# A Kernel Method to Extract Common Features Based on Mutual Information

Takamitsu Araki[1], Hideitsu Hino[2], and Shotaro Akaho[1]

[1] Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology,
Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 3058568, Japan
{tk-araki,s.akaho}@aist.go.jp
[2] Department of Computer Science, University of Tsukuba,
1-1-1 Tennodai, Tsukuba, Ibaraki 3058573, Japan
hinohide@cs.tsukuba.ac.jp

**Abstract.** Kernel canonical correlation analysis (CCA) aims to extract common features from a pair of multivariate data sets by maximizing a linear correlation between nonlinear mappings of the data. However, the kernel CCA tends to obtain the features that have only small information of original multivariates in spite of their high correlation, because it considers only statistics of the extracted features and the nonlinear mappings have high degree of freedom. We propose a kernel method for common feature extraction based on mutual information that maximizes a new objective function. The objective function is a linear combination of two kinds of mutual information, one between the extracted features and the other between the multivariate and its feature. A large value of the former mutual information provides strong dependency to the features, and the latter prevents loss of the feature's information related to the multivariate. We maximize the objective function by using the Parallel Tempering MCMC in order to overcome a local maximum problem. We show the effectiveness of the proposed method via numerical experiments.

**Keywords:** Kernel canonical correlation analysis, mutual information, Parallel Tempering.

## 1 Introduction

Recently, we can obtain data from multiple sources simultaneously such as electroencephalography (EEG) and near infra-red spectroscopy (NIRS) measurements of brain activity. Canonical correlation analysis is known as a linear method to extract common features in which source-specific noise is reduced from the original observations.

Kernel canonical correlation analysis (Kernel CCA; [1], [8], [3]) is a nonlinear extension of canonical correlation analysis with positive definite kernels. Given a pair of multivariates $\mathbf{x}$ and $\mathbf{y}$, the kernel CCA aims to extract the common features from them by finding nonlinear mappings $f(\mathbf{x})$ and $g(\mathbf{y})$ such that

C.K. Loo et al. (Eds.): ICONIP 2014, Part II, LNCS 8835, pp. 26–34, 2014.

the correlation coefficient is maximized. The kernel CCA has been applied for extracting nonlinear relations between the multivariates in various data, e.g., genomic data and functional magnetic resonance imaging (fMRI) brain images. The kernel CCA is one of kernel methods ([11]) that use nonlinear mappings of the multivariates instead of linear transformations used in traditional multivariate analysis.

The kernel CCA uses flexible nonlinear mappings to extract nonlinear relation between the multivariates. However, since large degrees of freedom of the mappings are devoted to their correlation maximization, the *features*, the mappings of the multivariates, lose a large amount of information related to the multivariates in many cases. The kernel CCA evaluates only the relation between the features and fails to detect the relationship between the multivariates.

The kernel CCA also extracts the features that have no interdependency, even though the correlation coefficient of them is high. The correlation coefficient cannot evaluate the dependency correctly when the data follow a non-Gaussian distribution. Thus, the correlation coefficient is not an appropriate criterion that evaluates the dependency of the features of the data that have nonlinear structure.

In this study, we propose a kernel method for common feature extraction that maximizes a new objective function based on mutual information. The objective function is a linear combination of mutual information between the features and that between the multivariate and its feature for each data set. The mutual information can evaluate essential dependency between the variables distributed with any distribution, so that the mutual information is a suitable criterion for the relation between the features and between the feature and the multivariate. A large value of the former mutual information provides the highly interdependent features and the latter prevents the features from losing a large amount of information about the original multivariate.

We apply the proposed method to an analysis of synthetic data, and show that our method can extract the true nonlinear structure of the data, which cannot be extracted by the conventional kernel CCA.

## 2 Kernel Canonical Correlation Analysis

Suppose there is a pair of multivariates $\mathbf{x} \in \mathbb{R}^{\mathbf{n_x}}$ and $\mathbf{y} \in \mathbb{R}^{\mathbf{n_y}}$, the kernel CCA aims to find a pair of nonlinear mappings $f(\mathbf{x})$ and $g(\mathbf{y})$ such that their correlation coefficient is maximized, where $f$ and $g$ belong to the respective reproducing kernel Hilbert spaces (RKHS) $\mathcal{H}_x$ and $\mathcal{H}_y$.

Since the maximization is ill-posed when the dimensionalities of the RKHS $\mathcal{H}_x$ and $\mathcal{H}_y$ are large, we introduce a quadratic regularization term $\eta(\|f\|^2_{\mathcal{H}_x}+\|g\|^2_{\mathcal{H}_y})$, where $\eta > 0$ is a regularization parameter. The kernel CCA maximizes the objective function that consists of the correlation coefficient between $f(\mathbf{x})$ and $g(\mathbf{y})$ and the regularization term.

The kernel CCA does not always extract essential common structure of a pair of multivariate data. For example, a common factor underlying the synthetic

data shown in Fig. 1 is not extracted by the kernel CCA. The synthetic data sets, $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{50}$, were generated from the two dimensional circle-shaped distributions derived from common random angle (details of this data are in Section 5). However, the values of the common angle are mixed in the features extracted by the kernel CCA with Gaussian kernel (Fig. 2(a)). This is because the kernel CCA finds redundant nonlinear mappings that often produce features having small information about the multivariate (Fig. 2(b.1)-(b.2)).

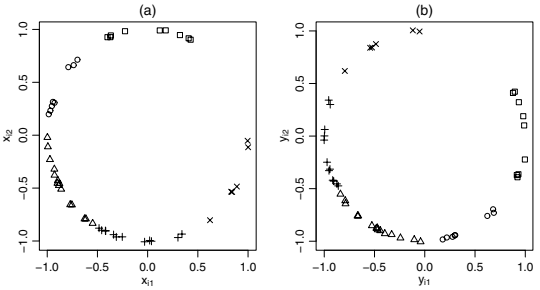


**Fig. 1.** Synthetic data sets, (a) $\mathbf{x}_i$, (b) $\mathbf{y}_i$. The data in each data set are grouped into five sections by intervals of the common angle value. The five groups are denoted by the five different marks.

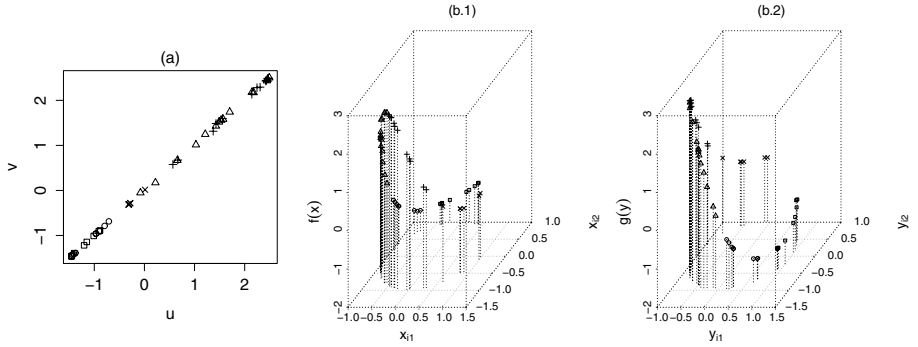


**Fig. 2.** (a) The features extracted by the kernel CCA, $u_i = \hat{f}(\mathbf{x}_i)$, $v_i = \hat{g}(\mathbf{y}_i)$, where $\hat{f}$ and $\hat{g}$ are the nonlinear mappings estimated by the kernel CCA. The marks correspond to those in Fig. 1. (b.1) $\hat{f}(\mathbf{x}_i)$, (b.2) $\hat{g}(\mathbf{y}_i)$.

This result shows that the kernel CCA dedicates the degree of freedom of nonlinear mappings to maximize the correlation of the features, and then the features' information on the multivariates is sacrificed. Therefore, the kernel CCA extracts only little information shared by the data set even if it obtains highly correlated features, because the features have little information about the multivariates.

The kernel CCA with another set of the regularization parameter and the Gaussian kernel's parameter extracts the non-informative features whose correlation coefficient is high (Fig. 3). This indicates that the correlation coefficient is not appropriate to evaluate the interdependence of the features.

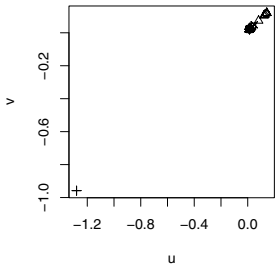


**Fig. 3.** The features extracted by the kernel CCA with other parameters

## 3 Mutual Information Based Objective Function

We propose a kernel method for common feature extraction based on mutual information that maximizes a new objective function. The objective function is constructed as follows.

We substitute the correlation coefficient between the features, $u = f(\mathbf{x})$ and $v = f(\mathbf{y})$, with the mutual information between them. The mutual information between two random variables, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, is $I(x, y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p(x, y) \log \frac{p(x,y)}{p(x)p(y)} dxdy$, where $p$ denotes a density function. The mutual information quantifies interdependence between random variables, and is interpreted as a measure of information shared by the random variables from an information theoretic point of view.

Since the maximization of the mutual information between the features also causes information loss of the features related to the multivariate, we add a constraint that the mutual information between the feature and the multivariate is enough large for each data set. That is, we consider an optimization problem

$$\max_{f \in \mathcal{H}_x, g \in \mathcal{H}_y} I(u, v), \quad \text{subject to} \quad I(u, \mathbf{x}) \geq s_x, I(v, \mathbf{y}) \geq s_y, \tag{1}$$

where $u = f(\mathbf{x})$, $v = g(\mathbf{y})$ and $s_x, s_y > 0$.

The problem in (1) is solved by maximizing the objective function

$$L_\lambda(f, g) = I(u, v) + \lambda_x I(u, \mathbf{x}) + \lambda_y I(v, \mathbf{y}),$$

where $\lambda = (\lambda_x, \lambda_y)$ are regularization parameters and $\lambda_x, \lambda_y > 0$. The regularization parameters $\lambda$ control the amount of the mutual information between the feature and the multivariate. The regularization term, $\lambda_x I(u, \mathbf{x}) + \lambda_y I(v, \mathbf{y})$, prevents loss of the information shared by the feature and the multivariate. The large

value of the mutual information between the features that have enough information on the multivariate provides the features capturing an essential nonlinear relation between the multivariates.

The mutual information does not depend on scale of random variables, that is, $I(cX, cY) = I(X, Y)$, for $c \neq 0$. Therefore, we maximize the objective function under the constraint $\|f\|^2_{\mathcal{H}_x} = \|g\|^2_{\mathcal{H}_y} = 1$.

In practice, we have to find the desired mappings from finite amount of data, $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, so that we estimate the mutual information by Mean Nearest Neighbor (MeanNN) method ([4],[6]). The mutual information between $\mathbf{x}$ and $\mathbf{y}$ estimated by the MeanNN method is

$$\hat{I}(\mathbf{x}, \mathbf{y}) = \hat{H}(\mathbf{x}) + \hat{H}(\mathbf{y}) - \hat{H}(\mathbf{x}, \mathbf{y}),$$

where $\hat{H}(\mathbf{x}) = \frac{\mathrm{n_x}}{N(N-1)} \sum_{i \neq j} \log \|\mathbf{x}_i - \mathbf{x}_j\| + \text{const.}$

Therefore, we maximize the objective function $\hat{L}_\lambda(f, g)$ under $\|f\|^2_{\mathcal{H}_x} = \|g\|^2_{\mathcal{H}_y} = 1$, where $\hat{L}_\lambda(f, g) = \hat{I}(u, v) + \lambda_x \hat{I}(u, \mathbf{x}) + \lambda_y \hat{I}(v, \mathbf{y})$. The solution of this problem is represented as $f(\cdot) = \sum_{i=1}^N \alpha_i k_x(\mathbf{x}_i, \cdot)$ and $g(\cdot) = \sum_{i=1}^N \beta_i k_y(\mathbf{y}_i, \cdot)$ by the Representer theorem ([10]), where $k_x$ and $k_y$ are kernel functions, and $\alpha_i, \beta_i \in \mathbb{R}$. We assume that $f$'s orthogonal part to the span of $k_x(\mathbf{x}_i, \cdot)$ is zero as well as $g$. The objective function of $\alpha = (\alpha_1, \ldots, \alpha_N), \beta = (\beta_1, \ldots, \beta_N)$, $\hat{L}_\lambda(\alpha, \beta)$, is obtained by applying the representation of the solution to $f, g$. The constraint is also represented by $\alpha^T K_x \alpha = \beta^T K_y \beta = 1$, where $(K_x)_{ij} = k_x(\mathbf{x}_i, \mathbf{x}_j)$ and $(K_y)_{ij} = k_y(\mathbf{y}_i, \mathbf{y}_j)$. However, since the constraint space is too complex to find the solution, we impose the simplified constraint $\|\alpha\|^2 = \|\beta\|^2 = 1$ in place of the constraint above for the sake of computational efficiency.

## 4 Algorithm

Since the objective function $\hat{L}_\lambda(\alpha, \beta)$ has many local maximum points, simple optimization methods such as a gradient method do not find a reasonable solution. To cope with this localization problem, we employ the Parallel Tempering ([5],[7]), which is one of the Markov chain Monte Carlo (MCMC) methods ([9]).

The MCMC methods efficiently generate samples from a target probability distribution by simulating a Markov chain that converges to the distribution. The Parallel Tempering introduces auxiliary distributions with a parameter called the temperature, generates multiple MCMC samples from target and the auxiliary distributions in parallel, and exchanges the positions of two samples.

The target distribution and the auxiliary distributions with inverse temperatures $t_l$ are

$$\pi_{t_l}(\alpha_l, \beta_l) \propto \exp\left(t_l \hat{L}_\lambda(\alpha_l, \beta_l)\right), \quad l = 1, \ldots, L,$$

where $t_1 > \cdots > t_L > 0$, $\pi_{t_1}(\alpha_1, \beta_1)$ is a target distribution, and the others are the auxiliary distributions.

The Parallel Tempering executes either of the parallel step and the exchange step at each iteration. The parallel step generates the $L$ samples according to $\pi_{t_l}(\alpha_l, \beta_l)$ for each by using a Metropolis algorithm. The Metropolis algorithm uses a proposal distribution that generates a sample candidate, which becomes the MCMC sample if accepted, and is rejected otherwise. We employ a von Mises-Fisher distribution, which is a probability distribution on the $(N-1)$-dimensional sphere in $\mathbb{R}^N$, as the proposal distribution. The proposal distribution enables us to directly generate samples of $\alpha$ on the constraint space, $\|\alpha\| = 1$, as well as $\beta$. The exchange step randomly chooses adjacent two samples and exchanges them with a Metropolis acceptance probability.

Since the target distribution is maximized if and only if the objective function is maximized, we find the solution from the samples generated by the Parallel Tempering.

## 5 Numerical Validation

Our method and the kernel CCA were applied to the circle data (analysed in Section 2) in order to show that our method can extract the essential nonlinear structure of a pair of data which the kernel CCA cannot extract.

The circle data, $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2$, $i = 1, \ldots, 50$, were generated as follows. First $\theta_i$ is generated from the uniform distribution on $(1, 2\pi)$, and then $\mathbf{x}_i$ and $\mathbf{y}_i$ were generated by,

$$\mathbf{x}_i = \begin{pmatrix} \cos(\theta_i) \\ \sin(\theta_i) \end{pmatrix} + \epsilon_i^x \quad \text{and} \quad \mathbf{y}_i = \begin{pmatrix} \sin(\theta_i) \\ \cos(\theta_i) \end{pmatrix} + \epsilon_i^y,$$

where $\epsilon_i^x, \epsilon_i^y$ are independent two dimensional Gaussian noises with a mean 0 and a standard deviation 0.01.

In this experiment, we used the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$ both for $\mathbf{x}$ and $\mathbf{y}$. The Parallel Tempering was run for $5 \times 10^5$ iterations, and the first inverse temperature was $t_1 = 1$, the other parameters were determined by one simulation of the adaptive Parallel Tempering ([2]).

Our method extracted the features as reconstructions of the common factor $\theta_i$ between $\mathbf{x}_i$ and $\mathbf{y}_i$ (Fig. 4(a)), while the kernel CCA could not extract (Fig. 5-6(a)). The nonlinear mappings estimated by our method provide enough information about the multivariates to the features (Fig. 4 (b.1)-(b.2)). On the other hand, those obtained by the kernel CCA provide only a part of information about the multivariates to the features (Fig. 5-6 (b.1)-(b.2)).

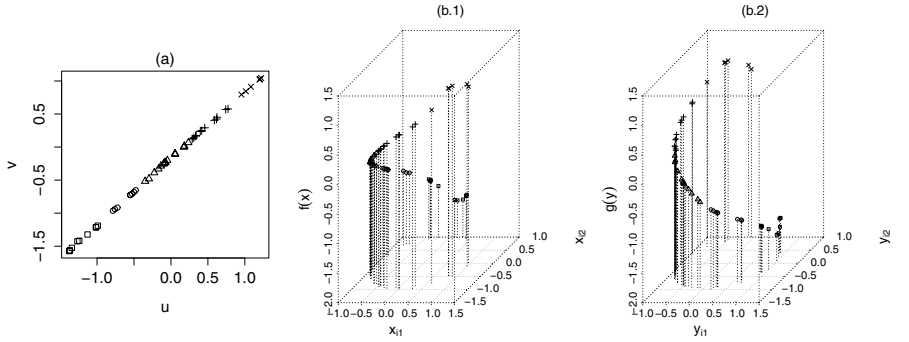


**Fig. 4.** (a) The features extracted by our method with parameters $\sigma = 0.45$ (the same value used in Section 2) and $\lambda_x = \lambda_y = 1$. (b.1)-(b.2) The nonlinear mappings estimated by our method, $\hat{f}(\mathbf{x}_i)$ (b.1) and $\hat{g}(\mathbf{y}_i)$ (b.2). (The marks are defined in Fig. 1 in Section 2.)

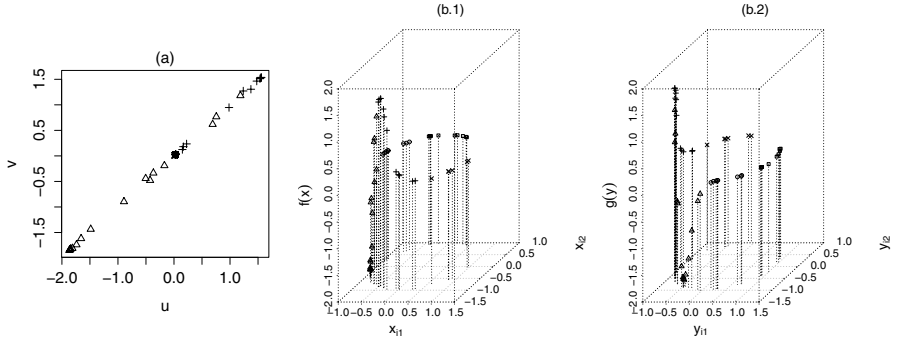


**Fig. 5.** (a) The features extracted by the kernel CCA with parameters $\sigma = 0.15$, $\eta = 3$. (b.1)-(b.2) The nonlinear mappings estimated by the kernel CCA.

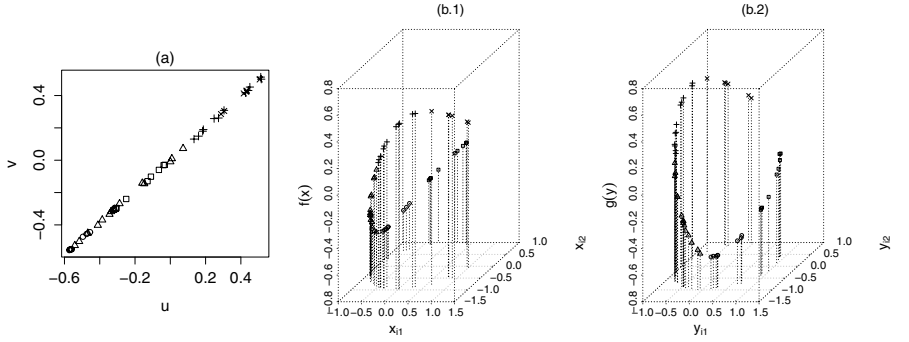


**Fig. 6.** (a) The features extracted by the kernel CCA with parameters $\sigma = 1.3$, $\eta = 7$. (b.1)-(b.2) The nonlinear mappings estimated by the kernel CCA.

## 6 Conclusions

We proposed the kernel method based on mutual information to extract common features from a pair of data sets. The proposed method maximizes a new objective function that consists of the mutual information between the features and those between the feature and the multivariate. The maximization of the objective function provides the features that represent nonlinear structure of a pair of the multivariate data set, because a large value of the mutual information between the feature and the multivariate provides the enough multivariate's information to the feature and the mutual information between the features is enlarged.

We also showed that our method can extract the common feature of the circle data which the kernel CCA cannot extract. This is because our method solves the essential problem of the kernel CCA that it tends to extract the features that have small information on the multivariates.

Our information-based method is difficult to apply to the extremely high dimensional data because estimation of entropy becomes unstable in such situation. However, our method is useful in adequate dimensional cases, in which it is difficult to extract nonlinear relations by the conventional kernel CCA.

The proposed method extracts only one component, so that we will extend the proposed method to the method that can extract multiple components. We will also develop faster algorithm that maximizes the proposed objective function, since the Parallel Tempering spends relatively large computational time.

**Acknowledgments.** Part of this work was supported by MEXT KAKENHI No.25120011 and JSPS KAKENHI No.25870811.

## References

1. Akaho, S.: A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of the Psychometric Society (IMPS 2001) (2001)
2. Araki, T., Ikeda, K.: Adaptive Markov chain Monte Carlo for auxiliary variable method and its application to Parallel Tempering. Neural Networks 43, 33–40 (2013)
3. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. The Journal of Machine Learning Research 3, 1–48 (2003)
4. Faivishevsky, L., Goldberger, J.: ICA based on a smooth estimation of the differential entropy. In: Advances in Neural Information Processing Systems, pp. 433–440 (2008)
5. Geyer, C.: Markov chain Monte Carlo maximum likelihood. In: Proc. 23rd Symp. Interface Comput. Sci. Statist., pp. 156–216 (1991)
6. Hino, H., Murata, N.: A conditional entropy minimization criterion for dimensionality reduction and multiple kernel learning. Neural Computation 22(11), 2887–2923 (2010)
7. Hukushima, K., Nemoto, K.: Exchange Monte Carlo method and application to spin glass simulations. Journal of the Physical Society of Japan 65(6), 1604–1608 (1996)

8. Melzer, T., Reiter, M.K., Bischof, H.: Nonlinear feature extraction using generalized canonical correlation analysis. In: Dorffner, G., Bischof, H., Hornik, K. (eds.) ICANN 2001. LNCS, vol. 2130, pp. 353–360. Springer, Heidelberg (2001)
9. Robert, C., Casella, G.: Monte Carlo Statistical Methods. Springer (2004)
10. Schölkopf, B., Herbrich, R., Smola, A.J.: A generalized representer theorem. In: Helmbold, D.P., Williamson, B. (eds.) COLT/EuroCOLT 2001. LNCS (LNAI), vol. 2111, pp. 416–426. Springer, Heidelberg (2001)
11. Schölkopf, B., Smola, A.: Learning with kernels. MIT Press, Cambridge (2002)