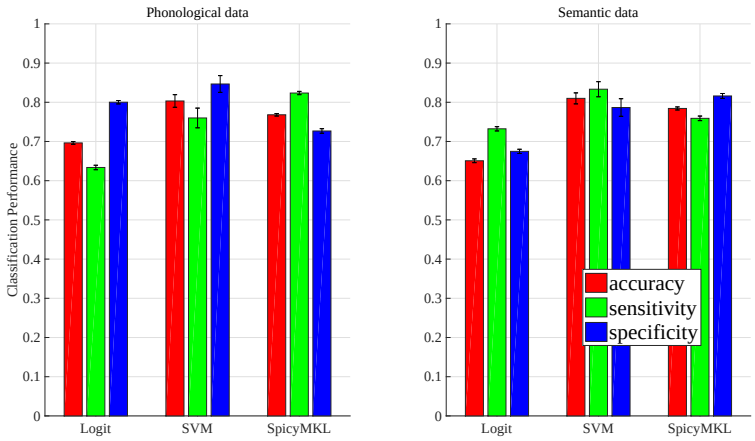

Figure 2: **Classification performance.** While my method significantly outperformed logistic regression with L1 regularization in accuracy, it was comparable with SVM.

Since Logit's classification was found to be unreliable, I compared only Spicy

Table 2: Classification performance in phonological data.

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logit	69.6±3.5	63.4±5.7	80.2±4.3
SVM	80.3±1.6	76.0±2.5	84.7±2.1
SpicyMKL	76.8±0.3	82.4±0.4	72.7±0.6

Table 3: Classification performance in semantic data.

	Accuracy (%)	Sensitivity (%)	Specificity (%)
Logit	65.1±0.5	73.2±0.5	67.5±0.5
SVM	81.0±1.4	83.3±1.9	78.7±2.3
SpicyMKL	78.4±0.4	75.9±0.6	81.6±0.6

MKL and SVM in terms of feature selection ability. Fig. 3 and 4 shows voxels contributing to the classification. For SVM, almost every voxel was selected as the contributing one, making it difficult to interpret relevant brain regions. In contrast, for Spicy MKL, the contributing voxels were located in *left postcentral gyrus* and *left middle frontal cortex* in phonological data ,and *left precentral gyrus*, *left precuneus*, and *left cerebellum crus1* in semantic data.

3.4 Discussion

Sparse multiple kernel learning algorithm using region-wise kernels defined in this study is consistent with group lasso if I consider voxels belonging to same anatomical region as a group [4, 53]. Our study identified relevant anatomical regions, consistent with the previous study [54].

The part of the *precentral gyrus* is the primary motor cortex in human brain that controls motor system. This region was pointed out to be related to a semantic task in fMRI study of aphasia [55]. The *left precuneus* is known as a hub of default mode network that shows synchronized deactivation during cognitive tasks and suggested to be associated with depression in resting state

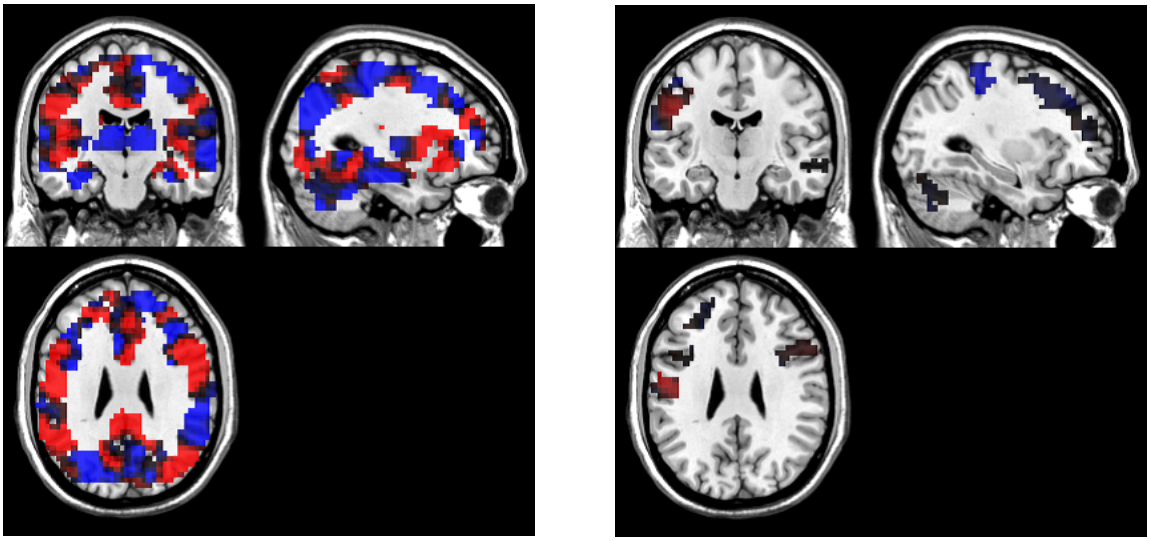

Figure 3: **Relevant brain regions in phonological verbal fluency task.** (Left) Relevant brain regions by SVM, distributing in whole-brain. (Right) Relevant brain regions by Spicy MKL with region-wise kernels, distributing in *left postcentral gyrus* and *left middle frontal cortex*. The red and blue voxels represent positive and negative weight, respectively.

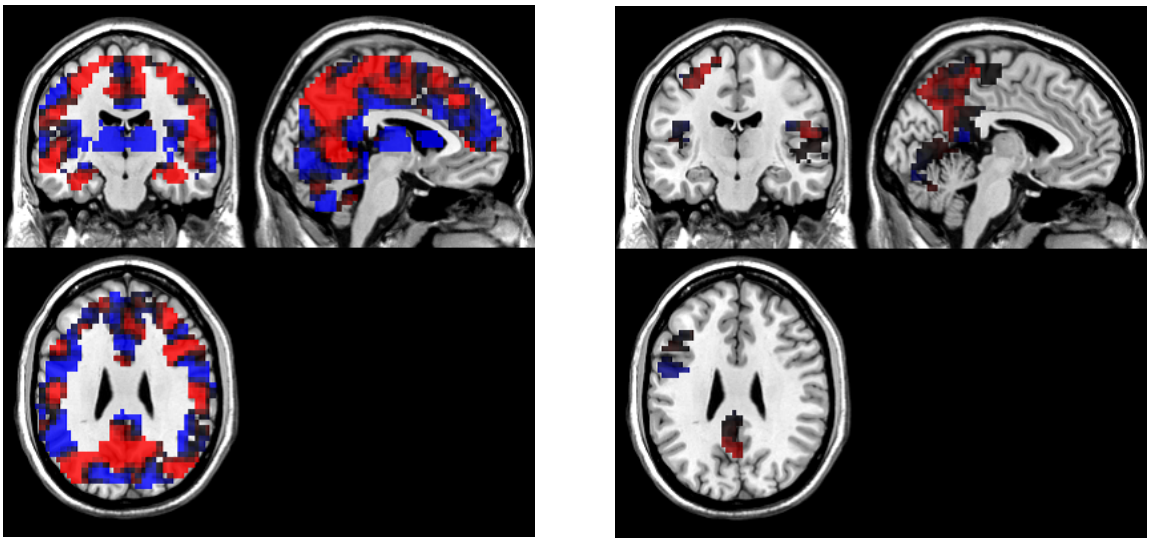

Figure 4: **Relevant brain regions in semantic verbal fluency task.** (Left) Relevant brain regions by SVM, distributing in whole-brain. (Right) Relevant brain regions by Spicy MKL with region-wise kernels, distributing in *left precentral gyrus*, *left precuneus*, and *left cerebellum crus1*. The red and blue voxels represent positive and negative weight, respectively.

fMRI studies [56, 57, 58]. In addition, increased activity in the *left precuneus* was found to be related to some risk-gene in depression [59] in fMRI study with a semantic verbal fluency task. While the *left cerebellum crus1* is usually considered to be responsible for motion controls in general my results indicated that it might be associated with depression in a semantic task. Some fMRI studies demonstrated that this area was responsible for various types of information processing [60, 61] and resting state functional connectivity study implied that *cerebellum* might be critical region for distinguish between depression patients and healthy controls [62, 63].

Our result suggests that these regions characterize the depressive patients in a semantic verbal fluency task.

3.5 Conclusion

I successfully classified depression patients and healthy controls with 83.9% accuracy, using region-wise kernels for sparse multiple kernel learning algorithm called Spicy MKL. The weight of each kernel demonstrated that *left precentral gyrus*, *left precuneus*, and *left cerebellum crus1* were relevant anatomical regions for depression.

4 Prediction of Clinical Depression Scores and Detection of Changes in Whole-brain using Resting-state Functional MRI Data with Partial Least Squares Regression

4.1 Introduction

Advances in analyzing large datasets with machine learning algorithms promote their application in medical diagnosis. In particular, their use in objective diagnosis of psychiatric disorders using brain imaging and other biological data is now being actively studied [54]. A major challenge in applying statistical machine learning algorithms to brain imaging or genetic data is the high dimensionality of the input variables, such as the number of voxels and the number of possible genetic polymorphisms. Even though algorithms such as support vector machine (SVM) and L1-regularized classifiers (LASSO) manage the issue of high-dimensionality, the problem of co-linearity in brain imaging data remains. Neural activities in nearby voxels or in the same functional network are highly correlated, which makes the results of commonly used regression or classification tools unreliable. In this thesis, I propose the use of partial least squares (PLSR) regression [64, 23, 65, 66, 24, 27] with multiple clinical measures to address this problem. Here, I project resting-state functional magnetic resonance imaging (rs-fMRI) data and clinical scores from clinically depressed patients and healthy control subjects into a low-dimensional space and use them to predict depression-related clinical measures and thereafter, to classify subjects.

Use of rs-fMRI is gaining attention in diagnosis of psychiatric disorders because it makes few cognitive demands in measurements and because it can be applied to multiple disorders [67]. In depressed patients, functional connec-

tivities (FCs) between brain areas estimated using rs-fMRI show distributed changes throughout the entire brain [68, 69, 70, 71]. Zeng et al. (2012) [63] demonstrated that $\sim 94\%$ of 53 subjects could be correctly classified as patients or healthy controls using FCs and linear SVM, and they reported that the majority of discriminating FCs were distributed within or across the default mode network, the affective network, visual cortical areas, and the cerebellum. While the aforementioned study sought to discern differences between patients and healthy controls in a binary manner, Zhang et al. (2011) [72] tried to predict clinical measures of the Beck Depression Inventory II (BDI-II) [73] by regressing fMRI signals acquired during a face-watching task. They showed that true and predicted BDI-II were significantly correlated ($r = 0.55$) and using the standard threshold of 14 for the predicted BDI-II, 89% of the automated diagnoses agreed with those of psychiatrists.

Clinical depression is characterized by multiple, related symptoms [74]. There are various clinical measures for assessing symptoms, such as the Snaith-Hamilton Pleasure Scale (SHAPS) [75] for anhedonia and Positive and Negative Affect Schedule (PANAS) [76] for altered mood. In addition, the age of subjects is important for diagnosis since aging increases the risk of depression in general [77].

Here, I consider a two-step method which predicts multiple measures of clinical depression from rs-fMRI in the first step, and then uses results of the first step for diagnosis. For the first step, I explore a regression model to predict BDI-II, SHAPS, PANAS(n), and age from functional connectivity data. Although this could be done using ordinary least squares regression, in order to tackle the issue of high-dimensionality and co-linearity of the input, I explore the use of partial least squares regression (PLSR) [64, 23, 65, 66, 24, 27], which maps input and output variables to low-dimensional spaces so that the covariance of data in the latent spaces is maximized. I compare the classification performance of

the two-step method using PLSR with other classification methods. Thereafter, I consider the use of subject age by testing (i) a model with age as a response variable (output-age model), (ii) a model with age as a predictor (input-age model), and (iii) a model that does not consider age (no-age model).

In this chapter, I further develop the basic idea presented in [78] to overcome limitations of linear methods and perform objective diagnosis. In section 4.2, I illustrate the details of rs-fMRI and clinical measures for subjects. Section 4.3 provides the mathematical basis of PLSR and its kernel variants. In addition, it is extended to classification models for the purpose of objective diagnosis. In section 4.4, I illustrate the efficacy of my application in predicting clinical measures, discriminating between patients and healthy controls, and interpreting derived coefficients. Finally, I offer my conclusions and discuss future work in section 4.5.

4.2 Data Set

This study was approved by the Human Subjects Research Review Committee at the Okinawa Institute of Science of Technology, as well as the Research Ethics Committee of Hiroshima University (permission nr. 172). Written consent was obtained from all subjects participating in the study.

4.2.1 Subjects

58 patients (age $26 - 73$, average 42.8 ± 11.9, 33 female) with major depression disorder were recruited by the Psychiatry Department of Hiroshima University and collaborating medical institutions, based on the Mini-international neuropsychiatric interview (M.I.N.I [79]), which enables doctors to identify psychiatric disorders, according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV [80]). As a healthy control group, 65 sub-

jects (ages $20-66$, average 34.8 ± 13.0, 28 female) with no history of mental or neurological disease, were recruited via advertisements in local newspapers.

Clinical Measures

The following interview- and questionnaire-based measures are used for determination of disease presence and quantification of the severity of two primary symptoms I wish to predict, namely, anhedonia (loss of motivation, loss of pleasure, etc.) and negative mood (low mood, guilty feelings, suicidal thoughts, etc.).

Beck Depression Inventory II (BDI-II)

This measure evaluates the presence and severity of depression based on a self-report questionnaire [73]. Subjects are asked to answer 21 questions about feelings of punishment or guilt, suicidal thoughts, etc. Each answer is scored with a value between 0 and 3, with 3 being the most serious. High scores indicate severe symptoms. The standardized score of $\geq$14 indicates that a subject is suffering from depression.

Snaith-Hamilton Pleasure Scale (SHAPS)

This measure was developed to evaluate the level of anhedonia [75]. Subjects are asked to answer 14 questions about hedonic capacity, with scores between 1 and 4. High scores indicate more severe anhedonia.

Positive and Negative Affect Schedule (PANAS)

This widely used measure evaluates positive and negative moods of subjects [76, 81]. In this study, I considered only scores related to negative mood items. This measure is generally known as PANAS(n). Subjects are asked to respond to 10 questions about their moods, with answers between 0 and 5. The sum of

Table 4: Mean (± standard deviation) of clinical measures

	Controls	Patients
Number of subjects	65	58
Age	34.8 (±13.0)	42.8 (±11.9)
BDI-II	6.92 (±5.9)	30.9 (±9.0)
SHAPS	23.3 (±6.2)	37.8 (±5.5)
PANAS(n)	8.5 (±6.4)	25.1 (±7.9)

all scores indicates the strength of their negative moods. Due to an evaluation issue, one subject's response could not be assessed, so that score was replaced with the mean of the remaining subjects.

Table 4 summarizes scores exhibited for each measure by each group in my study. Although most patients showed both anhedonia and negative mood, some exhibited only one trait. Correspondingly, the scores of the BDI-II, SHAPS, and PANAS(n) are highly, but not completely correlated. As decreased mental function results from aging, the age of the subjects is expected to correlate with BDI-II, SHAPS, and PANAS(n) as well.

I verified these correlations by calculating the correlation coefficients (Table 5). Strong correlations between clinical measures are reflected in coefficients above 0.7. Weaker correlations between age and individual clinical measures were around 0.3. In my regression model, BDI-II, SHAPS, PANAS(n), and age of each subject are considered as responses in order to correct for their natural correlation, resulting from functional connectivity. I will show that the introduction of subject age as an output rather than as an input is beneficial with respect to classification accuracy.

Table 5: The Pearson's correlation coefficients between the clinical measures and the subjects' age

	Age	BDI-II	SHAPS	PANAS(n)
BDI-II	0.2451	-	0.7883	0.8005
SHAPS	0.3221	0.7883	-	0.7497
PANAS(n)	0.2480	0.8005	0.7497	-

4.2.2 Functional Connectivity of resting-state fMRI

Functional MRI measurements were acquired on a 3T GE Signa HDx scanner with a 2D EP/GR (TR = 3s, TE=27ms, FA=90deg, matrix size 64x64x32, voxel size 4x4x4 mm, no gap, interleaved). Subjects were instructed to lie with their eyes open, to think of nothing in particular, and to remain awake. They were also instructed to refrain from taking caffeine, nicotine, and alcohol in the day of experiment.

For each subject, acquired images were processed with SPM8 (Wellcome Trust Centre for Neuroimaging, UCL, London) following standard procedures. I first performed slice timing correction, motion correction, co-registration to anatomical MRI, normalization with standard brain and smoothing (Gaussian of full-width at half-maximum 8mm). I confirmed that there were no significant differences in six motion parameters between two diagnostic groups in order to reject a possible effect of spurious functional connectivity due to head motion [82, 83]. Voxels were assigned to 116 brain regions, according to the automatic anatomical labeling atlas (AAL) [52]. Mean activation time series in each brain region were obtained by averaging MRI signal time series over all voxels assigned to each region. Note that I also tested finer brain atlases, such as the Brainvisa Sulci Atlas (BSA) [84] and extended BSA (BAL) [67], however, they did not provide better prediction performance than by AAL. Finally, functional connectivity between each pair of regions was computed as the cross

correlation of the corresponding time-series.

4.3 Methods

Partial least squares regression (PLSR) is a method for modeling a relationship between two sets of multivariate data via a latent space, and of performing least squares regression in that space. PLSR can handle high-dimensional co-linear datasets because of its underlying assumption that the two datasets are generated by a small number of latent components. In this process, latent components are formed by maximizing the covariance between the two datasets (see section 2.2). Conveniently, it can be easily extended to nonlinear regression models using a kernel trick so I briefly review the kernel variants of PLSR in this section [21, 85].

4.3.1 Kernel Partial Least Squares Regression (KPLSR)

Let $\phi : \mathbb{R}^p \to \mathcal{H}$ be a nonlinear transformation of the predictor, $\mathbf{x} \in \mathbb{R}^p$, into a feature vector, $\phi(\mathbf{x}) \in \mathcal{H}$, where $\mathcal{H}$ is a high-dimensional feature space. Define a Gram matrix K as inner products of points in feature space, i.e., $K = \Phi\Phi^T$, where $\Phi = (\phi(\mathbf{x}_1), ..., \phi(\mathbf{x}_n))^T$ represents the predictor matrix in feature space. In general, the number of columns of Φ is so large that with the explicit form of Φ, the same procedure can not be performed as in the linear case. However, due to the kernel trick, the explicit form of Φ becomes unnecessary.

The deflation procedure is performed as follows:

$$K \leftarrow (I_n - \mathbf{t}\mathbf{t}^T)K(I_n - \mathbf{t}\mathbf{t}^T) \tag{45}$$

$$Y \leftarrow Y - \mathbf{t}\mathbf{t}^T Y, \tag{46}$$

where I_n represents an n-dimensional identity matrix.

I obtain the prediction on the training data:

$$\hat{Y} = \Phi B \tag{47}$$

$$= \Phi\Phi^T U(T^T\Phi\Phi^T U)^{-1}T^T Y \tag{48}$$

$$= KU(T^T KU)^{-1}T^T Y. \tag{49}$$

To exclude the bias term, I assumed that the responses and the predictors were set to have zero mean in the feature space by applying the following procedure to test kernel K_t and training kernel K [86]:

$$K_t \leftarrow (K_t - \frac{1}{n}\mathbf{1}_{n_t}\mathbf{1}_n^T K)(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T) \tag{50}$$

$$K \leftarrow (I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)K(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T), \tag{51}$$

where $\mathbf{1}_n$ represents the n-length vector whose n elements are 1. Note that n and n_t represent the number of training and test samples, respectively.

In the following part of this thesis, I investigated three kernel functions: 1) a second order polynomial kernel $k(x, x^{'}) = (x^T x^{'} + 1)^2$, referred to as KPLS-Poly(2), 2) a third order polynomial kernel $k(x, x^{'}) = (x^T x^{'} + 1)^3$, referred to as KPLS-Poly(3), 3) a Gaussian kernel $k(x, x^{'}) = \exp(-\gamma||x - x^{'}||)^2)$, referred to as KPLS-Gauss, where γ is a hyper parameter and set to the inverse of the median of the Euclidian distance of data points.

4.3.2 Classification

In addition to predicting clinical measures, I investigated classification of subjects into depressed patients and healthy controls using the predicted value of clinical measures for objective diagnosis. I evaluated generalization of binary classifiers using linear discriminant analysis (LDA). Given the training data

$\mathbb{D}_{tr} = \{\mathbf{x}_{tr}, \mathbf{y}_{tr}, \mathbf{z}_{tr}\}$ and test data $\mathbb{D}_{te} = \{\mathbf{x}_{te}, \mathbf{y}_{te}, \mathbf{z}_{te}\}$, $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^q$, and $\mathbf{z} \in \{0, 1\}$ represent functional connectivity as predictors, clinical measures as responses, and binary labels (i.e. 0 is patients and 1 is healthy controls), respectively. In the prediction phase, my objective is to learn the function $f_B : \mathbb{R}^p \to \mathbb{R}^q$, which, given predictors, $\mathbf{x}_{tr}$, and responses, $\mathbf{y}_{tr}$, assigns predictors to the most probable values of $\mathbf{y}$. The prediction on the training dataset is $\hat{\mathbf{y}}_{tr} = f_B(\mathbf{x}_{tr})$. In the next classification phase, my objective is to learn the classifier $g_w : \mathbb{R}^q \to \{0, 1\}$, which, given predicted responses, $\hat{\mathbf{y}}_{tr}$, and binary labels, $\mathbf{z}_{tr}$, assigns predicted responses to the most probable labels. Assigned labels on the test dataset are obtained as $\hat{\mathbf{z}}_{te} = g_w(\hat{\mathbf{y}}_{te}) = g_w(f_B(\hat{\mathbf{x}}_{te}))$. It is important to stress that the binary classifier is not trained on actual clinical measures, $\mathbf{y}_{tr}$, but on predicted values of $\hat{\mathbf{y}}_{tr}$.

In a previous study [63], the authors only identified the binary classifier $g_w^{'} : \mathbb{R}^p \to 0, 1$, which, given functional connectivity, $\mathbf{x}_{tr}$, and binary labels, $\mathbf{z}_{tr}$, assigns functional connectivity directly to binary labels. By exploiting the predicted result of clinical measures, it may be possible to improve classification performance. I compared two scenarios, i.e. i) classification of patients and healthy controls using LDA from predicted clinical measures with KPLSR (with KPLS-Gauss, KPLS-Poly(3), and KPLS-Poly(2)), PLSR, and ordinary least squares regression (OLS), ii) classification of patients and healthy controls by means of LDA and SVM from functional connectivity directly. Note that I performed feature selection before scenario 2) by calculating connection-wise t-tests to determine the connections with different group means, represented by t-scores. I selected the M functional connections with the highest absolute t-scores. M was optimized by cross validations.

Pre-screening

Even though PLSR can cope with high-dimensional, co-linear datasets, I pre-screened variables depending on their relevance to responses in the following way.

Based on Pearson correlation coefficients, ρ_{rl}, between the r-th functional connection and the l-th clinical measures, I defined the empirical relevance of the r-th functional connection as follows:

$$R_r = \sum_{l=1}^{4} \rho_{rl}^2, \qquad r = 1, \ldots, p, \tag{52}$$

where p is the total number of functional connections.

These functional connections were ranked according to their empirical relevance, $\{R_r\}_{r=1}^{p}$, and only M relevant functional connections were used in following procedure. The optimal number for M was determined through nested leave-one-out cross-validation.

Nested leave-one-out cross validation

Conventionally, cross validation is employed to assure generalization ability of a model or to evaluate optimal parameters. Since I had to account for both generalization ability and parameter optimization, I made use of nested leave-one-out cross validation (LOOCV), which consisted of outer and inner LOOCV. The outer LOOCV repeated iterations that divided the whole set of samples into a single outer validation sample used to evaluate the generalization ability, and an outer training set for model estimation. The inner loop of LOOCV was performed on the outer training set to optimize two parameters, M and L, the number of selected predictor variables and the number of components, respectively. The pair of parameters that achieved the lowest root mean squared error based on the inner validation sample were adopted as optimal parameters

and used to evaluate the model using the outer LOOCV. These steps were repeated until each sample had served as the validation sample.

Age

Age was significantly correlated with three clinical measures (Table 5). In general, age matching performed on different diagnostic groups reduces sample size, causing poor performance. To avoid this problem, I investigated three models, i.e. (i) a model with age as a response (denoted by output-age), (ii) a model with age as a predictor (denoted by input-age), and (iii) a model without age (denoted by no-age). By incorporating age into my model, I could cope with age differences among subjects and can fairly evaluate prediction performance.

Interpretation

Interpretation of each latent component projected from input and output data gives novel insights into the relationship between functional connectivity and clinical measures. In the framework of PLSR, loading matrices, P and C, indicate contributions from predictor variables and response variables to each latent component (see Eq. 10 and 11). The (i, j)-element of the loading matrix, P, represents the contribution of the i-th functional connection to the j-th latent component. Similarly, the (i, j)-element of the loading matrix, C, represents the contribution of the i-th clinical measure to the j-th latent component. Note that due to subject variability, values of P_{ij} and C_{ij} vary depending on the training set used.

4.4 Results

4.4.1 Regression Performance

I compared the prediction performance of PLSR, its kernel variants, and other methods by means of the root mean squared error (RMSE) of the predicted clinical measures in nested leave-one-out cross validation (see Methods). Kernel PLSR with a second-order polynomial kernel (KPLS-Poly(2)) achieved the lowest RMSE (9.56 for BDI-II, 6.11 for SHAPS, and 7.29 for PANAS(n)) (Fig. 5). This performance was significantly better than that of ordinary least squares regression (OLS) (11.6 for BDI-II, 7.33 for SHAPS, and 8.91 for PANAS(n)) and comparable to that of other variants of PLSR applied in my study, suggesting that projection of data into a low-dimensional space was beneficial to regression performance. All statistical comparisons were adjusted for multiplicity using the Bonferroni-Holm method with significance level, $\alpha = 0.05$.

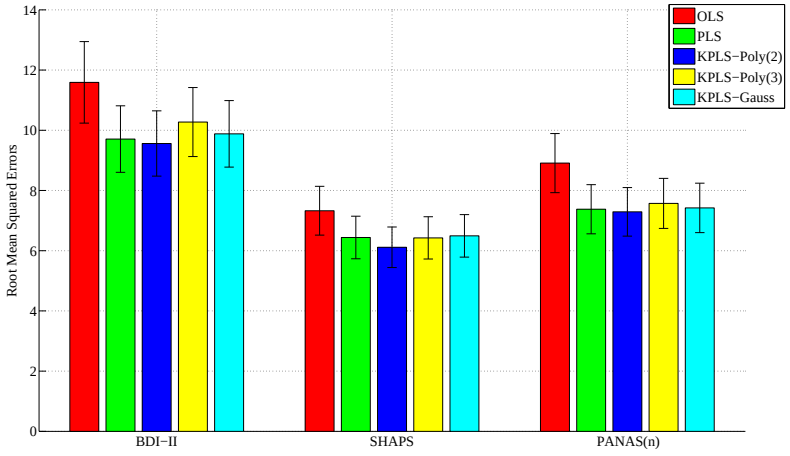

Figure 5: **Comparison of predicted performance by means of the root mean squared errors.** Linear and kernel variants of PLSR achieved significantly better performance than did OLS in all clinical scores. Subject age was used as the output along with clinical scores (output-age model).

Next, to evaluate the best way of incorporating age into my regression models, I compared RMSE of the output-age, input-age, and no-age models. In my study, incorporating age into my regression model as a response (output-age) achieved significantly better performance than that of the input-age and no-age models (Fig. 6, Table S1). All statistical comparisons were adjusted for multiplicity using the Bonferroni-Holm method with significance level, $\alpha = 0.05$.

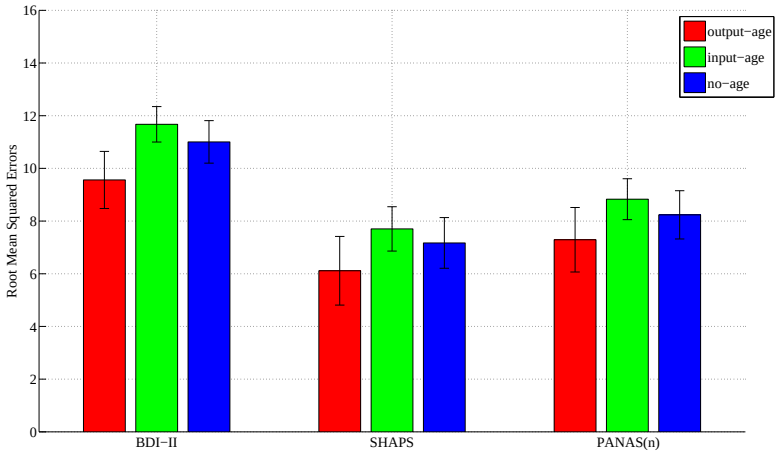

Figure 6: **Root mean squared errors in KPLS-Poly(2).** KPLS-Poly(2) achieved significantly better performance in output-age model than in other models.

The correlation coefficient of actual and predicted values for BDI-II, SHAPS, and PANAS(n) in the case of KPLS-Poly(2) were $r = 0.541, 0.591, 0.563$, respectively. Fig. 7 exemplifies the relationship between predicted and actual values of BDI-II for KPLS-Poly(2). This result was comparable to that of Zhang et al. (2011) [72]; however, the number of subjects in my study was larger than in theirs, reconfirming validity of the results.

The optimal number of retained features M^* identified by pre-screening and using the latent component L^* identified with nested LOOCV were 40

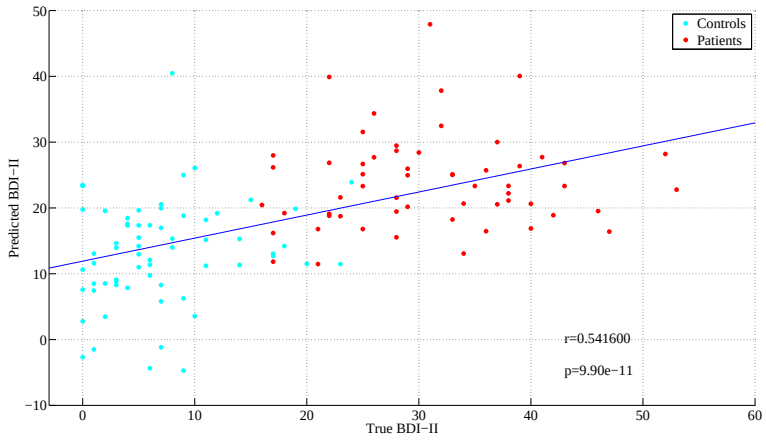

Figure 7: **Actual and predicted values of BDI-II.** BDI-II were well predicted by KPLS-Poly(2) with RMSE=9.56 and $r = 0.541$ ($p < 10^{-10}$). Red and blue points represent patients and healthy controls, respectively.

and 3, respectively, suggesting that reduction of feature size was relevant for improvement of PLSR accuracy.

4.4.2 Classification Performance

Projecting the original data onto a low-dimensional space was expected to improve classification accuracy. To verify the benefit of projection, several classification methods were performed and evaluated using accuracy, sensitivity, and specificity (Fig. 8 and Table S2). In my study, KPLS-Poly(2) followed by LDA achieved the best accuracy 80.5% (sensitivity 81.0% and specificity 80.0%), which is significantly better than the 57.7% accuracy of direct LDA (sensitivity 53.4%, and specificity 61.5%) and 69.1% accuracy of direct SVM (sensitivity 69.0%, and specificity 69.2%). This result indicates that it was beneficial to exploit the prediction model for clinical measures in order to build a classification model. In addition, KPLS-Poly(2) followed by LDA also achieved significantly better accuracy than the 62.6% accuracy of OLS followed by LDA (sensitiv-

ity 62.1% and specificity 63.1%), indicating that considering a latent space in a regression model was beneficial to final classification. Accuracy did not differ significantly between PLSR and kernel variants. All statistical tests were based on approximation with the normal and adjusted for multiplicity using the Bonferroni-Holm method with significance level $\alpha = 0.05$.

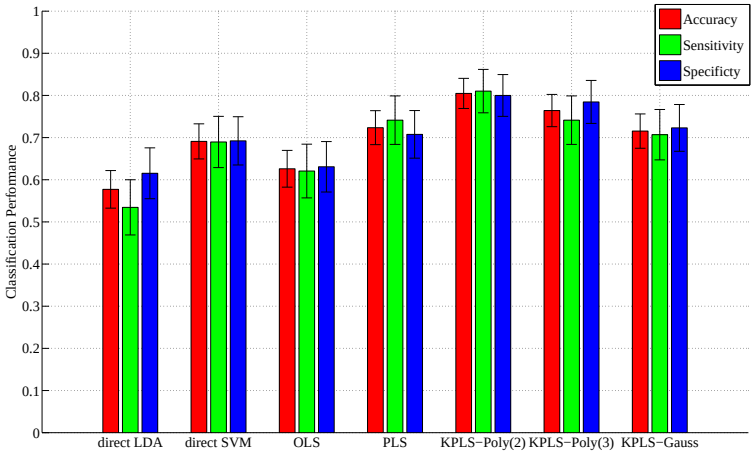

Figure 8: **Classification accuracy, sensitivity, and specificity.** KPLS-Poly(2) followed by LDA achieves the best performance (accuracy=80.5%, sensitivity=81.0%, and specificity=80.0%).

4.4.3 Interpretation

In my study, three clinical scores showed almost equally positive influences on the first component, and age also had a positive influence as well. However, age showed a strong negative influence on the second component, in contrast to the clinical scores (Table. 6).

Latent space representation of subjects showed that the first component explained most depression severity in comparison with the second component (Fig. 9). This is consistent with the results of loading matrix C. Note that since

Table 6: Loading matrix C. Mean $\pm$ standard deviation.

	BDI-II	SHAPS	PANAS(n)	age
1st	$7.23\pm8.17\times10^{-2}$	$7.50\pm7.56\times10^{-2}$	$7.40\pm8.54\times10^{-2}$	$4.55\pm1.35\times10^{-1}$
2nd	$1.21\pm3.69\times10^{-1}$	$1.24\pm3.77\times10^{-1}$	$2.53\pm5.88\times10^{-1}$	-5.26±1.43

the optimal number of latent components, in terms of minimizing regression error, was 3, the second and the third components are thought to contain some information about scores.

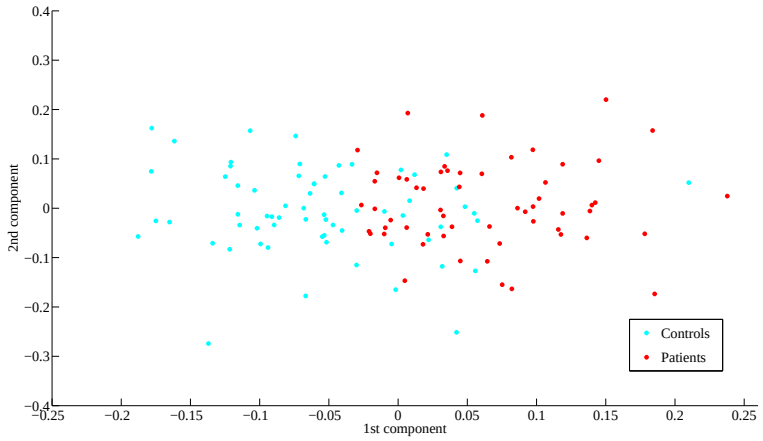

Figure 9: Scatter plot of the latent variables in the first two latent components generated from KPLS-Poly(2). Red and blue dots represent patients and healthy controls, respectively. The two groups are separated mainly by the first latent component.

In order to validate the effect of age, especially in the second component, all subjects were grouped into young (age $20-31$, 41 subjects), middle (age $31-43$, 41 subjects), and old (age $44-73$, 41 subjects) groups. Note I simply divided the subjects in three equal-sized groups for convenience, "young", "middle", and "old". They are relative, not absolute age classes. Latent variables of old subjects in the second component were significantly lower than those of young

and middle subjects ($p < 10^{-5}$ by Wilcoxon Rank-Sum Test), suggesting that old patients have distinctive patterns in the second latent space [66] (Fig. 10).

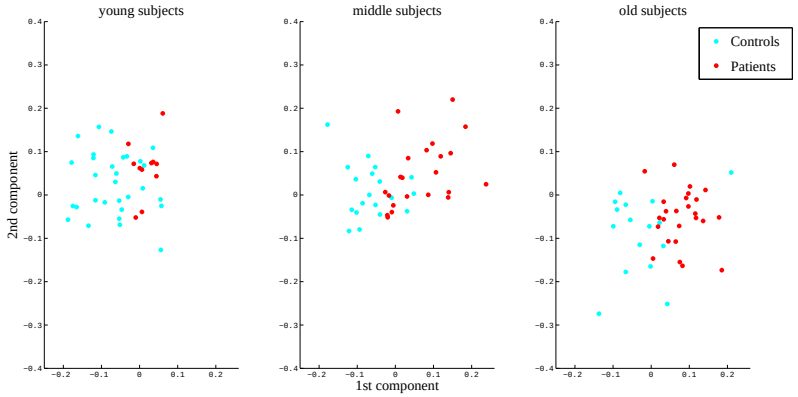

Figure 10: Scatter plot of subjects separated on the basis of the two latent components generated from KPLS-Poly(2) for young, middle, and old subjects. Old subjects have significantly lower values in the second component ($p < 10^{-5}$ by Wilcoxon Rank-Sum Test).

Evaluation of loading matrix, P, reveals functional connections relevant to each latent component. Especially, the first column of P, corresponding to the first component responsible for discrimination of each diagnostic group, is expected to reveal useful insights about the effect of functional connections on depression symptoms. Even though the performance of KPLS-Poly(2) in prediction and classification was comparable to or better than that of linear PLSR, patterns of significant loadings were consistent in my experiments. For reasons of interpretation, I therefore focus on the loading matrix of the linear terms in the following part.

BrainNet Viewer [87] (http://www.nitrc.org/projects/bnv/) was used to visualize the top 10 connections with positive and negative loadings for the first component (Fig. 11 and 12). In this figure, many regions involved in the default mode network (DMN), as well as *the left supplementary motor area*, *the*

right superior frontal gyrus, and *the insula*, were relevant. In addition, some functional connectivity between *the right cuneus* and regions involved in *the cerebellum* were negatively correlated with the first component.

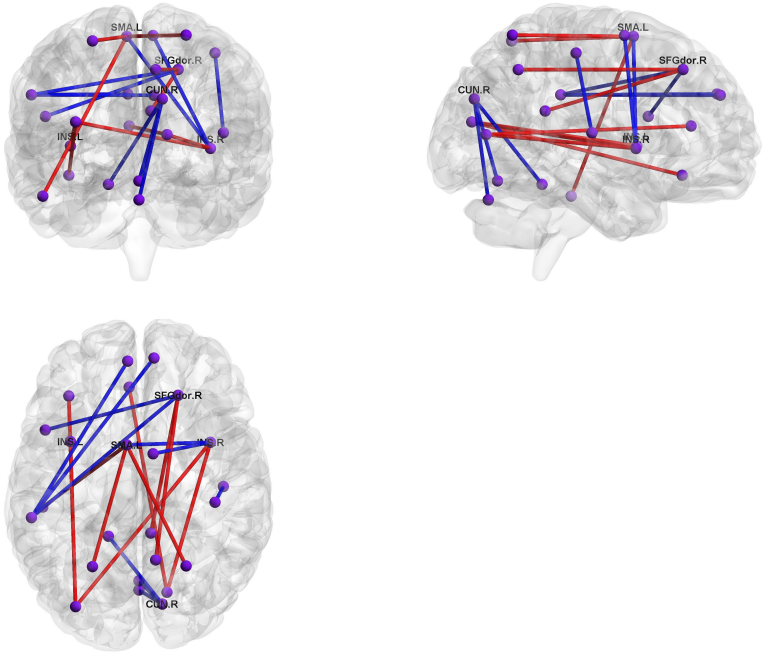

Figure 11: Contributing functional connectivity in first latent component. Red and blue lines represent positive and negative loadings, respectively. SFGdor.R: *right superior frontal gyrus*, INS: *insula*, SMA.L: *left supplementary motor area*, CUN.R:*right cuneus*.

4.5 Discussion

MacIntosh et al. (1996) first introduced partial least squares method into the field of neuroimaging in order to extract common information between brain activity and exogenous information, such as experimental or behavioral data [23, 65]. In particular, behavioral data are increasingly used to extract associated brain activity patterns for various types of psychological diseases, such

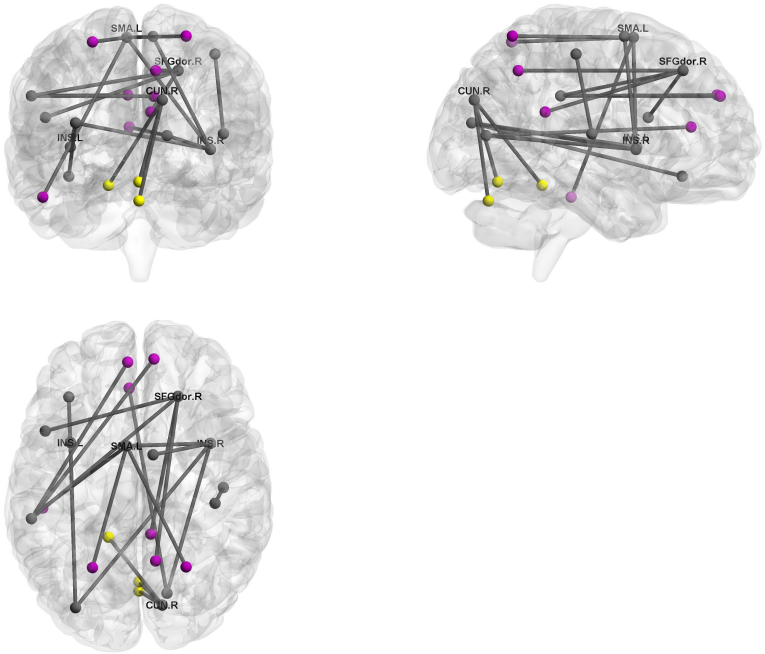

Figure 12: Contributing functional connectivity in first latent component. Purple and yellow nodes represent brain areas within the default mode network and *cerebellum*, respectively. SFGdor.R: *right superior frontal gyrus*, INS: *insula*, SMA.L: *left supplementary motor area*, CUN.R:*right cuneus*.

as Alzheimer's disease [25], obsessive-compulsive disorder [88], and schizophrenia [89]. In these studies, neuropsychological test scores are used as behavioral data, in addition to the labels that represent diagnostic groups and age. To the best of my knowledge, this is the first study to investigate associations between functional connectivity in whole brain and multiple clinical measures for depressed patients, using PLSR and its kernel variants.

Diagnosis based on resting-state functional connectivity is a challenging task due to the high-dimensionality and co-linearity of data. Recent studies have demonstrated that depressed patients can be distinguished from healthy controls by means of their functional connectivity by applying conventional methods,

such as support vector machine [63, 68]. Since binary labels are ultimately abstracted information about depression that ignores the severity of symptoms, it is worth considering more detailed information, such as BDI-II, SHAPS, and PANAS(n) to build more sophisticated models. Our study demonstrated that projecting functional connectivity data into a low-dimensional latent space, can predict clinical measures, and can also improve depression diagnostic accuracy (see from Fig. 5 to Fig. 8).

To separately identify neural circuits associated with anhedonia and negative mood is a challenging task. A psychopathological study suggests that these primary symptoms result from different neural circuits and from alternation of different neurotransmitters [74]. Our results show that SHAPS and PANAS(n) are highly correlated and contributed quite similarly to each latent component (see Table 6), suggesting that further investigation and different methodes may be required to support psychopathological studies from the point of data driven analysis.

4.5.1 Contributing brain regions

Identification of relevant brain regions in functional connectivity analysis yielded the following three observations: (1) connections between the default mode network and other regions, such as *the right superior frontal gyrus* and *the left supplementary motor area* are relevant (2) *the left and right insulas* in both hemispheres are relevant, (3) connections between *the cerebellum* and *the right cuneus* are relevant.

First, the default mode network (DMN) shows synchronized deactivation during cognitive tasks and is thought to be related to major depressive disorder [90, 91, 58, 57]. My study supports these results, indicating many contributing connections related to the DMN, such as *the right posterior cingulum*, *the right precuneus*, and *the superior parietal gyrus*. The DMN contributes posi-

tive connections with *the right superior frontal gyrus* and *the left supplementary motor area*. *The superior frontal gyrus*, as a critical region in cognitive tasks, was previously reported to be associated with depression [92]. While *the supplementary motor area* is known to be responsible for motor control, it was also reportedly related to some subtype of depression [93]. Our results support these results.

Second, my results suggest that *the insula* is associated with depression. This supports some meta-analysis of PET and fMRI studies that revealed that *the insula* plays an important role in regulation of emotion [94, 95]. Similarly, other resting-state fMRI studies indicated that *the insula* is directly associated with depression [96, 97].

Finally, my results show that connections between *the right cuneus*, located in the visual cortical area, and *the cerebellum*, negatively influence depression. While visual processing is believed not to be affected in depression, some previous studies suggested that it was associated with bipolar disorder [98]. In addition, there are other study that showed regional homogeneity (ReHo) as a measure of localized synchrony in resting-state fMRI was decreased [99]. In general, *the cerebellum* is considered to be responsible for motion control, however my results indicate that it may also be involved in regulation of mood and cognitive processing associated with symptoms of depression. Some fMRI studies demonstrated that this area was responsible for various types of information processing [60, 61], and some resting-state functional connectivity studies implied that *the cerebellum* may be critical for the distinction between depressed patients and healthy controls [62, 63]. My result is consistent with these previous studies.

4.6 Conclusion

In summary, I employed partial least squares regression and its kernel variants to predict clinical measures of subjects using resting-state functional connectivity. Diagnosis of depression based on predicted clinical scores performed better than classification algorithms attempting diagnoses directly from functional connectivity. Moreover, analysis of latent variables identified functional networks relevant to the diagnosis of depression. These results suggest that a low-dimensional representation derived using PLSR is beneficial for objective diagnosis. Further investigations are required to separate the two neural circuits associated with two primary symptoms, anhedonia and negative mood.

5 Sparse Kernel Canonical Correlation Analysis for discovery of nonlinear interactions in High-Dimensional Data

5.1 Introduction

Canonical correlation analysis (CCA) [8] is a statistical method for finding common information from two different sources of multivariate data. This method optimizes linear projection vectors so that two random multivariate datasets are maximally correlated. With advances in high-throughput biological measurements, such as DNA sequencing, RNA microarrays, and mass spectroscopy, CCA has been extensively used for discovery of interactions between the genome, gene transcription, protein synthesis, and metabolites [43, 15, 9, 12]. Because CCA solution is reduced to an eigenvalue problem, multiple components of interactions with sparse constraints are readily introduced [9, 100, 101].

Kernel CCA (KCCA) was introduced to capture nonlinear associations between two blocks of multivariate data [102, 103, 42, 104]. Given two blocks of multivariate data $\mathbf{x}$ and $\mathbf{z}$, KCCA finds nonlinear transformations $f(\mathbf{x})$ and $g(\mathbf{z})$ in a reproducing kernel Hilbert space (RKHS) so that the correlation between $f(\mathbf{x})$ and $g(\mathbf{z})$ is maximized. While this method performs properly in a low-dimensional dataset, it is impossible to remove irrelevant features and avoid overfitting in a high-dimensional dataset. In order to solve the problem and to improve interpretability of results, sparse additive functional CCA (SAFCCA) [105] constrains $f(\mathbf{x})$ and $g(\mathbf{z})$ as sparse additive models and optimizes them using the biconvex back-fitting algorithm [106]. However, it is not straightforward to obtain multiple orthogonal transformations for extracting multiple components of associations.

In this thesis, I propose two-stage kernel CCA (TSKCCA), which enables

us (1) to select sparse features in high-dimensional data and (2) to obtain multiple nonlinear associations. In the first stage, I represent target kernels with a weighted sum of pre-specified sub-kernels and optimize their weight coefficients based on HSIC with sparse regularization. In the second stage, I apply standard KCCA using target kernels obtained in the first stage to find multiple nonlinear correlations.

I briefly review CCA, KCCA, and two-stage MKL, and then present TSKCCA algorithm. I apply TSKCCA to three synthetic datasets and nutrigenomic experimental data to show that the method discovers multiple nonlinear associations within high-dimensional data, and provides interpretation that are robust to irrelevant features.

5.2 CCA, Kernel CCA, and Multiple Kernel Learning

In this section, I briefly review the bases of my proposed method, namely, kernel CCA (KCCA), and multiple kernel learning (MKL). The basis of linear CCA was reviewed in section 2.2.

5.2.1 Kernel CCA

In Kernel CCA (KCCA), I suppose that the original data can be mapped into a feature space via nonlinear functions. Then linear CCA is applied in the feature space. More specifically, nonlinear functions $\phi_x : \mathbb{R}^p \to \mathbb{H}_x$ and $\phi_z : \mathbb{R}^q \to \mathbb{H}_z$ transform the original data $\{(\mathbf{x}_n, \mathbf{z}_n)\}_{n=1}^N$ to feature vectors $\{(\phi_x(\mathbf{x}_n), \phi_z(\mathbf{z}_n))\}_{n=1}^N$ in reproducing kernel Hilbert spaces (RKHS) $\mathbb{H}_x$ and $\mathbb{H}_z$. Inner-product kernels for $\mathbb{H}_x$ and $\mathbb{H}_z$ are defined as $k_x(\mathbf{x}, \mathbf{x}') = \phi_x(\mathbf{x})^T \phi_x(\mathbf{x}')$, and $k_z(\mathbf{z}, \mathbf{z}') = \phi_z(\mathbf{z})^T \phi_z(\mathbf{z}')$.

Let us implement $f_w(\mathbf{x})$ and $g_v(\mathbf{z})$ by projections $f_w(\mathbf{x}) \equiv \mathbf{w}^T \phi_x(\mathbf{x})$ and $g_v(\mathbf{z}) \equiv \mathbf{v}^T \phi_z(\mathbf{z})$. By introducing appropriate regularization terms, Eq. (18)

can be reformulated as the following optimization problem ([102, 103]):

$$\max_{\alpha\in\mathbb{R}^N,\beta\in\mathbb{R}^N} \quad \alpha^T K_x K_z \beta \tag{53a}$$

$$\text{subject to} \quad \boldsymbol{\alpha}^T \left(K_x + \frac{N\kappa}{2} I \right)^2 \boldsymbol{\alpha} = 1 \tag{53b}$$

$$\boldsymbol{\beta}^T \left(K_z + \frac{N\kappa}{2} I \right)^2 \boldsymbol{\beta} = 1, \tag{53c}$$

where K_x and K_z are N-by-N kernel matrices defined as $[K_x]_{nn'} = k_x(\mathbf{x}_n, \mathbf{x}_{n'})$ and $[K_z]_{nn'} = k_z(\mathbf{z}_n, \mathbf{z}_{n'})$ [1]. I is the N-by-N identity matrix and κ ($\kappa > 0$) is the regularization parameter.

Once having obtained the solution of Eq. (53), denoted by $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$, canonical variables for $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{z} \in \mathbb{R}^q$ are given by

$$f^*(\mathbf{x}) = \sum_{n=1}^{N} k_x(\mathbf{x}, \mathbf{x}_n)\alpha_n^* \tag{54a}$$

$$g^*(\mathbf{z}) = \sum_{n=1}^{N} k_z(\mathbf{z}, \mathbf{z}_n)\beta_n^*, \tag{54b}$$

respectively. As indicated by Eq. (53), the nonlinear functions, ϕ_x and ϕ_z, are not explicitly used in the computation of KCCA. Instead, the kernels k_x and k_z implicitly specify the nonlinear functions, and the main goal is to solve the constrained quadratic optimization problem with $2N$-dimensional variables.

5.2.2 Multiple Kernel Learning

Kernel methods usually require users to design a particular kernel, which critically affects the performance of the algorithm. To make the design more flexible, the framework of multiple kernel learning (MKL) was proposed for classification and regression problems [46, 47]. In MKL, I manually design M_x sub-kernels

[1] In this thesis, $[\cdot]_{nn'}$ denotes the (n, n')-th elements of the matrix enclosed by the brackets.

$\{k_x^{(m)}\}_{m=1}^{M_x}$, where each sub-kernel $k_x^{(m)}$ uses only a distinct set of features in $\mathbf{x}$. Also, M_z sub-kernels $\{k_z^{(l)}\}_{l=1}^{M_z}$ for $\mathbf{z}$ is also designed in the same manner. Then, k_x and k_z are represented as the weighted sum of those sub-kernels:

$$k_x(\mathbf{x}, \mathbf{x}') = \sum_{m=1}^{M_x} \eta_m k_x^{(m)}(\mathbf{x}, \mathbf{x}') \tag{55a}$$

$$k_z(\mathbf{z}, \mathbf{z}') = \sum_{l=1}^{M_z} \mu_l k_z^{(l)}(\mathbf{z}, \mathbf{z}'), \tag{55b}$$

where weight coefficients of sub-kernels, $\{\eta_m\}_{m=1}^{M_x}$ and $\{\mu_l\}_{l=1}^{M_z}$ are tuned to optimize an objective function.

A specific example of this framework is the two-stage MKL method [107, 108]: In the first stage, the weight coefficients are optimized based on a similarity criterion, such as the kernel target alignment; then, a standard kernel algorithm, such as support vector machine, is applied in the second stage.

5.3 Methods

In this section, I propose a novel nonlinear CCA method, two-stage kernel CCA (TSKCCA), inspired by the concepts of sparse multiple kernel learning and kernel CCA. In the following, I present the general framework of TSKCCA, followed by my solutions for practical issues in the implementation.

5.3.1 First Stage: Multiple Kernel Learning with HSIC and Sparse Regularizer

In TSKCCA, sub-kernels are restricted to the same class as Eq. (55), allowing us to express the kernel matrices K_x and K_z as follows:

$$K_x = \sum_{m=1}^{M_x} \eta_m K_x^{(m)} \tag{56a}$$

$$K_z = \sum_{l=1}^{M_z} \mu_l K_z^{(l)}, \tag{56b}$$

where $[K_x^{(m)}]_{nn'} = k_x^{(m)}(\mathbf{x}_n, \mathbf{x}_{n'})$ and $[K_z^{(l)}]_{nn'} = k_z^{(l)}(\mathbf{z}_n, \mathbf{z}_{n'})$. The goal of the first stage is to optimize the weight vector $\boldsymbol{\eta} = (\eta_1, ..., \eta_{M_x})^T$ and $\boldsymbol{\mu} = (\mu_1, ..., \mu_{M_z})^T$ so that kernel matrices K_x and K_z statistically depend on each other as much as possible, while irrelevant sub-kernels are filtered out.

The statistical dependence between K_x and K_z is evaluated by the Hilbert-Schmidt Independent Criterion (HSIC) and approximated by its empirical estimator [109]:

$$\mathbb{D}(K_x, K_z) = \frac{\mathrm{Tr}(K_x H K_z H)}{(N-1)^2}, \tag{57}$$

where H is an N-by-N matrix such that $[H]_{nn'} = \delta_{nn'} - \frac{1}{N}$, and $\delta_{nn'}$ is Kronecker's delta. $\mathrm{Tr}(\cdot)$ denotes the trace. In my setting, optimization problem is reduced to a simple bilinear form with respect to η and μ:

$$\mathbb{D}(K_x, K_z) = \boldsymbol{\eta}^{\mathrm{T}} M \boldsymbol{\mu}, \tag{58}$$

where M is a M_x-by-M_z matrix such that

$$[M]_{ml} = \frac{\mathrm{Tr}(K_x^{(m)} H K_z^{(l)} H)}{(N-1)^2}. \tag{59}$$

In addition to maximizing the dependency measure $\mathbb{D}(K_x, K_z)$, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ should be sparse in order to filter out irrelevant sub-kernel matrices. To this end, I determine optimal weight vectors as the solution of the following constrained optimization problem:

Algorithm 4 Penalized Matrix Decomposition for Learning Kernels

Input: M (Eq. 59), regularization c_1 and c_2
for $i = 1$ **to** $rank(M)$ **do**
  initialize $\boldsymbol{\eta}^i$ to a first left singular vector of M
  repeat
    $\boldsymbol{\mu}^{(i)} \leftarrow \frac{S((\boldsymbol{\eta}^{(i)T}M)_+,\Delta)}{\|S((\boldsymbol{\eta}^{(i)T}M)_+,\Delta)\|_2}$
    $\boldsymbol{\eta}^{(i)} \leftarrow \frac{S((M\boldsymbol{\mu}^{(i)})_+,\Delta)}{\|S((M\boldsymbol{\mu}^{(i)})_+,\Delta)\|_2}$
  until Convergence
  compute i-th singular value as $\sigma_i \leftarrow \boldsymbol{\eta}^{(i)T}M\boldsymbol{\mu}^{(i)}$
  obtain residual as $M \leftarrow M - \sigma_i\boldsymbol{\eta}^{(i)}\boldsymbol{\mu}^{(i)T}$
end for
Output: $\{\boldsymbol{\mu}^{(i)}\}_{i=1}^{rank(M_x)}$ and $\{\boldsymbol{\eta}^{(i)}\}_{i=1}^{rank(M_z)}$

$$\max_{\boldsymbol{\eta}\in\mathbb{R}^{M_x},\boldsymbol{\mu}\in\mathbb{R}^{M_z}} \quad \mathbb{D}(K_x, K_z) = \boldsymbol{\eta}^{\mathrm{T}} M \boldsymbol{\mu} \tag{60a}$$

$$\text{subject to} \quad \boldsymbol{\eta} \geq 0,\ \boldsymbol{\mu} \geq 0,$$

$$\|\boldsymbol{\eta}\|_2 = \|\boldsymbol{\mu}\|_2 = 1, \tag{60b}$$

$$\|\boldsymbol{\eta}\|_1 \leq c_1, \|\boldsymbol{\mu}\|_1 \leq c_2, \tag{60c}$$

where $\|\mathbf{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ is the L^p-norm of the vector $\mathbf{x}$ and c_1 and c_2 are parameters (See also **Sec. Parameter Tuning by A Permutation Test**). This optimization problem is an example of penalized matrix decomposition with non-negativity constraints [9]. Accordingly, I can obtain optimal weight coefficients by performing singular value decomposition of matrix M under constraints. In this process, the i-th left singular vector $\boldsymbol{\eta}^{(i)} = (\eta_1^{(i)}, \ldots, \eta_{M_x}^{(i)})^T$ as well as the right singular vector $\boldsymbol{\mu}^{(i)} = (\mu_1^{(i)}, \ldots, \mu_{M_z}^{(i)})^T$ are obtained iteratively by Algorithm 4.

In Algorithm 4, S denotes the element-wise soft-thresholding operator: The m-th element of $S(\mathbf{a}, c)$ is given by $\mathrm{sign}(a_m)(|a_m| - c)_+$, where $(x)_+$ is x if $x \geq 0$ and 0 if $x < 0$. In each step, Δ is chosen by a binary search so that L1 constraints

$\|\boldsymbol{\eta}\|_1 \leq c_1$ and $\|\boldsymbol{\mu}\|_1 \leq c_2$ are satisfied. In general, the above iteration does not necessarily converge to a global optimum. For each iteration, I initialize $\boldsymbol{\eta}^{(i)}$ with a non-sparse, left singular vector of M, following the previous study, to obtain reasonable solutions [9].

5.3.2 The Second Stage: Kernel CCA

After learning kernels via penalized matrix decomposition as above, I perform the second stage of standard kernel CCA [102, 103] to obtain optimal coefficients $\boldsymbol{\alpha}^*$ and $\boldsymbol{\beta}^*$ (Eq. 54) with parameter κ for each pair of singular vectors $\{\boldsymbol{\eta}^{(i)}, \boldsymbol{\mu}^{(i)}\}_{i=1}^{rank(M)}$. Given test kernel $\{K_{x,test}^{(m)}\}_{m=1}^{M_x}$ and $\{K_{z,test}^{(l)}\}_{l=1}^{M_z}$, test correlation corresponding to the i-th singular vectors is defined as correlation between $\sum_{m=1}^{M_x} \eta_m K_{x,test}^{(m)} \boldsymbol{\alpha}^*$ and $\sum_{l=1}^{M_z} \mu_l K_{z,test}^{(l)} \boldsymbol{\beta}^*$.

5.3.3 Practical Solutions for TSKCCA Implementation

TSKCCA still has several options for sub-kernels to be designed manually. In this study, I focus on feature-wise kernel and pair-wise kernel defined in the following part.

Feature-wise kernel

Feature-wise kernel was introduced to perform feature-wise nonlinear Lasso [110]. In the previous study, using feature-wise kernels as sub-kernels in sparse MKL resulted in sparsity in terms of features since each sub-kernel corresponds to each feature. With x_{nm} and z_{nl} representing the m-th feature for $\mathbf{x}_n$ and l-th feature for $\mathbf{z}_n$, respectively, I adopt the following Gaussian kernel in this study:

$$[K_x^{(m)}]_{nn'} = \exp\left\{-\gamma_x (x_{nm} - x_{n'm})^2\right\} \tag{61a}$$

$$[K_z^{(l)}]_{nn'} = \exp\left\{-\gamma_z (z_{nl} - z_{n'l})^2\right\}, \tag{61b}$$

where γ_x and γ_z are width parameters. By applying feature-wise kernels, projection functions are restricted to additive models defined as $f^*(\mathbf{x}) = \sum_{m=1}^{p} f_m(\mathbf{x}_{\cdot m})$ and $g^*(\mathbf{z}) = \sum_{l=1}^{q} g_l(\mathbf{z}_{\cdot l})$, where $f_m : \mathbb{R} \to \mathbb{R}$ $(m = 1, \ldots, p)$ and $g_l : \mathbb{R} \to \mathbb{R}$ $(l = 1, \ldots, q)$ are certain nonlinear functions [2]. Note that the number of sub-kernels, M_x and M_z, are equivalent to the number of features, p and q, respectively.

Pair-wise kernel

I introduce pair-wise kernels as sub-kernels to consider cross-feature interactions among all possible pairs of features. Since the sparseness is induced to the weight of sub-kernels, the pair-wise kernels result in selecting relevant cross-feature interactions. Projection functions are defined as $f^*(\mathbf{x}) = \sum_{m<m'}^{p} f_{m,m'}(\mathbf{x}_{\cdot m}, \mathbf{x}_{\cdot m'})$ and $g^*(\mathbf{z}) = \sum_{l<l'}^{q} g_{l,l'}(\mathbf{z}_{\cdot l}, \mathbf{z}_{\cdot l'})$, where $f_{m,m'} : \mathbb{R}^2 \to \mathbb{R}$ and $g_{l,l'} : \mathbb{R}^2 \to \mathbb{R}$ are certain nonlinear functions with two dimensional inputs. Note that the number of sub-kernels, M_x and M_z, are, $p(p-1)/2$ and $q(q-1)/2$, respectively.

5.3.4 Preprocessing for MKL

I normalize the sub-kernels to have uniform variance in RKHS. This is an important procedure in the context of MKL because each feature-wise kernel has a different scale. This makes it difficult to evaluate weight coefficients [111]. To compensate for that, I calculate the variance σ^2 in RKHS as follows:

[2] In this thesis, $\mathbf{x}_{\cdot m}$ denotes the m-th feature of $\mathbf{x}$.

$$\sigma^2 = \frac{1}{N}\sum_{n}^{N}\|\phi(\mathbf{x}_n) - \frac{1}{N}\sum_{n'}^{N}\phi(\mathbf{x}_{n'})\|_2^2 \tag{62a}$$

$$= \frac{1}{N}\sum_{n}^{N}\phi(\mathbf{x}_n)^T\phi(\mathbf{x}_n) - \frac{1}{N^2}\sum_{n,n'}^{N}\phi(\mathbf{x}_{n'})^T\phi(\mathbf{x}_{n'}) \tag{62b}$$

$$= \frac{1}{N}\sum_{n}^{N}[K]_{nn} - \frac{1}{N^2}\sum_{n,n'}^{N}[K]_{nn'}. \tag{62c}$$

Dividing each sub-kernel by its variance $K \leftarrow \frac{K}{\sigma^2}$, I can achieve normalization of each sub-kernel.

5.3.5 Parameter Tuning by a Permutation Test

When the kernel matrix K_x (or K_y) is full rank, as is typically my case, KCCA with a small κ ($\kappa \ll 1$) can always find a solution such that the maximum canonical correlation nearly equals one. This property makes it difficult to tune the regularization parameters for the first stage c_1 and c_2. To solve the issue, I introduce a simple heuristics.

The key idea is to conduct a permutation test for deciding whether to reject a null hypothesis that the maximal canonical correlation induced by i-th singular vectors is no more than those attained when $\mathbf{x}$ and $\mathbf{z}$ are statistically independent. Since the p-value of this test is interpreted as the deviance between the actual outcome and those expected under the null hypothesis, I use it as a score to evaluate the significance of i-th singular vectors where smaller p-value is more significant.

Algorithm 5 summarizes my implementation for the permutation test. Only for the first singular vectors $\boldsymbol{\eta}^{(1)}$ and $\boldsymbol{\mu}^{(1)}$, this procedure is applied to various pairs of (c_1, c_2) that satisfy the constraints of $1 \leq c_1 \leq \sqrt{M_x}$ and $1 \leq c_2 \leq \sqrt{M_y}$ [9]. Among them, the pair with the lowest p-value is chosen as the optimal

Algorithm 5 A Permutation Test

Input: $\{K_x^{(m)}\}_{m=1}^{M_x}$, $\{K_z^{(l)}\}_{l=1}^{M_z}$, c_1, and c_2
$c = \mathrm{Cor}(\{K_x^{(m)}\}_{m=1}^{M_x}, \{K_z^{(l)}\}_{l=1}^{M_z}, \boldsymbol{\eta}^{(i)}, \boldsymbol{\mu}^{(i)})$
for $b = 1$ **to** B **do**
  permute the samples of $\mathbf{x}$ and calculate $\{\tilde{K}_x^{(m)}\}_{m=1}^{M_x}$
  obtain $\tilde{M}$ where $[\tilde{M}]_{ij} = \frac{\mathrm{Tr}(\tilde{K}_x^{(m)} H K_z^{(l)} H)}{(N-1)^2}$
  perform the first stage; matrix decomposition of $\tilde{M}$ to obtain $\tilde{\boldsymbol{\eta}}^{(i)}$ and $\tilde{\boldsymbol{\mu}}^{(i)}$
  calculate $c_b = \mathrm{Cor}(\{\tilde{K}_x^{(m)}\}_{m=1}^{M_x}, \{K_z^{(l)}\}_{l=1}^{M_z}, \tilde{\boldsymbol{\eta}}^{(i)}, \tilde{\boldsymbol{\mu}}^{(i)})$
end for
$p = \frac{\sum_{b=1}^{B} I(|c_b| > |c|)}{B+1}$
Output: p

parameters of c_1 and c_2.

For simplicity, other parameters, such as γ in the Gaussian kernel and κ in KCCA, are fixed heuristically. γ^{-1} is set to the median of the Euclidean distance between data points and κ is set to 0.02 as recommended in the previous study [103].

5.4 Results

In this section, I experimentally evaluated the performance of my proposed TSKCCA, SAFCCA [105], and other methods using synthetic data and nutrigenomic experimental data.

5.4.1 Dataset 1: Single nonlinear association

To evaluate the ability to extract a single nonlinear association, we generated simple synthetic data which consisted of a single pair of relevant features in quadratic association and noise, in which standard CCA and KCCA are known to performance poorly [105]. Let $N(\mu, s^2)$ and $U(\mathcal{A})$ denote the normal distribution with mean μ, variance s^2, and uniform distribution supported in $\mathcal{A}$, respectively. The synthetic data were generated as

$$
\begin{aligned}
\mathbf{x}_{.m} &\sim U([-0.5, 0.5]) \qquad m = 1, ..., D \\
\mathbf{z}_{.1} &= \mathbf{x}_{.1}^2 + \boldsymbol{\epsilon} \\
\mathbf{z}_{.l} &\sim U([-0.5, 0.5]) \qquad l = 2, ..., D \\
\boldsymbol{\epsilon} &\sim N(0, s^2),
\end{aligned}
$$

where D was the total number of dimensions and ϵ was independent noise.

The optimal model in each method was trained using N training samples. Here, I assumed $c_1 = c_2$ in the range of $1 \leq c_1, c_2 \leq \frac{\sqrt{D}}{2}$ and obtained optimal values using a permutation test with $B = 100$. The test correlation was evaluated with separate 100 test samples, averaged over 100 simulation runs as I varied the number of dimensions, the sample size, and the noise level.

Fig. 13 shows the test correlations achieved by TSKCCA and SAFCCA with different data dimensions D, sample size N, and noise level s. In the first stage, my method selected only two sub-kernels, corresponding to $\mathbf{x}_1$ and $\mathbf{z}_1$, among $2 \times D$ sub-kernels in the first stage, especially in the case of $N = 100$ and $N = 150$. As a result, it achieved better test correlation than SAFCCA, especially with high-dimensional data, indicating that my method was sufficiently robust.

In addition, Fig. 14 shows average computation time for each method over 100 simulation runs with dataset 1. Computation time of TSKCCA was comparable with that of SAFCCA, and could scale up with the feature size. Note that all the experiments were performed on a MacBook Pro with Intel Core i7 (2.9GHz dual core processor with 4MB L3 cache) with 8GB main memory. All the simulation programs were implemented in MATLAB®.

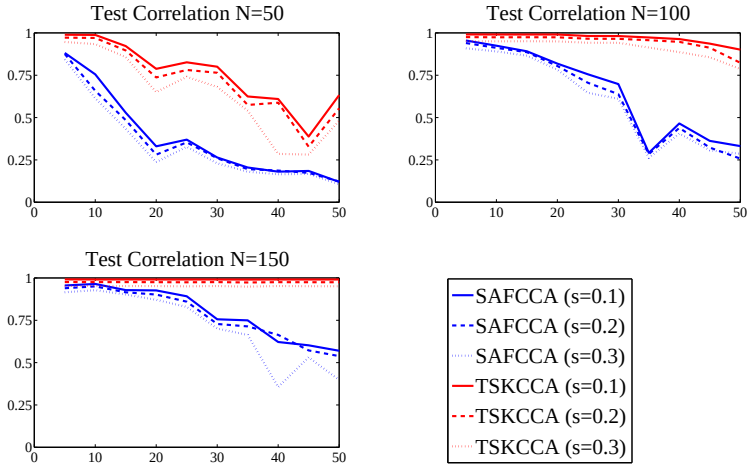

Figure 13: Comparison of test correlation averaged over simulation runs in Data 1. The horizontal axis denotes the number of dimensions D, and the vertical axis denotes test correlations. The number of training samples is 50, 100, and 150. TSKCCA outperforms SAFCCA, especially with high-dimensional data.

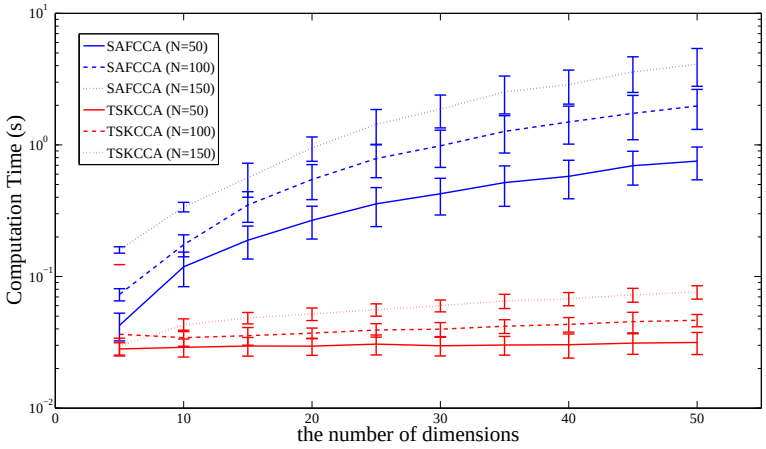

Figure 14: Comparison of computation time for Data 1. The horizontal axis denotes the number of dimensions D, and the vertical axis denotes computation time in log-scale. The number of training samples is 50, 100, 150 for SAFCCA and TSKCCA. Computation time of TSKCCA is moderate and can be scaled.

5.4.2 Dataset 2: Multiple nonlinear associations

To test whether my method could extract multiple nonlinear associations precisely, I generated the following data:

$$
\begin{aligned}
\mathbf{x}_{\cdot m} &\sim U([-0.5, 0.5]) \qquad m = 1, ..., 25 \\
\mathbf{z}_{\cdot 1} &= \mathbf{x}_{\cdot 1} + \exp(-\mathbf{x}_{\cdot 4}^2) + \boldsymbol{\epsilon}_1 \\
\mathbf{z}_{\cdot 2} &= \mathbf{x}_{\cdot 2}^2 + \sin(\pi \mathbf{x}_{\cdot 5}/2) + \boldsymbol{\epsilon}_2 \\
\mathbf{z}_{\cdot 3} &= |\mathbf{x}_{\cdot 3}| + 1/(1 + \exp(-5\mathbf{x}_{\cdot 6})) + \boldsymbol{\epsilon}_3 \\
\mathbf{z}_{\cdot l} &\sim U([-0.5, 0.5]) \qquad l = 4, ..., 25 \\
\boldsymbol{\epsilon}_l &\sim N(0, 0.1^2) \qquad l = 1, 2, 3.
\end{aligned}
$$

First, I performed a permutation test with $B = 1000$ for ten singular vectors $\{\boldsymbol{\eta}^{(i)}, \boldsymbol{\mu}^{(i)}\}_{i=1}^{10}$ corresponding to the ten highest singular values of M given by Eq. (59). P-values of the top three were significant ($p < 0.001$) and the rest were non-significant. This result suggests that only the three singular vectors included nonlinear associations.

Fig. 15 shows the transformations $f(\mathbf{x})$ and $g(\mathbf{z})$ obtained with TSKCCA. In the first singular vectors, the contributions of η_1^1, η_4^1 and μ_1^1 were dominant, indicating that $\mathbf{x}_{\cdot 1}$,$\mathbf{x}_{\cdot 4}$ and $\mathbf{z}_{\cdot 1}$ were associated. The contributions of η_2^2, η_5^2 and μ_2^2 in the second singular vectors were also dominant, indicating that $\mathbf{x}_{\cdot 2}$, $\mathbf{x}_{\cdot 5}$ and $\mathbf{z}_{\cdot 2}$ were associated. Finally, the contributions of η_3^3, η_6^3 and μ_3^3 in the third singular vectors were dominant, indicating that $\mathbf{x}_{\cdot 3}$, $\mathbf{x}_{\cdot 6}$ and$\mathbf{z}_{\cdot 3}$ were associated. Some singular vectors averaged over 100 simulation runs are listed in Table 7. My results suggest that TSKCCA achieved feature selection precisely.

I further evaluated test correlation, precision, and recall averaged over 20 simulation runs. Table 8 shows that SAFCCA failed to detect all relevant fea-

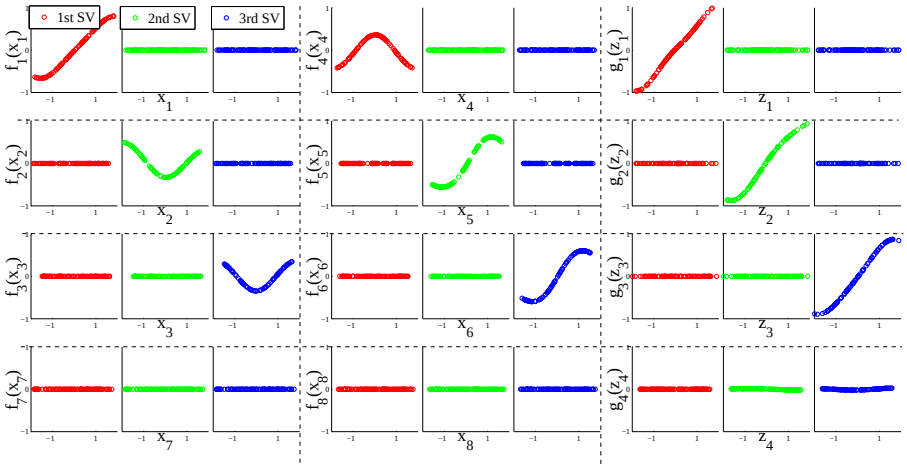

Figure 15: Transformations $f(\mathbf{x})$ and $g(\mathbf{z})$ obtained with TSKCCA. The top three rows and the bottom row show the resulting functions corresponding to relevant and irrelevant features, respectively.

Table 7: Feature selection through singular vectors (SVs) in Data 2. These results show mean weight coefficients (standard deviation) in 100 simulation runs. Significant weight coefficients are bold faced.

	1st SV ($\eta^{(1)}$)	2nd SV ($\eta^{(2)}$)	3rd SV ($\eta^{(3)}$)
η_1	**0.98 (0.002)**	0.00 (0.018)	0.00 (0.001)
η_2	0.00 (0.003)	**0.21 (0.033)**	0.00 (0.001)
η_3	0.00 (0.001)	0.00 (0.010)	**0.22 (0.029)**
η_4	**0.22 (0.013)**	0.00 (0.017)	0.00 (0.005)
η_5	0.00 (0.000)	**0.98 (0.004)**	0.00 (0.005)
η_6	0.00 (0.004)	0.00 (0.002)	**0.98 (0.003)**
	1st SV ($\mu^{(1)}$)	2nd SV ($\mu^{(2)}$)	3rd SV ($\mu^{(3)}$)
μ_1	**0.99 (0.005)**	0.01 (0.022)	0.01 (0.014)
μ_2	0.01 (0.027)	**0.99 (0.004)**	0.01 (0.015)
μ_3	0.01 (0.024)	0.01 (0.018)	**0.99 (0.003)**
μ_4	0.01 (0.023)	0.01 (0.026)	0.01 (0.017)

Table 8: Comparison of test correlation, precision, and recall in Data 2. TSKCCA can identify most relevant features through three significant singular vectors, while SAFCCA can only identify a small set of them.

	Correlation	Precision	Recall
TSKCCA	0.9670	0.9163	1
	0.9636		
	0.9732		
SAFCCA	0.7585	0.6350	0.4375

tures because it is not able to obtain multiple canonical correlations, while my method detected 9 relevant sub-kernels out of 50 in the first stage in most runs. Note that the precision is the fraction of retrieved features that are relevant and recall is the fraction of relevant features that are retrieved.

5.4.3 Dataset 3: Feature interactions

To assess the capability of TSKCCA in discovering nonlinear interactions, I generated data with a product term:

$$\begin{aligned} \mathbf{x}_{\cdot m} &\sim U([-0.5, 0.5]) \qquad m = 1, ..., D \\ \mathbf{z}_{\cdot 1} &= \mathbf{x}_{\cdot 1}\mathbf{x}_{\cdot 2} + \boldsymbol{\epsilon} \\ \mathbf{z}_{\cdot l} &\sim U([-0.5, 0.5]) \qquad l = 2, ..., D \\ \boldsymbol{\epsilon} &\sim N(0, 0.1^2), \end{aligned}$$

where D was the number of dimensions. For this dataset, I used feature-wise kernels and pair-wise kernels as sub-kernels in order to handle both single feature effects and cross-feature interactions like the term $\mathbf{x}_{\cdot 1}\mathbf{x}_{\cdot 2}$. There were $D + D \times (D-1)/2$ sub-kernels, the weight coefficients of which were optimized in my method.

First, to evaluate the performance of my method with feature-wise and pair-wise kernels, I obtained test correlations evaluated by individual test data ($N =$

100) in different numbers of dimensions D. Next, to evaluate the accuracy of feature selection of the model, I assessed recall and precision. Average test correlations, recall, and precision over 100 simulation runs are shown in Fig. 16. Our results illustrate that in the case of $D < 10$ (i.e. the number of sub-kernels is less than $10+10\times 9/2 = 55$), my method successfully determined the relation between $\mathbf{z}_{\cdot 1}$ and $\mathbf{x}_{\cdot 1}\mathbf{x}_{\cdot 2}$.

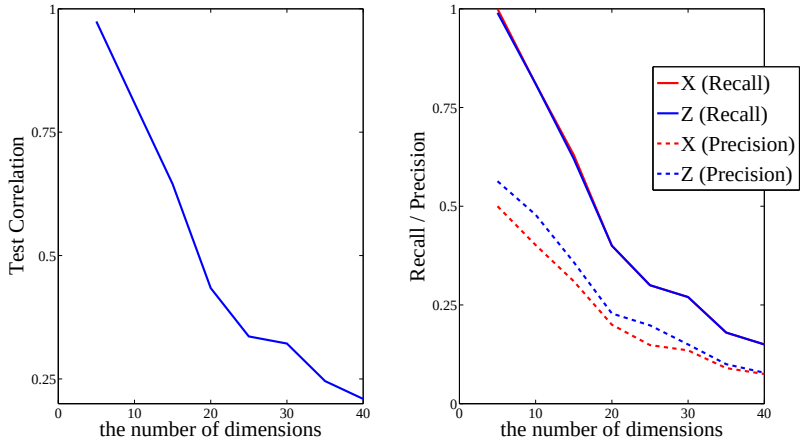

Figure 16: The performance of pair-wise kernels in Data 3. (Left) Test correlations averaged over 100 simulation runs in different numbers of dimensions. (Right) Recall and precision averaged over 100 simulation runs in different numbers of dimensions. Our method successfully extracts nonlinear associations with relevant features.

5.4.4 Dataset 4: Nutrigenomic data

I then analyzed a nutrigenomic dataset from a previous mouse study [112, 11]. In this study, expression of 120 genes in liver cells that would be relevant in the context of nutrition and concentrations of 21 hepatic fatty acids were measured on 20 wild-type mice and 20 PPARα-deficient mice. Mice of each genotype were fed 5 different diets with different levels of fat. For matrix notation, gene expression data were denoted by $X \in \mathbb{R}^{40\times 120}$, and data regarding concentrations

of fatty acids was denoted by $Z \in \mathbb{R}^{40\times 21}$. Data were standardized to have a mean of zero and unit variance in each dimension. Several linear correlations between X and Z were detected by applying a regularized version of the linear CCA [12, 11].

First, I performed a permutation test for sparse CCA, KCCA, SAFCCA, and TSKCCA on parameters defined by equally-spaced grid points in order to identify significant associations in these data. In KCCA and SAFCCA, there were no significant associations; thus, I focused on sparse CCA and TSKCCA in the following analysis. I identified two significant linear associations in sparse CCA ($p < 0.001$ using a permutation test) and one nonlinear association in TSKCCA ($p = 0.0067$ using a permutation test) with $c_1 = 2.6257$ and $c_2 = 1.9275$.

Fig. 17 and 18 show the results of feature selection of sparse CCA and TSKCCA, respectively. Genes selected by the first singular vector of my method have different expression levels in different genotypes (marked with asterisk), suggesting that my method successfully extracted the nonlinear correlation associated with genotypes.

For further analysis, cross-validation was performed in 100 runs. In each runs, 40 samples were randomly split into 30 training samples used for fitting models and 10 validation samples used for evaluating the canonical correlation for fitted models. Fig. 19 shows box plots of correlation coefficients in sparse CCA and TSKCCA. Left one represents the first canonical correlation coefficient in sparse CCA and right one represents correlation coefficient obtained with the first singular vectors. Significantly higher test correlation ($p < 10^{-6}$ with a t-test) were achieved by the first singular vectors of TSKCCA, indicating that it avoided overfitting despite having nonlinearity.

To account for interactions between features into my model, I calculated

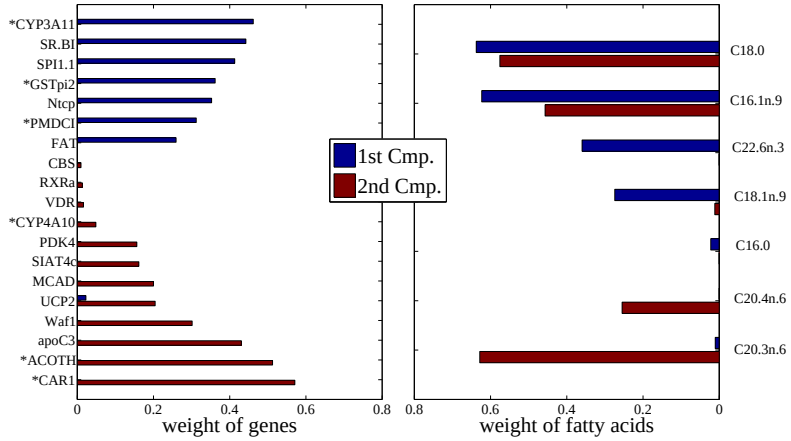

Figure 17: Feature selection of sparse CCA in nutrigenomic data. Left and right panels show selected genes and fatty acids, respectively. Genes marked with asterisks show significantly different expression in different genotypes.

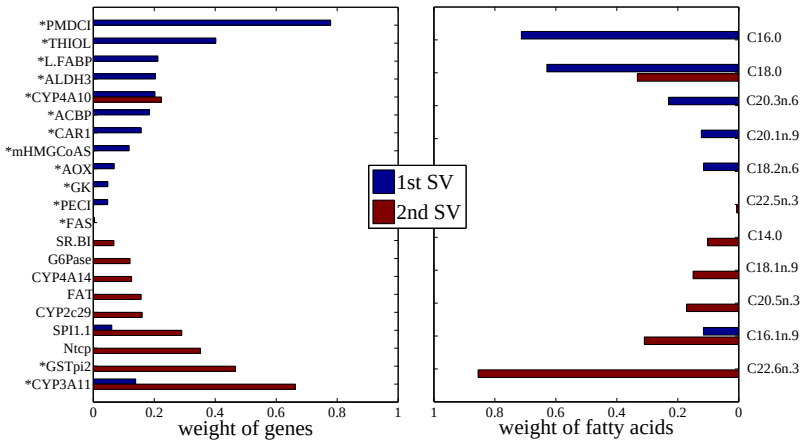

Figure 18: Feature selection of TSKCCA using nutrigenomic data. Left and right panels show selected genes and fatty acids, respectively. Genes marked with asterisks show significantly different expression in different genotypes. The left panel shows that the 1st singular vector extracts nonlinear correlations associated with the genotype.

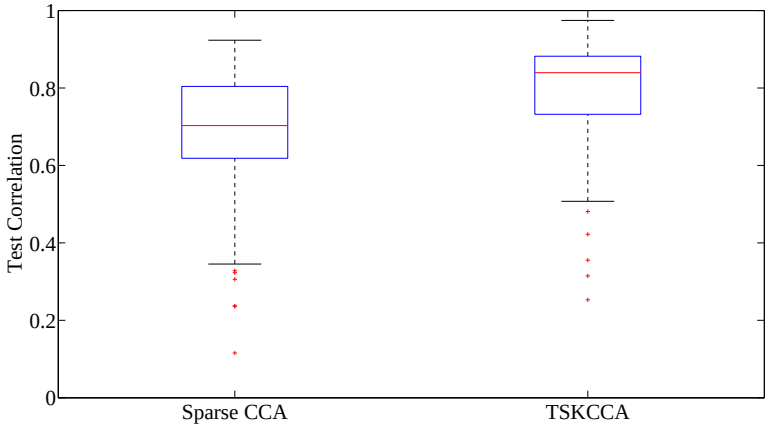

Figure 19: Box plot of test correlations in nutrigenomic data. Left and right panels show the box plot of 100 times test correlation using sparse CCA and TSKCCA, respectively. TSKCCA achieves significantly higher test correlation through its first weight vector ($p < 10^{-6}$ with a t-test).

pair-wise kernels for nutrigenomic data. Although the number of sub-kernels was huge ($120 + 120 \times 119/2 = 7260$ sub-kernels for genes, $21 + 21 \times 20/2 = 231$ sub-kernels for fatty acids), TSKCCA successfully extracted a significant association ($p < 0.001$ using a permutation test). To evaluate the stability of feature selection, I performed TSKCCA on 1000 runs with data generated by random sampling of empirical data with replacement. Table 9 shows the frequencies of features (i.e. pairs of features) selected across 1000 runs, suggesting that *PMDCI* played an important role within the interactions.

5.5 Discussion

Other researchers have employed the sparse additive model [106] to extend KCCA to high-dimensional problems, and have defined two equivalent formulations, such as sparse additive functional CCA (SAFCCA) and sparse additive kernel CCA (SAKCCA) [105]. The former was defined in a second order

Table 9: Frequency of selection per sub-kernel corresponding to genes (left) and fatty acids (right) in nutrigenomic data.

genes / pair of genes	freq.	fatty acids / pair of fatty acids	freq.
PMDCI	643	C16.0-C18.0	622
CAR1-PMDCI	564	C18.0	485
PMDCI-THIOL	563	C16.0-C20.3n.6	429
ACBP-PMDCI	473	C16.0	340
L.FABP-PMDCI	451	C18.0-C20.3n.6	315
CYP4A10-PMDCI	379	-	-
CYP3A11-PMDCI	370	-	-
ALDH3-PMDCI	369	-	-
Ntcp-PMDCI	354	-	-
PMDCI-SPI1.1	347	-	-
ACOTH-PMDCI	330	-	-
PMDCI-SR.BI	306	-	-

Sobolev space and solved using the biconvex back-fitting procedure. The latter, defined in RKHS, was derived by applying representer theorem to the former. Given some function $f_m \in \mathbb{H}_m$, these algorithms optimize the additive model, $f_1 \in \mathbb{H}_1, f_2 \in \mathbb{H}_2, ..., f_p \in \mathbb{H}_p$. In contrast, my formulation supposes an additive kernel, such as $\sum \eta_m K_m$ associated with RKHS $\mathbb{H}_{add}$ and finds correlations in this space. This method enables us to reveal multiple components of associations.

Some problems specific to KCCA, such as choosing two parameters (i.e. regularization parameter κ and the width parameter γ) and the number of components, remain unsolved. While cross validation is applicable to set these values [113], they are fixed for simplicity in my study, based on the previous study [103].

Next, I discuss the validity of feature selection in nutrigenomic data performed using sparse CCA and TSKCCA. In the original study, the authors focused on the role of PPARα as a major transcriptional regulator of lipid metabolism and determined that PPARα regulates the expression of many genes

in mouse liver under lower dietary fat conditions [112]. They provided a list of genes that have significantly different expression levels between wild-type and PPARα-deficient mice. While only a few genes selected by sparse CCA were included in the list, 13 out of 14 genes selected with the 1st singular vector in TSKCCA were included in the list. This result shows that TSKCCA successfully extracts meaningful nonlinear associations induced by PPARα-deficiency.

Moreover, in my analysis of pair-wise kernels, most of the frequently selected pairs of genes retained *PMDCI* known as a sort of enoyl-CoA isomerases involved in β-oxidation of polyunsaturated fatty acids. This implies that the interactions of *PMDCI* and other genes contribute to lipid metabolism in PPARα-deficient mice.

Many variants of sub-kernels, such as string kernels or graph kernels, can be employed in the same framework. In the field of biomedical engineering, Yamanishi et al. adopted integrated KCCA (IKCCA), which exploited the simple sum of multiple kernels to combine many sorts of biological data [104]. This technique can be improved by optimizing weight coefficients of each kernel in the frame of TSKCCA. Finally, if kernels are defined on groups of features, it enables us to perform group-wise feature selection, just like group sparse CCA [4, 53, 16]. It is beneficial to consider group-wise feature selection for biomarker detection problems.

5.6 Conclusions

This thesis proposesed a novel extension of kernel CCA that I call two-stage kernel CCA, which is able to identify multiple canonical variables from sparse features. This method optimizes the sparse weight coefficients of pre-specified sub-kernels as a sparse matrix decomposition before performing standard kernel CCA. This procedure enables us to achieve interpretability by removing

irrelevant features in the context of nonlinear correlational analysis.

Through three numerical experiments, I have demonstrated that TSKCCA is more useful for higher dimensional data and for extracting multiple nonlinear associations than an existing method, SAFCCA. Using nutrigenomic data, my results show that TSKCCA can retrieve information about genotype and may reveal an interactive mechanism of lipid metabolism in PPARα-deficient mice.

6 Conclusion Remarks

In this thesis, to evaluate the state of depression patients based on objective physiological data such as brain activity, I explored interpretable machine learning algorithms in high-dimensional datasets.

In chapter 3, I exploited some priori knowledge about anatomical segmentation of human brain and examined a method using sub-kernels that corresponds one-to-one to the anatomical brain regions. I confirmed that it was superior to the existing method such as logistic regression with L1 regularization for the purpose of judging the presence or absence of disease from fMRI data of depressed patients and healthy controls. Although the accuracy was almost comparable with SVM, thanks to the application of the regional-wise sub-kernels, I successfully identified the relevant anatomical regions, such as *left precentral gyrus* and *left precuneus*.

In chapter 4, I assumed that the resting-state functional connectivity data and depression-related clinical scores, which have strong correlations with each other, could be projected on a common low-dimensional latent space. In order to take account of this assumption, I examined the application of the partial least squares regression and obtained higher accuracy in comparison with ordinary least squares regression. Moreover, by looking at the regression coefficients, I successfully identified some relevant functional connectivity to depression.

In chapter 5, I examined the problems that could occur when nonlinear correlation analysis was applied to high dimensional datasets. I introduced a novel method to obtain an appropriate design of kernels that remove irrelevant information in the framework of multiple kernel learning as in chapter 3. I applied this novel method to some synthetic data and mouse nutrigenomic data to evaluate the capability of removing noise and obtaining nonlinear associations.

In this thesis, I investigated the relationship between diagnoses, clinical

scores and fMRI data in chapters 3 and 4. Since these studies were not enough to obtain deep understanding of major depression in a data driven manner, other types of physiological data such as SNPs should be considered like [114]. In future work, my proposed method called two stage kernel CCA in chapter 5 can be useful to combine multiple high dimensional datasets as a multimodal method. To that end, there are still many remaining issues. The most important one is about the design of the sub-kernels. For the purpose of handling multiple high dimensional datasets such as SNPs and brain activity, further design of sub-kernels should be investigated. One of the valuable candidate is group sparseness applied to analysis of SNPs and fMRI data in [16, 115]. As a natural extension of group sparseness, it is possible to divide variables into several groups and define the sub-kernels for each group. Since the final interpretability of results depends on how to divide the variables into groups, further investigation is required in future study. As a conclusion, further investigation is required to combine multiple datasets using two stage kernel CCA to find some new insights about depression.

Appendix

Table 10: Root mean squared error (output-age).

	BDI-II	SHAPS	PANAS(n)	age
OLS	11.6±1.35	7.33±0.81	8.91±0.982	9.89±1.15
PLS	9.71±1.11	6.44±0.77	7.38±0.817	9.40±1.08
KPLS-Poly(2)	**9.56±1.08**	**6.11±0.673**	**7.29±0.807**	9.51±1.09
KPLS-Poly(3)	10.3±1.15	6.43±0.702	7.57±0.831	9.41±1.07
KPLS-Gauss	9.88±1.11	6.49±0.706	7.42±0.821	**9.29±1.05**

Table 11: Root mean squared error (input-age).

	BDI-II	SHAPS	PANAS(n)	age
OLS	12.9±1.48	7.90±0.895	9.98±1.11	-
PLS	11.7±1.30	7.52±0.822	8.74±0.962	-
KPLS-Poly(2)	11.7±1.30	7.70±0.842	8.83±0.963	-
KPLS-Poly(3)	12.3±1.36	7.93±0.863	9.21±1.01	-
KPLS-Gauss	10.3±1.17	6.78±0.759	7.69±0.853	-

Table 12: Root mean squared error (no-age).

	BDI-II	SHAPS	PANAS(n)	age
OLS	12.5±1.43	7.69±0.868	9.69±1.07	-
PLS	11.4±1.27	7.44±0.804	8.59±0.946	-
KPLS-Poly(2)	11.0±1.22	7.17±0.777	8.24±0.916	-
KPLS-Poly(3)	11.5±1.31	7.32±0.812	8.61±0.973	-
KPLS-Gauss	11.6±1.29	7.87±0.853	8.97±0.971	-

Table 13: Classification Performance (Mean±SE over nested leave-one-out cross validation). KPLS-Poly(2) followed by LDA significantly outperformed direct LDA, SVM, and OLS followed by LDA in accuracy (adjusted for multiplicity using the Bonferroni-Holm method with significance level $\alpha = 0.05$).

	accuracy (%)	sensitivity (%)	specificity (%)
direct LDA	57.7±4.45	53.4±6.55	61.5±6.03
direct SVM	69.1±4.17	69.0±6.07	69.2±5.72
OLS+LDA	62.6±4.36	62.1±6.37	63.1±5.99
PLS+LDA	72.4±4.03	74.1±5.75	70.8±5.65
KPLS-Poly(2)+LDA	**80.5±3.57**	**81.0±5.15**	**80.0±4.96**
KPLS-Poly(3)+LDA	76.4±3.83	74.1±5.75	78.5±5.10
KPLS-Gauss+LDA	71.5±4.07	70.7±5.98	72.3±5.55

Table 14: Classification Performance without subjects over 60 (Mean±SE over nested leave-one-out cross validation). KPLS-Poly(2) followed by LDA significantly outperformed direct LDA and OLS followed by LDA in accuracy (adjusted for multiplicity using the Bonferroni-Holm method with significance level $\alpha = 0.05$).

	accuracy (%)	sensitivity (%)	specificity (%)
direct LDA	66.1±4.27	66.7±6.19	65.6±5.89
direct SVM	67.8±4.21	74.1±5.75	62.3±6.01
OLS+LDA	61.7±4.38	61.1±6.40	62.3±6.01
PLS+LDA	70.4±4.11	72.2±5.88	68.9±5.74
KPLS-Poly(2)+LDA	**78.3±3.72**	**77.1±5.29**	**80.0±5.22**
KPLS-Poly(3)+LDA	73.9±3.96	68.5±6.10	78.7±5.08
KPLS-Gauss+LDA	70.4±4.11	68.5±6.10	72.1±5.56

Acknowledgement

I would first like to thank Dr. Kenji Doya at Okinawa Institute of Science and Technology Graduate University for his financial support and academic supervision. I could achieve results of better quality thanks to him.

I am very grateful to Dr. Shigeto Yamawaki, Dr. Yasumasa Okamoto, Dr. Go Okada, and Dr. Masahiro Takamura of Department of Psychiatry and Neurosciences at Hiroshima University for offering their fMRI datasets used in my research and valuable feedback from the perspective of psychiatry.

I would also like to thank Dr. Junichiro Yoshimoto of Graduate School of Information Science at Nara Institute of Science and Technology for his practical advice and patient support. I could complete my thesis thanks to valuable discussions with him.

I would like to express my gratitude to Dr. Shin Ishii, Dr. Manabu Kano, and Dr. Hidetoshi Shimodaira for their kind feedback to improve the quality of my thesis.

At the end of my acknowledgement, I would like to thank all members in Neural Computation Unit at Okinawa Institute of Science and Technology Graduate University and Integrated Systems Biology Laboratory at Kyoto University. My thesis would not have been completed without their kind and patient support.

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.

[2] Fu WJ. Penalized regressions: the bridge versus the lasso. Journal of computational and graphical statistics. 1998;7(3):397–416.

[3] Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. The Annals of statistics. 2004;32(2):407–499.

[4] Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006;68(1):49–67.

[5] Shi W, Lee KE, Wahba G. Detecting disease-causing genes by LASSO-Patternsearch algorithm. In: BMC proceedings. vol. 1. BioMed Central; 2007. p. 1.

[6] Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009;25(6):714–721.

[7] Puniyani K, Kim S, Xing EP. Multi-population GWA mapping via multi-task regularized regression. Bioinformatics. 2010;26(12):i208–i216.

[8] Hotelling H. Relations between two sets of variates. Biometrika. 1936;p. 321–377.

[9] Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics. 2009 Jul;10(3):515–34.

[10] Soneson C, Lilljebjörn H, Fioretos T, Fontes M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. BMC bioinformatics. 2010;11(1):1.

[11] González I, Déjean S, Martin PG, Baccini A, et al. CCA: An R package to extend canonical correlation analysis. Journal of Statistical Software. 2008;23(12):1–14.

[12] González I, Déjean S, Martin PG, Gonçalves O, Besse P, Baccini A. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. Journal of Biological Systems. 2009;17(02):173–199.

[13] Correa NM, Li YO, Adali T, Calhoun VD. Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to detection of associative networks in schizophrenia. IEEE journal of selected topics in signal processing. 2008;2(6):998–1007.

[14] Correa NM, Adali T, Li YO, Calhoun VD. Canonical correlation analysis for data fusion and group inferences. IEEE signal processing magazine. 2010;27(4):39–50.

[15] Waaijenborg S, Verselewel de Witt Hamer PC, Zwinderman AH. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. Statistical Applications in Genetics and Molecular Biology. 2008;7(1).

[16] Lin D, Zhang J, Li J, Calhoun VD, Deng HW, Wang YP. Group sparse canonical correlation analysis for genomic data integration. BMC bioinformatics. 2013;14(1):245.

[17] Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Medical image analysis. 2014;18(6):891–902.

[18] Wold H. Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach Perspectives in Probability and Statistics. In: Gani J, editor. Papers in Honour of M. S. Bartlett. Academic Press, London; 1975. p. 520–540.

[19] Wold H. Partial least squares. Encyclopedia of statistical sciences. 1985;.

[20] Manne R. Analysis of two partial-least-squares algorithms for multivariate calibration. Chemometrics and Intelligent Laboratory Systems. 1987;2(1):187–197.

[21] Rosipal R, Trejo LJ. Kernel partial least squares regression in reproducing kernel hilbert space. The Journal of Machine Learning Research. 2002;2:97–123.

[22] Rännar S, Lindgren F, Geladi P, Wold S. A PLS kernel algorithm for data sets with many variables and fewer objects. Part 1: Theory and algorithm. Journal of Chemometrics. 1994;8(2):111–125.

[23] McIntosh A, Bookstein F, Haxby JV, Grady C. Spatial pattern analysis of functional brain images using partial least squares. NeuroImage. 1996;3(3):143–157.

[24] Ziegler G, Dahnke R, Winkler AD, Gaser C. Partial least squares correlation of multivariate cognitive abilities and local brain structure in children and adolescents. NeuroImage. 2013;82:284 – 294.

[25] Price J, Ziolko S, Weissfeld L, Klunk W, Lu X, Hoge J, et al. Quantitative and statistical analyses of PET imaging studies of amyloid deposition in

humans. In: Nuclear Science Symposium Conference Record, 2004 IEEE. vol. 5. IEEE; 2004. p. 3161–3164.

[26] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics. 2002;18(1):39–50.

[27] Giessing C, Fink GR, Rösler F, Thiel CM. fMRI data predict individual differences of behavioral effects of nicotine: a partial least square analysis. Journal of cognitive neuroscience. 2007;19(4):658–670.

[28] Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics. 2002;18(12):1625–1632.

[29] Schölkopf B, Herbrich R, Smola AJ. A generalized representer theorem. In: International Conference on Computational Learning Theory. Springer; 2001. p. 416–426.

[30] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press; 2000.

[31] Mukherjee S, Tamayo P, Slonim D, Verri A, Golub T, Mesirov J, et al. Support vector machine classification of microarray data. 1999;.

[32] Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences. 2000;97(1):262–267.

[33] Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures.

Proceedings of the National Academy of Sciences. 2001;98(26):15149–15154.

[34] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16(10):906–914.

[35] Zavaljevski N, Stevens FJ, Reifman J. Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. Bioinformatics. 2002;18(5):689–696.

[36] Bradford JR, Westhead DR. Improved prediction of protein–protein binding sites using a support vector machines approach. Bioinformatics. 2005;21(8):1487–1494.

[37] Mourão-Miranda J, Hardoon DR, Hahn T, Marquand AF, Williams SC, Shawe-Taylor J, et al. Patient classification as an outlier detection problem: an application of the one-class support vector machine. NeuroImage. 2011;58(3):793–804.

[38] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine learning. 2002;46(1-3):389–422.

[39] Kondor RI, Lafferty J. Diffusion kernels on graphs and other discrete input spaces. In: ICML. vol. 2; 2002. p. 315–322.

[40] Leslie CS, Eskin E, Cohen A, Weston J, Noble WS. Mismatch string kernels for discriminative protein classification. Bioinformatics. 2004;20(4):467–476.

[41] Gordon J, Towsey M. SVM based prediction of bacterial transcription start sites. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer; 2005. p. 448–453.

[42] Vert JP, Kanehisa M. Graph-driven feature extraction from microarray data using diffusion kernels and kernel CCA. In: Advances in neural information processing systems; 2002. p. 1425–1432.

[43] Yamanishi Y, Vert JP, Kanehisa M. Protein network inference from multiple genomic data: a supervised approach. Bioinformatics. 2004;20(suppl 1):i363–i370.

[44] Marquand AF, Mourão-Miranda J, Brammer MJ, Cleare AJ, Fu CH. Neuroanatomy of verbal working memory as a diagnostic biomarker for depression. Neuroreport. 2008;19(15):1507–1511.

[45] Fu CH, Mourao-Miranda J, Costafreda SG, Khanna A, Marquand AF, Williams SC, et al. Pattern classification of sad facial processing: toward the development of neurobiological markers in depression. Biological psychiatry. 2008;63(7):656–662.

[46] Lanckriet GR, Cristianini N, Bartlett P, Ghaoui LE, Jordan MI. Learning the kernel matrix with semidefinite programming. Journal of Machine learning research. 2004;5(Jan):27–72.

[47] Bach FR, Lanckriet GR, Jordan MI. Multiple kernel learning, conic duality, and the SMO algorithm. In: Proceedings of the twenty-first international conference on Machine learning. ACM; 2004. p. 6.

[48] Hinrichs C, Singh V, Xu G, Johnson SC, Initiative ADN, et al. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. Neuroimage. 2011;55(2):574–589.

[49] Castro E, Gómez-Verdejo V, Martínez-Ramón M, Kiehl KA, Calhoun VD. A multiple kernel learning approach to perform classification of groups from complex-valued fMRI data analysis: Application to schizophrenia. NeuroImage. 2014;87:1–17.

[50] Suzuki T, Tomioka R. SpicyMKL: a fast algorithm for multiple kernel learning with thousands of kernels. Machine learning. 2011;85(1-2):77–108.

[51] Tomioka R, Sugiyama M. Dual-augmented Lagrangian method for efficient sparse reconstruction. IEEE Signal Processing Letters. 2009;16(12):1067–1070.

[52] Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage. 2002;15(1):273–289.

[53] Bach FR. Consistency of the group lasso and multiple kernel learning. The Journal of Machine Learning Research. 2008;9:1179–1225.

[54] Shimizu Y, Yoshimoto J, Toki S, Takamura M, Yoshimura S, Okamoto Y, et al. Toward probabilistic diagnosis and understanding of depression based on functional MRI data analysis with logistic group LASSO. PloS one. 2015;10(5):e0123524.

[55] Heath S, McMahon KL, Nickels L, Angwin A, MacDonald AD, van Hees S, et al. Neural mechanisms underlying the facilitation of naming in aphasia using a semantic task: an fMRI study. BMC neuroscience. 2012;13(1):1.

[56] Utevsky AV, Smith DV, Huettel SA. Precuneus is a functional core of the default-mode network. The Journal of Neuroscience. 2014;34(3):932–940.

[57] Mulders PC, van Eijndhoven PF, Schene AH, Beckmann CF, Tendolkar I. Resting-state functional connectivity in major depressive disorder: A review. Neuroscience and Biobehavioral Reviews. 2015;56:330 – 344.

[58] Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA. Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. JAMA psychiatry. 2015;72(6):603–611.

[59] Krug A, Nieratschker V, Markov V, Krach S, Jansen A, Zerres K, et al. Effect of CACNA1C rs1006737 on neural correlates of verbal fluency in healthy individuals. Neuroimage. 2010;49(2):1831–1836.

[60] O'Reilly JX, Beckmann CF, Tomassini V, Ramnani N, Johansen-Berg H. Distinct and overlapping functional zones in the cerebellum defined by resting state functional connectivity. Cerebral Cortex. 2010;20(4):953–965.

[61] Moulton EA, Elman I, Pendse G, Schmahmann J, Becerra L, Borsook D. Aversion-related circuitry in the cerebellum: responses to noxious heat and unpleasant images. Journal of Neuroscience. 2011;31(10):3795–3804.

[62] Liu L, Zeng LL, Li Y, Ma Q, Li B, Shen H, et al. Altered cerebellar functional connectivity with intrinsic connectivity networks in adults with major depressive disorder. PLoS ONE. 2012;7(6):e39516.

[63] Zeng LL, Shen H, Liu L, Wang L, Li B, Fang P, et al. Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis. Brain. 2012;135(5):1498–1507.

[64] Wold H, et al. Soft modeling by latent variables: the nonlinear iterative partial least squares approach. Perspectives in Probability and Statistics, papers in honour of MS Bartlett. 1975;p. 520–540.

[65] McIntosh AR, Lobaugh NJ. Partial least squares analysis of neuroimaging data: applications and advances. NeuroImage. 2004;23:S250–S263.

[66] Chen K, Reiman EM, Huan Z, Caselli RJ, Bandy D, Ayutyanont N, et al. Linking functional and structural brain images with multivariate network analyses: a novel application of the partial least square method. NeuroImage. 2009;47(2):602–610.

[67] Yahata N, Morimoto J, Hashimoto R, Lisi G, Shibata K, Kawakubo Y, et al. A small number of abnormal brain connections predicts adult autism spectrum disorder. Nature Communications. 2016;7.

[68] Craddock RC, Holtzheimer PE, Hu XP, Mayberg HS. Disease state prediction from resting state functional connectivity. Magnetic Resonance in Medicine. 2009;62(6):1619–1628.

[69] Greicius MD, Flores BH, Menon V, Glover GH, Solvason HB, Kenna H, et al. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. Biological Psychiatry. 2007;62(5):429–437.

[70] Veer IM, Beckmann CF, Van Tol MJ, Ferrarini L, Milles J, Veltman DJ, et al. Whole brain resting-state analysis reveals decreased functional connectivity in major depression. Frontiers in Systems Neuroscience. 2010;4.

[71] Zhang J, Wang J, Wu Q, Kuang W, Huang X, He Y, et al. Disrupted brain connectivity networks in drug-naive, first-episode major depressive disorder. Biological Psychiatry. 2011;70(4):334–342.

[72] Zhang X, Yaseen ZS, Galynker II, Hirsch J, Winston A. Can depression be diagnosed by response to mother's face? A personalized attachment-based paradigm for diagnostic fMRI. PLoS ONE. 2011;6(12):e27253.

[73] Beck AT, Steer RA, Ball R, Ranieri WF. Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. Journal of Personality Assessment. 1996;67(3):588–597.

[74] Hasler G, Northoff G. Discovering imaging endophenotypes for major depression. Molecular Psychiatry. 2011;16(6):604–619.

[75] Snaith R, Hamilton M, Morley S, Humayan A, Hargreaves D, Trigwell P. A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. The British Journal of Psychiatry. 1995;167(1):99–103.

[76] Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. Journal of Personality and Social Psychology. 1988;54(6):1063.

[77] Katona CL, Watkin V. Depression in old age. Reviews in Clinical Gerontology. 1995;5(04):427–441.

[78] Yoshida K, Shimizu Y, Yoshimoto J, Toki S, Okada G, Takamura M, et al. Resting state functional connectivity explains individual scores of multiple clinical measures for major depression. In: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on. IEEE; 2015. p. 1078–1083.

[79] Lecrubier Y, Sheehan DV, Weiller E, Amorim P, Bonora I, Harnett Sheehan K, et al. The Mini International Neuropsychiatric Interview (MINI). A short diagnostic structured interview: reliability and validity according to the CIDI. European Psychiatry. 1997;12(5):224–231.

[80] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR). American Psychiatric Association; 2000.

[81] Ehring T, Tuschen-Caffier B, Schnülle J, Fischer S, Gross JJ. Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. Emotion. 2010;10(4):563.

[82] Van Dijk KR, Sabuncu MR, Buckner RL. The influence of head motion on intrinsic functional connectivity MRI. NeuroImage. 2012;59(1):431–438.

[83] Zeng LL, Wang D, Fox MD, Sabuncu M, Hu D, Ge M, et al. Neurobiological basis of head motion in brain imaging. Proceedings of the National Academy of Sciences. 2014;111(16):6058–6062.

[84] Perrot M, Rivière D, Mangin JF. Cortical sulci recognition and spatial normalization. Medical Image Analysis. 2011;15(4):529–550.

[85] Bennett K, Embrechts M. An optimization perspective on kernel partial least squares regression. NATO Science Series sub series III Computer and Systems Sciences. 2003;190:227–250.

[86] Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation. 1998;10(5):1299–1319.

[87] Xia M, Wang J, He Y. BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS ONE. 2013;8(7):e68910.

[88] Menzies L, Achard S, Chamberlain SR, Fineberg N, Chen CH, Del Campo N, et al. Neurocognitive endophenotypes of obsessive-compulsive disorder. Brain. 2007;130(12):3223–3236.

[89] Nestor PG, O'Donnell BF, McCarley RW, Niznikiewicz M, Barnard J, Shen ZJ, et al. A new statistical method for testing hypotheses of neuropsychological/MRI relationships in schizophrenia: partial least squares analysis. Schizophrenia Research. 2002;53(1):57–66.

[90] Zhu X, Wang X, Xiao J, Liao J, Zhong M, Wang W, et al. Evidence of a dissociation pattern in resting-state default mode network connectivity in first-episode, treatment-naive major depression patients. Biological Psychiatry. 2012;71(7):611–617.

[91] Zeng LL, Shen H, Liu L, Hu D. Unsupervised classification of major depression using functional connectivity MRI. Human Brain Mapping. 2014;35(4):1630–1641.

[92] Fitzgerald PB, Laird AR, Maller J, Daskalakis ZJ. A meta-analytic study of changes in brain activation in depression. Human Brain Mapping. 2008;29(6):683–695.

[93] Exner C, Lange C, Irle E. Impaired implicit learning and reduced pre-supplementary motor cortex size in early-onset major depression with melancholic features. Journal of Affective Disorders. 2009 Dec;119(1-3):156–62.

[94] Phan KL, Wager T, Taylor SF, Liberzon I. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. NeuroImage. 2002;16(2):331–348.

[95] Hamilton JP, Etkin A, Furman DJ, Lemus MG, Johnson RF, Gotlib IH. Functional neuroimaging of major depressive disorder: a meta-analysis and new integration of baseline activation and neural response data. American Journal of Psychiatry. 2012;.

[96] Yao Z, Wang L, Lu Q, Liu H, Teng G. Regional homogeneity in depression and its relationship with separate depressive symptom clusters: a resting-state fMRI study. Journal of Affective Disorders. 2009;115(3):430–438.

[97] Liu Z, Xu C, Xu Y, Wang Y, Zhao B, Lv Y, et al. Decreased regional homogeneity in insula and cerebellum: a resting-state fMRI study in patients with major depression and subjects at high risk for major depression. Psychiatry Research: Neuroimaging. 2010;182(3):211–215.

[98] Haldane M, Cunningham G, Androutsos C, Frangou S. Structural brain correlates of response inhibition in Bipolar Disorder I. Journal of Psychopharmacology. 2008;.

[99] Iwabuchi SJ, Krishnadas R, Li C, Auer DP, Radua J, Palaniyappan L. Localized connectivity in depression: a meta-analysis of resting state functional imaging studies. Neuroscience and Biobehavioral Reviews. 2015;51:77–86.

[100] Wilms I, Croux C. Sparse canonical correlation analysis from a predictive point of view. Biometrical Journal. 2015;57(5):834–851.

[101] Parkhomenko E, Tritchler D, Beyene J, et al. Sparse canonical correlation analysis with application to genomic data integration. Statistical Applications in Genetics and Molecular Biology. 2009;8(1):1–34.

[102] Akaho S. A Kernel Method For Canonical Correlation Analysis. In: In Proceedings of the International Meeting of the Psychometric Society (IMPS2001); 2001. .

[103] Bach FR, Jordan MI. Kernel independent component analysis. The Journal of Machine Learning Research. 2003;3:1–48.

[104] Yamanishi Y, Vert JP, Nakaya A, Kanehisa M. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. Bioinformatics. 2003;19(suppl 1):i323–i330.

[105] Balakrishnan S, Puniyani K, Lafferty JD. Sparse Additive Functional and Kernel CCA. In: Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012. The International Machine Learning Society; 2012. .

[106] Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009;71(5):1009–1030.

[107] Cristianini N, Shawe-taylor J, Elisseeff A, Kandola J. On Kernel-Target Alignment. In: Advances in Neural Information Processing Systems 14. Citeseer; 2001. .

[108] Cortes C, Mohri M, Rostamizadeh A. Two-Stage Learning Kernel Algorithms. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel; 2010. p. 239–246.

[109] Gretton A, Bousquet O, Smola AJ, Schölkopf B. Measuring Statistical Dependence with Hilbert-Schmidt Norms. In: Algorithmic Learning Theory, 16th International Conference, ALT 2005, Singapore, October 8-11, 2005, Proceedings. Springer-Verlag Berlin Heidelberg; 2005. p. 63–77.

[110] Yamada M, Jitkrittum W, Sigal L, Xing EP, Sugiyama M. High-dimensional feature selection by feature-wise kernelized lasso. Neural computation. 2014;26(1):185–207.

[111] Kloft M, Brefeld U, Sonnenburg S, Zien A. Lp-norm multiple kernel learning. The Journal of Machine Learning Research. 2011;12:953–997.

[112] Martin P, Guillou H, Lasserre F, Déjean S, Lan A, Pascussi J, et al. Novel aspects of PPARalpha-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. Hepatology (Baltimore, Md). 2007;45(3):767–777.

[113] Leurgans SE, Moyeed RA, Silverman BW. Canonical correlation analysis when the data are curves. Journal of the Royal Statistical Society Series B (Methodological). 1993;p. 725–740.

[114] Tamatam A BA Khanum F. Genetic biomarkers of depression. Indian Journal of Human Genetics. 2012;18(1):20–33.

[115] Lin D, Calhoun VD, Wang YP. Correspondence between fMRI and SNP data by group sparse canonical correlation analysis. Medical image analysis. 2014 August;18(6):891—902.

Interpretable machine learning approaches to high-dimensional date and their applications to biomedical engineering problems

Kosuke Yoshida

March 2018