LETTER Communicated by Shotaro Akaho

Multiple Kernel Learning with Gaussianity Measures

Hideitsu Hino
hideitsu.hino@toki.waseda.jp
School of Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

Nima Reyhani
nima.reyhani@aalto.fi
School of Science, Aalto University, FI-00076 Aalto, Finland

Noboru Murata
noboru.murata@eb.waseda.ac.jp
School of Science and Engineering, Waseda University, Shinjuku, Tokyo 169-8555, Japan

Uncorrected Proof

Kernel methods are known to be effective for nonlinear multivariate analysis. One of the main issues in the practical use of kernel methods is the selection of kernel. There have been a lot of studies on kernel selection and kernel learning. Multiple kernel learning (MKL) is one of the promising kernel optimization approaches. Kernel methods are applied to various classifiers including Fisher's discriminant analysis (FDA). FDA gives the Bayes optimal classification axis if the data distribution of each class in the feature space is a gaussian with a shared covariance structure. Based on this fact, an MKL framework based on the notion of gaussianity is proposed. As a concrete implementation, an empirical characteristic function is adopted to measure gaussianity in the feature space associated with a convex combination of kernel functions, and two MKL algorithms are derived. From experimental results on some data sets, we show that the proposed kernel learning followed by FDA offers strong classification power.

1 Introduction

Kernel methods such as support vector machines (SVMs) have been shown to be successful for a wide range of data analysis problems (Cristianini & Shawe-Taylor, 2000). However, the difficulty in choosing a suitable kernel function and its parameters for a given data set is a serious drawback of these methods. A number of efforts have been made to solve kernel selection problems. For example, Cristianini, Shawe-Taylor, and Kandola (2001) proposed using the ideal kernel made from the class labels of the

 © Massachusetts Institute of Technology

training data and aligning the base kernel matrix to the ideal kernel matrix. Amari and Wu (1999) proposed magnifying the feature space around the classification surface of the SVM using conformal transformation. In various kernel optimization methods, one of the notable approaches is multiple kernel learning (MKL), in which the optimal convex combination of a set of given kernels is considered. In that framework, kernel combination coefficients are learned so that the separability of data in different classes is maximized in the feature space associated with combined kernels. Lanckriet, Cristianini. Bartlett, Ghaoui, and Jordan (2004) proposed a framework to combine multiple kernel functions with several different loss functions and to optimize weights for data and coefficients for kernels by using semidefinite programming (SDP). Starting from this pioneering work, studies have been devoted to improving classification accuracy, learning efficiency, and feature interpretability (Sonnenburg, Rätsch, Schäfer, & Schölkopf, 2006; Rakotomamonjy, Bach, Canu, & Grandvalet, 2008; Do, Kalousis, Woznica, & Hilario, 2009; Kim, Magnani, & Boyd, 2006; Yan, Kittler, Mikolajczyk, & Tahir, 2009; Suzuki & Tomioka, 2011).

In statistics and multivariate analysis literature, Fisher discriminant analysis (FDA; Fisher, 1936) is one of the most popular linear classification methods. It is also regarded as a supervised dimensionality-reduction method and is capable of a wide range of applications, such as preprocessing for other data analyses or visualizations. In FDA, if the data distribution of each class is a gaussian with the same covariance structure in the feature space, we can obtain a Bayes optimal classifier (Duda, Hart, & Stork, 2000). (See appendix A for a brief proof of this fact.)

In this study, we propose optimizing the kernel coefficients so that all the data distributions of individual classes in the feature space are as close to gaussian distributions sharing the same covariance structure as possible. In the feature space associated with the optimally combined kernel functions, we expect to obtain a Bayes optimal classifier by applying FDA. Figure 1 shows a conceptual diagram of desirable and undesirable data distributions in different feature spaces. Intuitively, the data in the original space are mapped to feature spaces by a family of maps $\{\phi_{\boldsymbol{\beta}}\}_{\boldsymbol{\beta}\in\Delta_S}$, which is defined by a given finite set of kernel functions and parameterized by the combination parameter $\boldsymbol{\beta}$ in the $S-1$ simplex Δ_S. The mathematical symbols in the figure will be defined in the next section. The kernel combination coefficients are optimized so that the distribution of the mapped data in each class is as close as possible to a gaussian. We first propose a general framework of MKL with a gaussianity measure. Algorithmically, MKL is a problem of finding the best kernel combination coefficients in a certain sense. We show a simple formulation of the framework using the empirical characteristic function for measuring gaussianity, and develop two algorithms.

Most of existing MKL methods are oriented to sparse representation or interpretable feature selection under some sparseness constraints for the kernel combination coefficients. Since a kernel function determines a feature

Uncorrected Proof

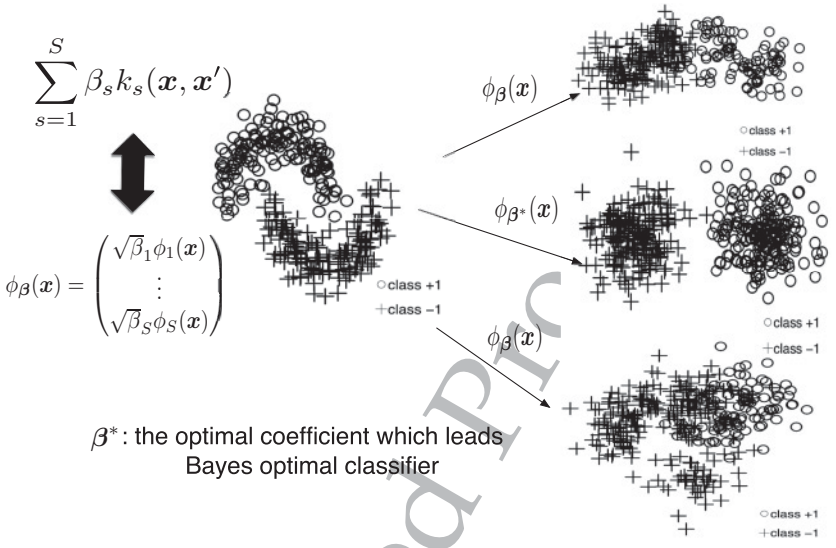


Figure 1: Conceptual diagram of desirable and undesirable data distributions obtained by feature mapping.

space, MKL can be regarded as a method to tailor the distribution of data by modifying the feature space. Though feature selection has always been an important issue in statistical data analysis, our focus will not be on selecting features but on combining features to achieve better discriminative power.

The rest of the letter is organized as follows. Section 2 briefly reviews the problem of MKL. Then a framework of MKL based on the gaussianity measure is proposed. In section 3, technical preliminaries to realize MKL methods based on the gaussianity measure are explained. The empirical characteristic function and its useful properties are utilized to define a gaussianity measure of the data distribution in the feature space. In section 4, two MKL algorithms with concrete algorithm descriptions are derived from the proposed MKL framework. Experimental results with artificial data and various kinds of benchmark data are given in section 5. The last section offers concluding remarks and notes future directions for research.

2 Gaussian Multiple Kernel Learning

In this section, multiple kernel learning for the binary classification problem is briefly reviewed. Then, the concept of multiple kernel learning with a gaussianity measure (Gaussian MKL; GMKL) is explained.

Uncorrected Pr

2.1 Multiple Kernel Learning Overview. Suppose we are given a set of observations $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1,\ldots,n}$ with class labels $\mathcal{L} = \{y_i\}_{i=1,\ldots,n}$, where $\boldsymbol{x}_i$ belongs to some input space $\mathcal{X}$ and y_i belongs to a class label set $\mathcal{Y} = \{+1, -1\}$. By $\mathcal{D}_{+1}$ and $\mathcal{D}_{-1}$, we denote subsets of $\mathcal{D}$ that belong to class $+1$ and class -1, respectively. When learning with multiple kernels, we are given S different feature mappings $\phi_1, \ldots, \phi_S$ from the input space $\mathcal{X}$ to corresponding feature spaces $\mathcal{H}_s$: $\phi_s : \mathcal{X} \to \mathcal{H}_s$, $s = 1, \ldots, S$. The dimensionality of these feature spaces is arbitrary, and they can be function spaces in general. We consider the case that each mapping ϕ_s gives a reproducing kernel k_s such that $k_s(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi_s(\boldsymbol{x}), \phi_s(\boldsymbol{x}') \rangle_{\mathcal{H}_s}$. We denote the inner product in the feature space $\mathcal{H}$ by $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ and omit the subscript $\mathcal{H}$ if there is no possible confusion. In the remainder of the letter, we use ϕ_s, k_s interchangeably for convenience and use $K_s = [k_s(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\ldots,n} = [\langle \phi_s(\boldsymbol{x}_i), \phi_s(\boldsymbol{x}_j) \rangle]_{i,j=1,\ldots,n}$ to represent a kernel matrix for $\mathcal{D}$. Then, denoting the $S-1$ simplex by Δ_S, we aim at finding an appropriate convex combination,

$$k_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{s=1}^{S} \beta_s k_s(\boldsymbol{x}, \boldsymbol{x}'), \quad \boldsymbol{\beta} \in \Delta_S, \tag{2.1}$$

which is written in the kernel matrix form as

$$K_{\boldsymbol{\beta}} = \sum_{s=1}^{S} \beta_s K_s, \quad \boldsymbol{\beta} \in \Delta_S, \tag{2.2}$$

and parameters $\boldsymbol{w}_s$, $s = 1, \ldots, S$ for a linear classification function,

$$f_{\boldsymbol{w},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{s=1}^{S} \sqrt{\beta_s} \langle \boldsymbol{w}_s, \phi_s(\boldsymbol{x}) \rangle, \tag{2.3}$$

where $\boldsymbol{w}_s \in \mathcal{H}_s$ and $\boldsymbol{w} = \{\boldsymbol{w}_s\}_{s=1,\ldots,S}$. We note that the feature space associated with the combined kernel functions becomes a tensor product of given feature spaces as $\mathcal{H}_{\boldsymbol{\beta}} = \mathcal{H}_1 \otimes \cdots \otimes \mathcal{H}_S$, where the corresponding feature mapping is written as $\phi_{\boldsymbol{\beta}}(\boldsymbol{x})$.

A common approach to MKL imposes an l_1-norm constraint on the combination coefficients, which usually results in sparse solutions lying on Δ_S. Under appropriate regularity conditions on the classification function, we can apply the representer theorem (Shawe-Taylor & Cristianini, 2004) to get the expression $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k_{\boldsymbol{\beta}}(\boldsymbol{x}, \boldsymbol{x}_i)$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$. Then, with a combined kernel matrix in equation 2.2, l_1-norm constrained MKL is generally formulated as the following optimization problem:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} L(\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \eta \Omega(\boldsymbol{\alpha}) \tag{2.4}$$

$$\text{subject to } \boldsymbol{\beta} \in \Delta_S, \tag{2.5}$$

Uncorrected Proof

where $L(\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is an arbitrary convex loss function, $\Omega(\cdot)$ is a strictly monotonically increasing function to regularize the complexity of the classifier, and $\eta > 0$ is a regularization parameter. The kernel combination coefficient $\boldsymbol{\beta}$ is constrained to be in a simplex Δ_S, and this is equivalent to l_1-norm constrained with a positivity constraint. A general result on the convexity of kernel learning of the form in equation 2.4 has been established in Micchelli and Pontil (2005). For $L(\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}_i))^2$ and $L(\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{x}_i))$, the classifiers behind the optimization problem are known to be a kernel regularized least square and an SVM, respectively. On the other hand, when $L(\mathcal{D}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\boldsymbol{\alpha}^\top V_{w,\boldsymbol{\beta}}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top V_{b,\boldsymbol{\beta}}\boldsymbol{\alpha}}$, where $V_{w,\boldsymbol{\beta}}$ and $V_{b,\boldsymbol{\beta}}$ are within- and between-class covariance matrices in the feature space, the classifier behind the optimization problem is a kernel Fisher discriminant analysis (Mika, Rätsch, Weston, Schölkopf & Müllers, 1999).

We note that conventional MKL methods include optimization with respect to both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Since our proposed kernel learning method tailors distributions of the data in the feature space by optimizing $\boldsymbol{\beta}$, we will concentrate on optimization with respect to $\boldsymbol{\beta}$. This is one of the features of the proposed MKL framework; that is, there is no need to perform simultaneous or alternating optimization for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

2.2 Learning Kernel Combination Based on Gaussianity. In FDA, it is desirable that each class-conditional distribution $p(\boldsymbol{x}|y), \boldsymbol{x} \in \mathcal{X}, y \in \mathcal{Y}$ is a gaussian with the same covariance structure in the feature space. That is, we would like to let the distributions of projected data $\phi_{\boldsymbol{\beta}}(\mathcal{D}_l), l \in \mathcal{Y} = \{+1, -1\}$ be as close to gaussian distributions as possible and to let the two covariance matrices of these data sets be as close as possible.

We formulate the MKL problem, which we call the GMKL (gaussian MKL) framework, as follows:

$$\min_{\boldsymbol{\beta}\in\Delta_S} \quad J(\boldsymbol{\beta}), \tag{2.6}$$

$$J(\boldsymbol{\beta}) = M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1})) + M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) + \eta M_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}), \phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})),$$

where M_G is a distance between a gaussian distribution and an empirical distribution of the observed data. M_V is a distance between empirical covariance matrices of two sets of data in the feature space, and $\eta > 0$ is a balancing parameter.

In this framework, the problem is which distance measure to select for M_G and M_V. There are some requirements for the distance measure. For example, it must be easily estimated from the given data, and it must be bounded below. It is also preferable that the measure be differentiable with respect to $\boldsymbol{\beta}$. For M_G, there are numerous quantities for measuring the gaussianity of the given data. In this letter, we choose a quantity based on an empirical characteristic function to measure gaussianity because it

Uncorrected Proof

is estimated using only kernel matrices, it is bounded below, and it is differentiable. We discuss other possibilities for M_G later. For M_V, there are also many candidates. Here, we simply take natural distance measures based on the empirical characteristic function.

3 Technical Preliminary

This section presents technical preliminaries to construct the GMKL framework.

3.1 Empirical Characteristic Function. The characteristic function of a d-dimensional random variable X is defined as

$$c(\boldsymbol{t}) = E_{p_X}[e^{i\boldsymbol{t}^\top X}], \tag{3.1}$$

where p_X is a probability density function of X and $\boldsymbol{t}$ is a d-dimensional vector in $\mathbb{R}^d$. In this letter, we consider only the case where the distribution has a density function. We note that $c(\boldsymbol{t})$ is nothing but a Fourier transformation of the probability density function of X. For independent and identically distributed (i.i.d.) samples $\mathcal{D} = \{\boldsymbol{x}_j\}_{j=1,\ldots,n}$ from p_X, we construct an empirical distribution $p_{\mathcal{D}}$, and the empirical characteristic function is then defined by

$$c_{\mathcal{D}}(\boldsymbol{t}) = E_{p_{\mathcal{D}}}[e^{i\boldsymbol{t}^\top X}] = \frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} e^{i\boldsymbol{t}^\top \boldsymbol{x}_j}. \tag{3.2}$$

The empirical characteristic function has several preferable properties. Specifically, under some general restrictions, $c_{\mathcal{D}}(\boldsymbol{t})$ converges uniformly almost surely to the population characteristic function $c(\boldsymbol{t})$ (Feuerverger & Mureika, 1977). For arbitrary dimension d, the convergence of the empirical characteristic function to the characteristic function is proved using the law of large numbers and the Glivenko-Cantelli theorem (Vapnik, 1998). In section 3.2, we define the empirical characteristic function in the feature space, which might be a functional space. Even in such cases, the characteristic functions in the feature space also completely characterize the probability distributions in that space. Furthermore, under some regularity conditions, the same convergence property as in the finite-dimensional case holds for empirical characteristic functions in infinite-dimensional Hilbert spaces. (See Ledoux & Talagrand, 1991, for details.) Because of its computational ease and theoretical soundness, the empirical characteristic function is applied to a goodness-of-fit test (Koutrouvelis, 1980), a test for the shape of distributions (Murota & Takeuchi, 1981), and ICA (Murata, 2001; Eriksson & Koivunen, 2003), for example.

Uncorrected Proof

Using the following theorem, when we consider the gaussian distribution, we have to calculate only the modulo (i.e., absolute value) of the characteristic function, ignoring its argument:

Theorem 1. *A distribution with a characteristic function $c(\boldsymbol{t})$ is a gaussian if and only if $-\log|c(\boldsymbol{t})|^2$ is of the form*

$$-\log|c(\boldsymbol{t})|^2 = \boldsymbol{t}^\top \Sigma \boldsymbol{t}, \tag{3.3}$$

where Σ is a positive-definite matrix.

The proof of the theorem is given in appendix B.

The characteristic function of a multivariate gaussian distribution is defined by

$$c^*(\boldsymbol{t}) = \exp(i\boldsymbol{t}^\top \boldsymbol{\mu}) \exp\left(-\frac{1}{2}\boldsymbol{t}^\top \Sigma \boldsymbol{t}\right), \tag{3.4}$$

where $\boldsymbol{\mu}$ is a mean vector and Σ is a covariance matrix. Then $|c^*(\boldsymbol{t})|^2 = \exp(-\boldsymbol{t}^\top \Sigma \boldsymbol{t})$. Under the assumption that the data distribution is a gaussian, the empirical counterpart of $c^*(\boldsymbol{t})$ will be also denoted $c^*_{\mathcal{D}}(\boldsymbol{t})$, and it is obtained by plugging the empirical estimate of the mean and the covariance matrix into equation 3.4.

3.2 Kernelized Empirical Characteristic Function. Since functions 3.2 and 3.4 are written with only the inner products, they can be computed with a kernel function $k_{\boldsymbol{\beta}}$ when we consider the mapped data by a feature map:

$$\begin{aligned} \phi_{\boldsymbol{\beta}} &: \mathcal{X} \to \mathcal{H}_{\boldsymbol{\beta}} \\ &\boldsymbol{x} \mapsto \phi_{\boldsymbol{\beta}}(\boldsymbol{x}). \end{aligned}$$

In our settings, we have only to consider distributions $p_{\phi_{\boldsymbol{\beta}}(X)}$ of $\phi_{\boldsymbol{\beta}}(X)$ in $\mathcal{H}_{\boldsymbol{\beta}}$, which is induced from the action of the map $\phi_{\boldsymbol{\beta}}$ on the random variable X. Let $\mathcal{H}'_{\boldsymbol{\beta}}$ be the dual space of the feature space $\mathcal{H}_{\boldsymbol{\beta}}$. For a random variable $\phi_{\boldsymbol{\beta}}(X) \in \mathcal{H}_{\boldsymbol{\beta}}$ and arbitrary element $\tau : \mathcal{H}_{\boldsymbol{\beta}} \to \mathbb{R}$ in the dual space $\mathcal{H}'_{\boldsymbol{\beta}}$, the characteristic function in the feature space is defined as

$$c(\tau) = E_{p_{\phi_{\boldsymbol{\beta}}(X)}}\left[\exp(i\tau(\phi_{\boldsymbol{\beta}}(X)))\right]. \tag{3.5}$$

We are considering the case that the feature space is associated with an inner product; hence, it is a Hilbert space. By the Riesz lemma (Reed & Simon, 1981), for any $\tau \in \mathcal{H}'_{\boldsymbol{\beta}}$, there exists some $\boldsymbol{\tau} \in \mathcal{H}_{\boldsymbol{\beta}}$ so that for any

Uncorrected Proof

$\phi_{\boldsymbol{\beta}}(X) \in \mathcal{H}_{\boldsymbol{\beta}}$, $\tau(\phi_{\boldsymbol{\beta}}(X)) = \langle \boldsymbol{\tau}, \phi_{\boldsymbol{\beta}}(X) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}}$. Then the characteristic function of a random variable $\phi_{\boldsymbol{\beta}}(X) \in \mathcal{H}_{\boldsymbol{\beta}}$ is written as

$$c(\boldsymbol{\tau}) = E_{p_{\phi_{\boldsymbol{\beta}}(X)}}[\exp(i\langle \boldsymbol{\tau}, \phi_{\boldsymbol{\beta}}(X) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}})], \tag{3.6}$$

for $\boldsymbol{\tau} \in \mathcal{H}_{\boldsymbol{\beta}}$, and the empirical characteristic function becomes

$$\begin{aligned} c_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}(\boldsymbol{\tau}) &= E_{p_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}}[\exp(i\langle \boldsymbol{\tau}, \phi_{\boldsymbol{\beta}}(X) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}})] \\ &= \frac{1}{n} \sum_{\phi_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \in \phi_{\boldsymbol{\beta}}(\mathcal{D})} \exp(i\langle \boldsymbol{\tau}, \phi_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}}) \qquad (3.7) \\ &= \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \exp(i\langle \boldsymbol{\tau}, \phi_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}}), \qquad (3.7) \end{aligned}$$

where the empirical distribution $p_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}$ in the feature space is constructed from $\phi_{\boldsymbol{\beta}}(\mathcal{D}) = \{\phi_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\}_{\boldsymbol{x}_i \in \mathcal{D}}$. When the value of two characteristic functions coincides in every point in its domain, the corresponding two random variables are under the same distribution. That is, the empirical characteristic function $c_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}(\boldsymbol{\tau})$ should be evaluated at all points $\boldsymbol{\tau} \in \mathcal{H}_{\boldsymbol{\beta}}$ in principle. However, to express the empirical characteristic function in the feature space using the value of kernel function, we consider only the point $\boldsymbol{\tau}$ such that there exists $\boldsymbol{t} \in \mathcal{X}$ satisfying $\phi_{\boldsymbol{\beta}}(\boldsymbol{t}) = \boldsymbol{\tau}$. Some arguments to validate this restriction are presented later. By the definition of the kernel function property, we then have $\langle \phi_{\boldsymbol{\beta}}(\boldsymbol{t}), \phi_{\boldsymbol{\beta}}(\boldsymbol{x}) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}} = k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x})$, and using only the values of the kernel functions, the empirical characteristic function in the feature space is written as

$$c_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}(\phi_{\boldsymbol{\beta}}(\boldsymbol{t})) = \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \exp(i\langle \phi_{\boldsymbol{\beta}}(\boldsymbol{t}), \phi_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \rangle_{\mathcal{H}_{\boldsymbol{\beta}}}) = \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \exp(ik_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j)). \tag{3.8}$$

To simplify notations and to show dependence on $\boldsymbol{\beta}$ explicitly, we write $c_{\phi_{\boldsymbol{\beta}}(\mathcal{D})}(\phi_{\boldsymbol{\beta}}(\boldsymbol{t}))$ as $c_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta})$ henceforth. Now the squared modulo of the empirical characteristic function is

$$\begin{aligned} |c_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta})|^2 &= \left\{ \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \cos\left(\sum_{s=1}^{S} \beta_s k_s(\boldsymbol{t}, \boldsymbol{x}_j) \right) \right\}^2 \\ &\quad + \left\{ \frac{1}{n} \sum_{\boldsymbol{x}_j \in \mathcal{D}} \sin\left(\sum_{s=1}^{S} \beta_s k_s(\boldsymbol{t}, \boldsymbol{x}_j) \right) \right\}^2. \qquad (3.9) \end{aligned}$$

Uncorrected Proof

The covariance matrix in the feature space is formally defined as

$$\hat{\Sigma}_{\mathcal{D}} = \frac{1}{n}\sum_{\boldsymbol{x}_i \in \mathcal{D}} \left(\boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - \frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \right) \left(\boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{x}_i) - \frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{x}_j) \right)^{\top}. \tag{3.10}$$

A quadratic form $S^*_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta})$ that characterizes the empirical characteristic function of the gaussian distribution in the feature space is obtained by replacing the inner product in the original data space with that in the feature space,

$$\begin{aligned} S^*_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta}) &= \langle \boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{t}),\, \hat{\Sigma}_{\mathcal{D}} \boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{t}) \rangle = \frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \left(k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) \right)^2 \\ &= \boldsymbol{\beta}^{\top} \left(\frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)^{\top} \right) \boldsymbol{\beta} = \boldsymbol{\beta}^{\top} V_{\mathcal{D}} \boldsymbol{\beta}, \end{aligned} \tag{3.11}$$

where $\hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) = \sum_{s=1}^{S} \beta_s \hat{\mu}_s(\boldsymbol{t})$ is the estimated mean evaluated at $\boldsymbol{\phi}_{\boldsymbol{\beta}}(\boldsymbol{t})$, and each component $\hat{\mu}_s(\boldsymbol{t})$ is defined by $\hat{\mu}_s(\boldsymbol{t}) = (1/n)\sum_{\boldsymbol{x}_i \in \mathcal{D}} k_s(\boldsymbol{t}, \boldsymbol{x}_i)$. Vectors $\boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)$ and a matrix $V_{\mathcal{D}}$ are defined by

$$\boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) = \begin{pmatrix} k_1(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_1(\boldsymbol{t}) \\ \vdots \\ k_S(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_S(\boldsymbol{t}) \end{pmatrix} \in \mathbb{R}^S, \quad V_{\mathcal{D}} = \frac{1}{n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j) \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{x}_j)^{\top}. \tag{3.12}$$

The most notable advantage of using the empirical characteristic function is that it enables us to estimate gaussianity using the value of the kernel function only. Furthermore, though the probability density function of the gaussian distribution involves the inverse of the covariance matrix, we can measure gaussianity without estimating the inverse of the covariance matrix when we use the empirical characteristic function. The empirical characteristic function of the gaussian distribution in the feature space is

$$c^*_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta}) = \exp(i\hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t})) \exp\left(-\frac{1}{2n}\sum_{\boldsymbol{x}_j \in \mathcal{D}} \left(k_{\boldsymbol{\beta}}(\boldsymbol{t}, \boldsymbol{x}_j) - \hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t}) \right)^2 \right), \tag{3.13}$$

Uncorrected Proof

and we obtain the squared modulo of the empirical characteristic function under the gaussian assumption:

$$\begin{aligned}|c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 &= \exp\left\{-\frac{1}{n}\sum_{\boldsymbol{x}_j\in\mathcal{D}}\left(k_{\boldsymbol{\beta}}(\boldsymbol{t},\boldsymbol{x}_j)-\hat{\mu}_{\boldsymbol{\beta}}(\boldsymbol{t})\right)^2\right\}\\ &= \exp\left\{-\boldsymbol{\beta}^\top\left(\frac{1}{n}\sum_{\boldsymbol{x}_j\in\mathcal{D}}\boldsymbol{v}(\boldsymbol{t},\boldsymbol{x}_j)\boldsymbol{v}(\boldsymbol{t},\boldsymbol{x}_j)^\top\right)\boldsymbol{\beta}\right\}\\ &= \exp\left\{-\boldsymbol{\beta}^\top V_{\mathcal{D}}\boldsymbol{\beta}\right\} = \exp\{-S^*_{\mathcal{D}}(\boldsymbol{t},\boldsymbol{\beta})\}. \end{aligned} \tag{3.14}$$

From the above equation, $-\log|c^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2 = S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = \boldsymbol{\beta}^\top V_{\mathcal{D}}\boldsymbol{\beta}$ holds, and the squared modulo of the characteristic function of gaussian distribution is transformed to a simple quadratic form by logarithmic transformation. Thus, we also transform the squared modulo of the empirical characteristic function $|c_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2$ defined in equation 3.9 by $-\log$ and denote it as

$$S_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta}) = -\log|c_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})|^2. \tag{3.15}$$

By equation 3.11, $S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})$ is regarded as a projection of covariance matrix by a vector $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$; hence, we will call $S^*_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})$ and $S_{\mathcal{D}}(\boldsymbol{t};\boldsymbol{\beta})$ projected variances henceforth.

The choice of the point $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ to evaluate the empirical characteristic function is a difficult problem. Theoretically, to claim that two distributions are the same, the corresponding two characteristic functions must be the same for all points in the feature space. A test statistic for ICA is proposed by a weighted integral of the characteristic function (Eriksson & Koivunen, 2003). Another work claims that empirically, evaluation at only one point is enough for a test of the shape of distributions (Murota & Takeuchi, 1981). It is natural to use every training data point $\{\phi_{\boldsymbol{\beta}}(\boldsymbol{x}_i)\}$ to evaluate the characteristic function. However, using all of the training data might be computationally inefficient. In this study, to save computational cost, we sample only one point from training data, $\boldsymbol{t}\in\mathcal{D}$, and take it as $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$. In appendix C, we show an empirical evaluation of the effect of $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ on both classification accuracy and resultant kernel combination parameter.

By selecting an arbitrary point $\boldsymbol{t}$ from the given data, we define a measure of Gaussianity in our framework of MKL in equation 2.6 by

$$M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) = \frac{|\mathcal{D}_l|}{n}\left(S^*_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta}) - S_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})\right)^2, \tag{3.16}$$

Uncorrected Proof

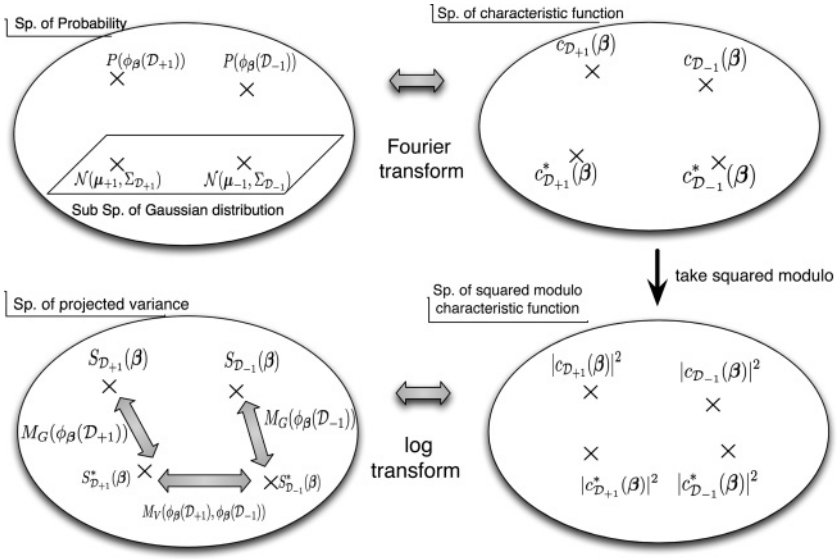


Figure 2: A conceptual diagram of data transformations and distance measures of gaussianity and covariance structure.

where the projected variances $S^*_{\mathcal{D}_l}$ and $S_{\mathcal{D}_l}$ are estimated using only data in $\mathcal{D}_l, l \in \mathcal{Y}$ and $|\mathcal{D}_l|$ is the cardinality of the set $\mathcal{D}_l$. It is easy to see that $M_G(\phi_\beta(\mathcal{D}_l))$ is bounded below and smooth with respect to $\boldsymbol{\beta}$. Figure 2 provides a conceptual diagram of the relationship between the space of probability density functions, the space of characteristic functions, the space of the squared modulo of the characteristic functions, and the space of projected variances.

3.3 Sharing Covariance Matrices. We next consider the gap between covariance matrices of two classes. As we stated in section 3.2, two data sets $\mathcal{D}_{+1}$ and $\mathcal{D}_{-1}$ are projected by $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$, and the corresponding covariance matrices are projected as $S^*_{\mathcal{D}_{+1}}(\boldsymbol{t};\boldsymbol{\beta})$ and $S^*_{\mathcal{D}_{-1}}(\boldsymbol{t};\boldsymbol{\beta})$, respectively. If variances $S^*_{\mathcal{D}_{+1}}(\boldsymbol{t};\boldsymbol{\beta})$ and $S^*_{\mathcal{D}_{-1}}(\boldsymbol{t};\boldsymbol{\beta})$ on the axis coincide for arbitrary projection $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$, two covariance matrices are the same. Thus, we can use $S^*_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})$ to define distance measures of two covariance matrices. By selecting arbitrary point $\boldsymbol{t}$ from the given data, we define a distance measure between empirical covariance matrices based on their projected variances:

$$\begin{aligned} M_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}),\phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) &= (S^*_{+1}(\boldsymbol{t},\boldsymbol{\beta}) - S^*_{-1}(\boldsymbol{t},\boldsymbol{\beta}))^2 \\ &= \left\{\boldsymbol{\beta}^\top \left(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}}\right)\boldsymbol{\beta}\right\}^2. \end{aligned} \tag{3.17}$$

Uncorrected

Algorithm 1: GGMKL: Gradient-Based Multiple Kernel Learning with Gaussianity Measure.

input: A labeled data set $\{\mathcal{D}_{+1}, \mathcal{D}_{-1}\}$. Candidate kernel functions $k_s, s = 1, \ldots, S$. A tuning parameter $\eta \geq 0$.

initialize: Set $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_S) = (1/S, \ldots, 1/S)$. For $l \in \mathcal{Y}$, sample $\boldsymbol{t}_l$ from $\mathcal{D}_l$, and re-define $\mathcal{D}_l$ by $\mathcal{D}_l \backslash \{\boldsymbol{t}_l\}$.

while: stopping criterion not met,**do**

compute gradient of $J(\boldsymbol{\beta}_t)$,

update $\boldsymbol{\beta}_t$ by $\boldsymbol{\beta}_t \leftarrow \boldsymbol{\beta}_t + \mu \nabla J(\boldsymbol{\beta}_t)$, where μ is found by line search.

end while

As another variant of the distance measure of covariance, we define

$$\tilde{M}_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}), \phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) = \boldsymbol{\beta}^\top \big(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}}\big)\big(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}}\big)^\top \boldsymbol{\beta}, \tag{3.18}$$

which is useful for deriving a quadratic approximation of the objective function. We note that with the positivity constraint for $\boldsymbol{\beta}$, the minimizers of M_V and $\tilde{M}_V$ are different in general.

4 Algorithm Description

In this section, we first describe two GMKL algorithms to learn the kernel combination coefficients. Then we introduce the empirical kernel feature map to perform FDA and evaluate gaussianity in the feature space associated with the learned kernel function.

4.1 Gradient-Based GMKL Algorithm. We present a gradient-based algorithm to perform GMKL (GGMKL) in algorithm 1, which solves the optimization problem 2.6 with

$$\begin{aligned} M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) &= \frac{|\mathcal{D}_l|}{n}(S^*_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta}) - S_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta}))^2, \\ M_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}), \phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) &= (S^*_{\mathcal{D}_{+1}}(\boldsymbol{t};\boldsymbol{\beta}) - S^*_{\mathcal{D}_{-1}}(\boldsymbol{t};\boldsymbol{\beta}))^2 \\ &= \{\boldsymbol{\beta}^\top (V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})\boldsymbol{\beta}\}^2. \end{aligned}$$

In this letter, we use ∇ and ∇^2 for $\frac{\partial}{\partial \boldsymbol{\beta}}$ and $\frac{\partial^2}{\partial \boldsymbol{\beta}^2}$, respectively. The gradient vector of the objective function $J(\boldsymbol{\beta})$ is given in appendix C.

Uncorrected Proof

Algorithm 2: SQGMKL: SQP-Based Multiple Kernel Learning with Gaussianity Measure.

input: A labeled data set $\{\mathcal{D}_{+1}, \mathcal{D}_{-1}\}$. Candidate kernel functions $k_s, s = 1, \ldots, S$. A tuning parameter $\eta \geq 0$.

initialize: Set $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_S) = (1/S, \ldots, 1/S)$. For $l \in \mathcal{Y}$, sample $\boldsymbol{t}_l$ from $\mathcal{D}_l$, and redefine $\mathcal{D}_l$ by $\mathcal{D}_l \backslash \{\boldsymbol{t}_l\}$.

while: stopping criterion not met, **do**

update $\boldsymbol{\beta}_t$ by the solution of quadratic problem 26 defined at $\boldsymbol{\beta}_t$.

end while

The objective function $J(\boldsymbol{\beta})$ is not a convex function with respect to $\boldsymbol{\beta}$, and the gradient method does not offer the global optimal solution. A common method to alleviate this problem is to execute the optimization from several different initial points and adopt the result with minimum objective function value. However, in our experiments, we use only one initial point $\boldsymbol{\beta}_0 = (1/S, \ldots, 1/S)$ for simplicity. We stopped the algorithm when $|J(\boldsymbol{\beta}_t) - J(\boldsymbol{\beta}_{t+1})| < 10^{-6}$ holds. For real-world data sets, the number of iterations required to match the stopping criterion was around 50 to 100, and the GGMKL algorithm is found to be relatively slow compared to other algorithms. Although it performs well in light of classification accuracy, we recommend using the faster GMKL algorithm derived in section 4.2 for large data sets. Enhancing the convergence property by more elaborate methods such as conjugate gradient method or Newton method remains future work.

4.2 Sequential Quadratic Programming for GMKL. We next present a variant of a sequential quadratic programming (SQP)–based algorithm to perform GMKL (SQGMKL) in algorithm 2, which solves a modified version of optimization problem 2.6, where M_G is defined by equation 3.16 and M_V is replaced with $\tilde{M}_V + \lambda\|\boldsymbol{\beta}\|^2,\ \lambda > 0$. Let the objective function be

$$\tilde{J}(\boldsymbol{\beta}) = \sum_{l \in \mathcal{Y}} M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) + \eta(\tilde{M}_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}), \phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) + \lambda\|\boldsymbol{\beta}\|^2). \tag{4.1}$$

Then at each step of the SQGMKL algorithm, we solve the following quadratic programming iteratively until convergence:

$$\min_{\boldsymbol{\beta} \in \Delta_S} \tilde{J}_t(\boldsymbol{\beta}) = \boldsymbol{c}_t^\top \boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\beta}^\top H_t \boldsymbol{\beta}, \tag{4.2}$$

Uncorrected Proof

where

$$\boldsymbol{c}_t = \nabla \tilde{J}(\boldsymbol{\beta})|_{\boldsymbol{\beta}_t} \in \mathbb{R}^S, \quad H_t = \nabla^2 \tilde{J}(\boldsymbol{\beta})|_{\boldsymbol{\beta}_t} \in \mathbb{R}^{S \times S}. \tag{4.3}$$

The explicit formulas for $\boldsymbol{c}_t$ and H_t are given in appendix D.

Many researchers have studied kernel optimization methods including MKLs (Bach, Lanckriet, & Jordan, 2004; Bennett, Momma, & Embrechts, 2002; Bi, Zhang, & Bennett, 2004; Bousquet & Herrmann, 2002; Cristianini et al., 2001; Crammer, Keshet, & Singer, 2002). In these studies, one of the main emphases is on formulating kernel learning as a tractable convex optimization problem. Our gaussianity measure M_G is not convex in general, but with a straightforward calculation, it is easily verified that its Hessian matrix is bounded. Since the difference measure of covariance matrices $\tilde{M}_V$ is a quadratic form with a positive semidefinite symmetric matrix $(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top$, with a small, positive constant $\lambda > 0$, the quadratic term $\boldsymbol{\beta}^\top((V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top + \lambda I_S)\boldsymbol{\beta}$ becomes convex. Using the concave-convex procedure (CCCP) theorem (Yuille & Rangarajan, 2003, theorem 1), optimization problem 4.1 becomes a convex minimization problem with sufficiently large $\eta > 0$. The sequential quadratic programming method for convex programming is guaranteed to converge to the global optimum. We note that since the matrix $(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top$ is positive semidefinite, the small, constant λ can be arbitrarily chosen. In the experiments in this letter, we set $\lambda = 0.01$. We note that the quadratic term $\|\boldsymbol{\beta}\|^2$ is minimized if $\boldsymbol{\beta} = (1/S, \ldots, 1/S)$. This additional term can be seen as a regularization term to avoid too sparse a solution, which might lead to poor generalization capability (Kloft, Brefeld, Laskov, & Sonnenburg, 2008; Yan, Kittler, Mikolajczyk, & Tahir, 2009).

4.3 Empirical Kernel Feature Map. In kernel methods, linear classifications such as PCA, SVM, and FDA are applied in the feature space. In general, the dimensionality of the feature space is very large, or it can be a function space. We use the notion of an empirical kernel feature map (Schölkopf et al., 1999; Xiong, Swamy, & Ahmad, 2005) to alleviate the difficulty of handling the distribution of the data in feature spaces and to obtain feature vectors explicitly.

Let $K \in \mathbb{R}^{n \times n}$ be a kernel matrix associated with a kernel function k and a given n data. We assume rank : $K = r$, and K can be decomposed as

$$K = P \Lambda P^\top, \quad P \in \mathbb{R}^{n \times r}, \tag{4.4}$$

where Λ is a diagonal matrix containing only the r positive eigenvalues of K in decreasing order. The empirical kernel feature map is defined by

$$\begin{aligned} &\Phi_r^e : \mathcal{X} \to \mathbb{R}^r \\ &\qquad \boldsymbol{x} \mapsto \Lambda^{-1/2} P^\top (k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n))^\top. \end{aligned}$$

Uncorrected Proof

Letting $Y = (\Phi_r^e(\boldsymbol{x}_1), \ldots, \Phi_r^e(\boldsymbol{x}_n))$, we obtain $YY^\top = KP\Lambda^{-1/2}\Lambda^{-1/2}P^\top K = K$, and we see that the empirical kernel feature map gives the same result of the kernel method applied to the data in the input space directly. We perform FDA for the transformed data set $\{\Phi_r^e(\boldsymbol{x}_i)\}_{i=1}^n$ to obtain a classification axis. That is, in the mapped space, we calculate the mean vectors of the two classes, between-class and total within-class covariance matrix, as

$$\boldsymbol{\mu}_l^e = \frac{1}{|\mathcal{D}_l|}\sum_{i\in\mathcal{D}_l}\Phi_r^e(\boldsymbol{x}_i), \quad l \in \mathcal{Y},$$

$$\Sigma_B^e = (\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})^\top,$$

$$\Sigma_W^e = \sum_{l\in\mathcal{Y}}\sum_{i\in\mathcal{D}_l}(\Phi_r^e(\boldsymbol{x}_i) - \boldsymbol{\mu}_l^e)(\Phi_r^e(\boldsymbol{x}_i) - \boldsymbol{\mu}_l^e)^\top,$$

respectively. Then, by solving the generalized eigenvalue problem,

$$\Sigma_B^e \boldsymbol{w} = \xi \Sigma_W^e \boldsymbol{w}, \quad \xi \in \mathbb{R}, \boldsymbol{w} \in \mathbb{R}^r, \tag{4.5}$$

and taking the first eigenvector, we can obtain the classification axis.

If the kernel function is optimized so that the data distribution of each class is a gaussian with the same covariance matrix, the resulting projection axis by discriminant analysis realizes a Bayes optimal classifier. We note that the result of FDA using the empirical kernel feature map is different from that of conventional KFDA (Mika et al., 1999), because KFDA requires a regularization to avoid rank degeneration. In our experiments, we let the rank r of a kernel matrix be the number of eigenvalues containing 90% of the total power.

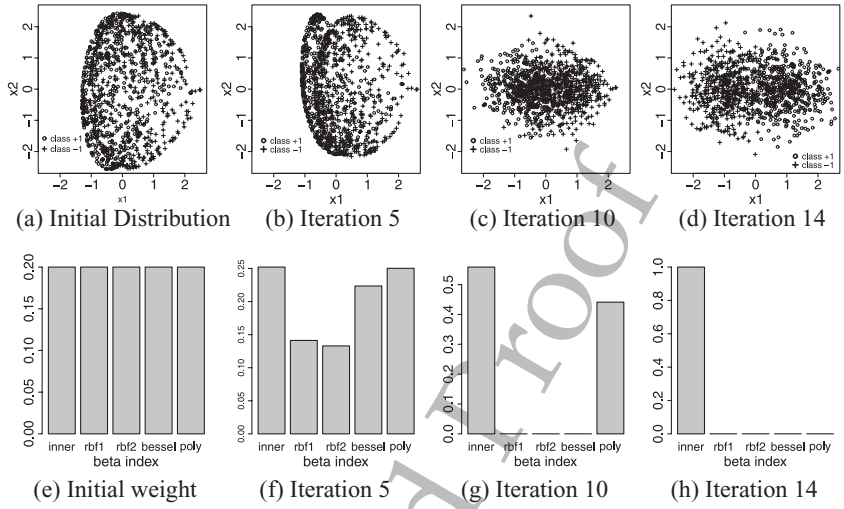
We also note that in principle, it is possible to optimize test statistics of gaussianity, such as Kolmogorov-Smirnov test or Shapiro-Wilk test statistics, calculated using the empirical kernel feature map. However, the effects of kernel combination coefficients for those test statistics can be highly nonlinear and complicated. On the other hand, we can evaluate gaussianity in the feature space easily with the empirical characteristic function. Thus, we can perform optimization with respect to $\boldsymbol{\beta}$ with ease.

5 Experimental Results

Numerical experiments for both artificial data and real-world data are presented in this section.

5.1 Illustrative Experiments with Artificial Data. We first show that the proposed GMKL methods work properly with artificial data. We generated two-dimensional dichotomy data from gaussian distributions with

Uncorrected Proof



(a) Initial Distribution (b) Iteration 5 (c) Iteration 10 (d) Iteration 14

(e) Initial weight (f) Iteration 5 (g) Iteration 10 (h) Iteration 14

Figure 3: (a–d) Data distribution in two-dimensional empirical kernel feature space. (e–g) Kernel combination coefficients. (h) *P*-values and objective values of minimization.

mean $\boldsymbol{\mu}_{+1} = (1, 0)^\top$ and $\boldsymbol{\mu}_{-1} = (-1, 0)^\top$ with the same covariance matrix $\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$ for each class. Five hundred samples are generated for each class of data. Five kernel functions, a linear kernel $k_1(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$, RBF kernels $k_2(\boldsymbol{x}, \boldsymbol{x}') = \exp(-2\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, $k_3(\boldsymbol{x}, \boldsymbol{x}') = \exp(-3\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$, a Bessel kernel $k_4(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|^2 J_2(2\|\boldsymbol{x} - \boldsymbol{x}'\|)$, where J_2 is the Bessel function of first kind, and a polynomial kernel $k_5(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^\top \boldsymbol{x}')^2$ are prepared for the combination of kernels. In Figures 3a and 3e, we show the distribution of the first two dimensions of the empirical kernel map of each class's data, transformed by a uniform combination of five kernels, and an initial combination of coefficients. In the initial uniform combination of the kernels, the data distribution in the empirical kernel space is not a gaussian, and two classes are completely mixed. We run the GGMKL algorithm for the combined kernel matrix to optimize the combination coefficients. Along with the iteration of the algorithm, the distribution of the data in the empirical kernel feature space gets close to a gaussian distribution, and combination coefficients for all kernels but the first linear kernel function get smaller. Finally, as shown in Figures 3d and 3h, the data distribution in each class becomes a gaussian, where the combination coefficient for the kernel k_1 is 1 and 0 for other four kernels. From this experiment, we see that GGMKL algorithm is able to optimize the kernel combination parameter so that the distribution of the data is a gaussian. Finally, Figure 4 shows the *p*-values obtained by Shapiro-Wilk test for the first two dimensions of the resulting

Uncorrected Proof

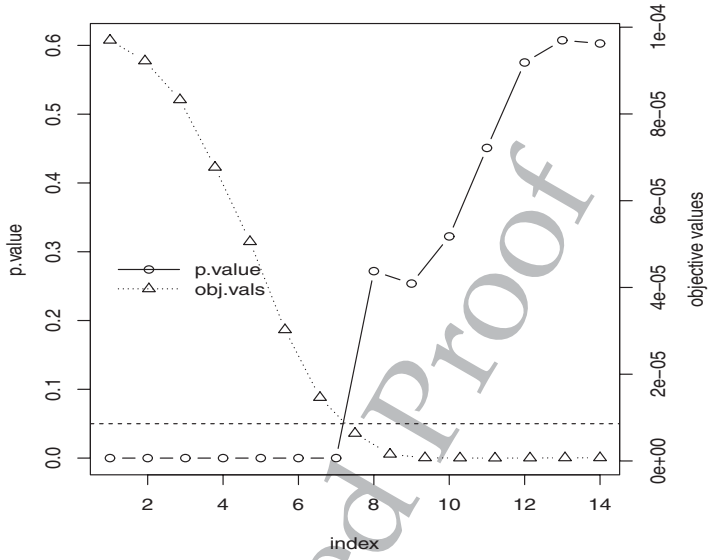


Figure 4: *P*-values and objective values of minimization.

Uncorrected Proof

empirical kernel feature map of the data and objective function values along with the iteration of the algorithm. The horizontal dashed line shows 0.05, which is usually taken as the significance level of the test of gaussianity. The objective function value uniformly decreased and the *p*-values increased as the progress of the algorithm. We note that a similar result was obtained by SQGMKL.

5.2 IDA Data Set. We employed the IDA data sets, which are standard binary classification data sets originally used in Rätsch, Onoda, and Müller (2001). The specifications of the data sets are shown in Table 1. We optimized the kernel function by the GMKL algorithm; then the test data are projected onto the axis found by FDA and classified by the large margin linear classifier in the same manner as Mika et al. (2000) described. We compared the GMKL algorithm to an equally weighted combination of kernels ($\beta_s = 1/S$), the best single kernel in candidate kernels, simpleMKL (Rakotomamonjy et al., 2008), SpicyMKL (Suzuki & Tomioka, 2011) with logit loss function, RMKL (Do et al., 2009), and l_2 MK-FDA (Yan et al., 2009). SimpleMKL is one of the most famous MKL methods, and SpicyMKL is its improved variant, which is especially effective when the number of kernel functions is large. For both methods, there are publicly available Matlab implementations. RMKL optimizes the kernel combination coefficients by both maximizing margin and minimizing the radius of the feature space. The concept of RMKL is similar to GMKL in the sense that it aims at

Table 1: IDA Data Specifications.

Data name	Input Data Dimensionality	Number of Train Samples	Number of Test Samples	Number of Realizations
Banana	2	400	4900	100
Breast-Cancer	9	200	77	100
Diabetes	8	468	300	100
Flare-Solar	9	666	400	100
German	20	700	300	100
Heart	13	170	100	100
Image	18	1300	1010	20
Ringnorm	20	400	7000	100
Splice	60	1000	2175	20
Thyroid	5	140	75	100
Titanic	3	150	2051	100
Twonorm	20	400	7000	100
Waveform	21	1000	1000	100

optimizing the geometry of the feature space. The l_2MK-FDA is a non-sparse MKL method based on Fisher's discriminant criterion. The method optimizes both kernel combination coefficients $\boldsymbol{\beta}$ and weight α_i for each datum $\boldsymbol{x}_i$ using semi-infinite programming (Hettich & Kortanek, 1993), while our proposed methods optimize only kernel combination coefficients.

We implemented all methods except simpleMKL, SpicyMKL, and l_2MK-FDA using R language (R Development Core Team, 2010). Quadratic programming in SQGMKL is solved by an interior point method equipped with a "kernlab" package (Karatzoglou, Smola, Hornik, & Zeileis, 2004) of R.

The combined kernel functions are the following 20 kernels:

- A linear kernel: $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top \boldsymbol{x}'$.
- Gaussian kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma |\boldsymbol{x} - \boldsymbol{x}'|^2), \sigma = 0.1, \ldots, 0.9$.
- Polynomial kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^\top \boldsymbol{x}' + 1)^d, d = 2, \ldots, 6$.
- Laplace kernels: $k(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\sigma |\boldsymbol{x} - \boldsymbol{x}'|), \sigma = 0.1, \ldots, 0.5$.

For all MKL methods, tuning parameters are chosen from candidate value sets so that the training error is minimized.

From Table 2, we can see that GGMKL and SQGMKL work favorably compared to other methods for many data in the light of classification accuracy.

Computational cost is another important aspect of learning methods. We implemented the proposed methods and RMKL methods using R programming language, and other MKL methods are mainly implemented using Matlab. The direct comparison is not possible; however, as a reference, we show comparative results on IDA data sets executed on the same

Uncorrected Proof

Uncorrected Proof

Table 2: Misclassification Rates (with Standard Deviations) by Various Classification Methods.

	Classification in Empirical Kernel Feature Space				Conventional MKLs			
Data name	GGMKL	SQGMKL	Uniform	Best Single	SimpleMKL	SpicyMKL	RMKL	l_2MK-FDA
Banana	**10.05**(0.42)	12.62(0.96)	18.12(6.08)	11.10(0.55)	11.47(0.60)	13.17(1.25)	**11.08**(0.53)	11.79(0.53)
Breast-Cancer	31.43(6.88)	**23.06**(3.62)	35.51(6.73)	28.97(4.31)	**25.77**(4.45)	26.99(4.77)	26.58(4.10)	28.60(4.29)
Diabetes	27.29(2.42)	24.55(1.81)	28.88(2.59)	26.78(2.83)	24.55(3.53)	**24.15**(1.68)	24.55(1.71)	**24.08**(1.89)
Flare-Solar	34.88(1.78)	**33.74**(1.63)	37.68(5.19)	**33.69**(1.84)	41.01(7.64)	35.07(1.76)	34.35(2.06)	34.43(1.55)
German	**22.37**(1.83)	**23.03**(2.10)	27.58(7.40)	23.98(2.64)	36.14(6.72)	27.03(2.09)	23.71(2.30)	23.55(2.23)
Heart	**14.83**(3.27)	15.83(3.28)	17.17(3.25)	16.11(2.92)	17.05(4.49)	**15.75**(3.25)	16.71(3.17)	16.81(3.68)
Image	**2.47**(0.52)	**2.40**(0.41)	3.35(1.50)	4.20(0.97)	10.90(1.13)	3.46(0.66)	3.87(0.74)	10.62(0.68)
Ringnorm	2.02(0.47)	1.59(0.12)	13.74(20.48)	**1.39**(0.31)	**1.53**(0.09)	1.63(1.55)	1.63(0.12)	2.67(0.44)
Splice	14.30(1.65)	**13.17**(0.76)	19.24(10.63)	14.43(1.75)	16.51(0.70)	**12.7**(0.72)	13.88(0.68)	15.92(0.87)
Thyroid	**3.65**(2.12)	**3.44**(1.93)	12.49(10.82)	5.35(2.53)	4.64(2.15)	5.29(2.83)	4.97(2.33)	5.17(2.23)
Titanic	**22.32**(1.12)	**21.92**(1.03)	32.24(2.79)	30.66(3.58)	23.20(3.00)	22.89(3.04)	22.93(0.99)	22.72(0.86)
Twonorm	3.83(3.75)	**2.31**(0.11)	2.46(0.15)	2.43(0.13)	2.55(0.15)	**2.37**(0.11)	2.49(0.15)	2.64(0.29)
Waveform	**9.44**(0.38)	**9.48**(0.52)	11.61(3.89)	10.36(0.90)	10.33(2.42)	13.16(0.60)	9.75(0.45)	9.73(0.63)

Note: The best and second-best results among the two methods in this table are shown in boldface.

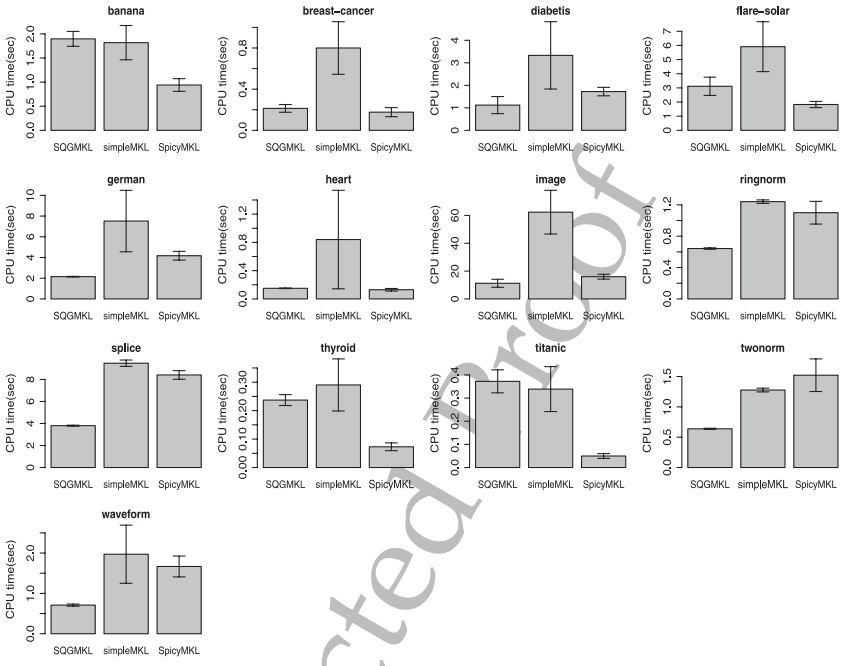


Figure 5: CPU time for learning kernel combination coefficients using SQGMKL, simpleMKL, and SpicyMKL methods.

computer in Figure 5.[1] We do not report the computational cost for GGMKL, RMKL, and l_2MK-FDA because it is more than five times slower than the slowest other methods. Figure 5 shows that the computational efficiency of the SQGMKL method is superior to that of simpleMKL for many data sets and comparable to SpicyMKL when a moderate number of kernels are combined.

6 Conclusion and Future Directions

In this study, we proposed to optimize kernel combination coefficients based on gaussianity in the feature space associated with the combined kernel functions. To our knowledge, there is currently no MKL method based on the notion of gaussianity in the feature space. Simple implementations of the proposed framework are given based on the empirical characteristic function. This function can be estimated using only a given set of kernel matrices

[1] All of numerical experiments in this letter are processed on an Intel machine with 2.93 GHz dual processors and 8 GB memory. The operation system is Mac OS X version 10.6.8.

Uncorrected Proof

and offers differentiable and bounded optimization objective. Through an experiment with artificial data, we see that the proposed algorithm maximizes gaussianity in the feature space associated with the learned kernel matrix. Using a number of benchmark data sets, we also show that the classification performances of the proposed GMKL methods are comparable or superior to other conventional MKL methods. One of the characteristics of the proposed GMKL framework is that learning of kernel combination coefficients is performed independently from the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ for each training datum. Intuitively, simultaneous optimization of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ leads to better classification accuracy; however, we know that we can obtain a Bayes optimal classifier by FDA if the data distribution is tailored to be a gaussian by GMKL, and the classification performance of the GMKL is comparable or superior to other methods. The proposed method is one of the optimal choices for kernel learning if FDA is used after learning the kernel. Application of the proposed GMKL methods followed by FDA to feature selection and visualization remains future work for us. Although we can obtain the Bayes optimal classifier by FDA under a certain condition on the data distributions, if the means of two classes are closely located, the Bayes error would be large. In this study, we proposed a basic concept of GMKL with a simple implementation; however, putting a term to separate the mean in each class would improve the classification accuracy with an additional cost of parameter tuning. It is important to investigate how to impose regularizations on the proposed GMKL framework to improve class separability.

In this letter, we adopted simple measures of gaussianity and the difference of covariance matrices, which are written in terms of the projected variances. There are other possibilities for those measures. For example, gaussianity and the difference of covariance matrices can be measured in the space of squared modulo of characteristic functions. Furthermore, it is also possible that M_G and M_V are defined using quantities of difference spaces. The classification accuracy and algorithmic stability might be improved by finding the optimal combination of measures.

Beside the realization of the framework based on the characteristic function, we can use other measures of gaussianity for the GMKL framework. For example, the entropy power inequality is a possible candidate for further investigation. Let the Shannon differential entropy of a random variable X be $h(X)$. In information theory, for n i.i.d. random variables, the following inequality is known:

$$N(X_1 + \cdots + X_n) \geq \sum_{i=1}^{n} N(X_i), \quad N(X) = \frac{1}{2\pi e} \exp\left(\frac{2}{d} h(X)\right).$$

The equality holds if and only if distributions of all the random variables are gaussians with proportional covariance matrices (Shannon, 1948; Stam, 1959). For two random variables, X_1 and X_2 correspond to two classes of

Uncorrected Proof

data; we can use the difference of the inequality as an objective function for gaussianity. The difficulty of this approach may be that we need entropy estimation. However, there are many entropy estimators based on the Euclidean distance of the observed data, and they might be formulated using only kernel function values. In addition to the alternative measure of gaussianity, the exploration of other measures of the gap between two covariance matrices M_V is an interesting work.

In the experiments, we obtained good classification accuracy, and the SQGMKL algorithm worked efficiently with a moderate number of kernel functions. Computational efficiency is one of the beneficial results of GMKL formulation, which concentrates only on the learning of kernel combination coefficients. In this study, we did not consider combining a huge number of kernels because the motivation for our GMKL framework was not feature selection via sparse representation but the construction of accurate classifiers by tailoring the distribution of the given data in the feature space. Unfortunately, the computational cost for SQGMKL would increase rapidly with the number of kernels to be combined because the method solves quadratic optimization programming iteratively, whereas by increasing the number of kernels, the ability of tailoring the feature space might be improved. Hence, development of a scalable MKL method based on the proposed framework will be explored in our future work.

Appendix A: Bayes Optimality of the Fisher Linear Discriminant

We show the following well-known fact for a two-class discriminant problem for the sake of completeness:

Theorem 2. *If the distributions of data in classes C_{+1} and C_{-1} are gaussians with means $\boldsymbol{\mu}_{+1}$ and $\boldsymbol{\mu}_{-1}$ and the same covariance matrix Σ, then the Fisher's discriminant analysis finds the best linear projection in the sense of Bayes error minimization.*

To prove the theorem, we recall that the linear projection vector $\boldsymbol{w}$ obtained by minimizing Fisher's criterion $J(\boldsymbol{w}) = \boldsymbol{w}^\top \Sigma_W \boldsymbol{w} / \boldsymbol{w}^\top \Sigma_B \boldsymbol{w}$ is given by $\boldsymbol{w}^* = \Sigma_W^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})$, where Σ_W and Σ_B are within-class and between-class covariance matrices, respectively. By the assumption of the theorem, $\Sigma_W = \Sigma$ and $\boldsymbol{w}^* = \Sigma^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})$.

We next consider a classifier defined by

$$f(\boldsymbol{x}) = \operatorname{sign}\left(\log \frac{P(C_{+1}|\boldsymbol{x})}{P(C_{-1}|\boldsymbol{x})}\right),$$

which achieves Bayes classification error by definition. By a simple calculation, we obtain

Uncorrected Proof

$$\begin{aligned}
\log \frac{P(C_{+1}|\boldsymbol{x})}{P(C_{-1}|\boldsymbol{x})} &= \log \frac{p(\boldsymbol{x}|C_{+1})P(C_{+1})/p(\boldsymbol{x})}{p(\boldsymbol{x}|C_{-1})P(C_{-1})/p(\boldsymbol{x})} = \log \frac{P(C_{+1})}{P(C_{-1})} + \log \frac{p(\boldsymbol{x}|C_{+1})}{p(\boldsymbol{x}|C_{-1})} \\
&= \log \frac{P(C_{+1})}{P(C_{-1})} \\
&\quad + \log \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{+1})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{+1}) \right. \\
&\quad \left. + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{-1})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{-1}) \right) \\
&= \{\Sigma^{-1}(\boldsymbol{\mu}_{+1}-\boldsymbol{\mu}_{-1})\}^\top \boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_{+1}\Sigma^{-1}\boldsymbol{\mu}_{+1} + \frac{1}{2}\boldsymbol{\mu}_{-1}\Sigma^{-1}\boldsymbol{\mu}_{-1} \\
&\quad + \log \frac{P(C_{+1})}{P(C_{-1})},
\end{aligned}$$

That is, under the assumption of the theorem 2, the Bayes optimal classifier is given by a linear classifier, and its projection vector is the same as that of obtained by Fisher's criterion.

Appendix B: Proof of Theorem 1

The necessity is obvious. Conversely, for a positive semidefinite matrix Σ, suppose $-\log|c(\boldsymbol{t})|^2 = \boldsymbol{t}^\top \Sigma \boldsymbol{t}$. This is equivalent to $|c(\boldsymbol{t})|^2 = c(\boldsymbol{t})c(-\boldsymbol{t}) = \exp(-\frac{1}{2}\boldsymbol{t}^\top 2\Sigma \boldsymbol{t})$. That is, $|c(\boldsymbol{t})|^2$ is nothing but the characteristic function of $N(0, 2\Sigma)$, and it is decomposed into two factors, $c(\boldsymbol{t})$ and $c(-\boldsymbol{t})$. Cramer's theorem claims that if the characteristic function of a gaussian distribution is factorized, then each factor characteristic function is also the characteristic function of a gaussian distribution (Lukacs, 1960). From Cramer's theorem, sufficiency of the theorem 1 is proved.

Appendix C: Effect of Evaluation Point in Empirical Characteristic Function

We show an empirical evaluation result of the effect of $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ in the empirical characteristic function $c_{\mathcal{D}}(\boldsymbol{t}; \boldsymbol{\beta})$. As we explained in section 3.2, we save one point in the given data set for $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ to evaluate the empirical characteristic function. Using the first realization of all the IDA data sets, we conduct a classification experiment using the different points in the given data set in the same setting as section 5.2. We show the mean and standard deviation of classification accuracies in the upper-left panel of Figure 6. We also show the means of obtained kernel combination coefficients in each element with

Uncorrected Proof

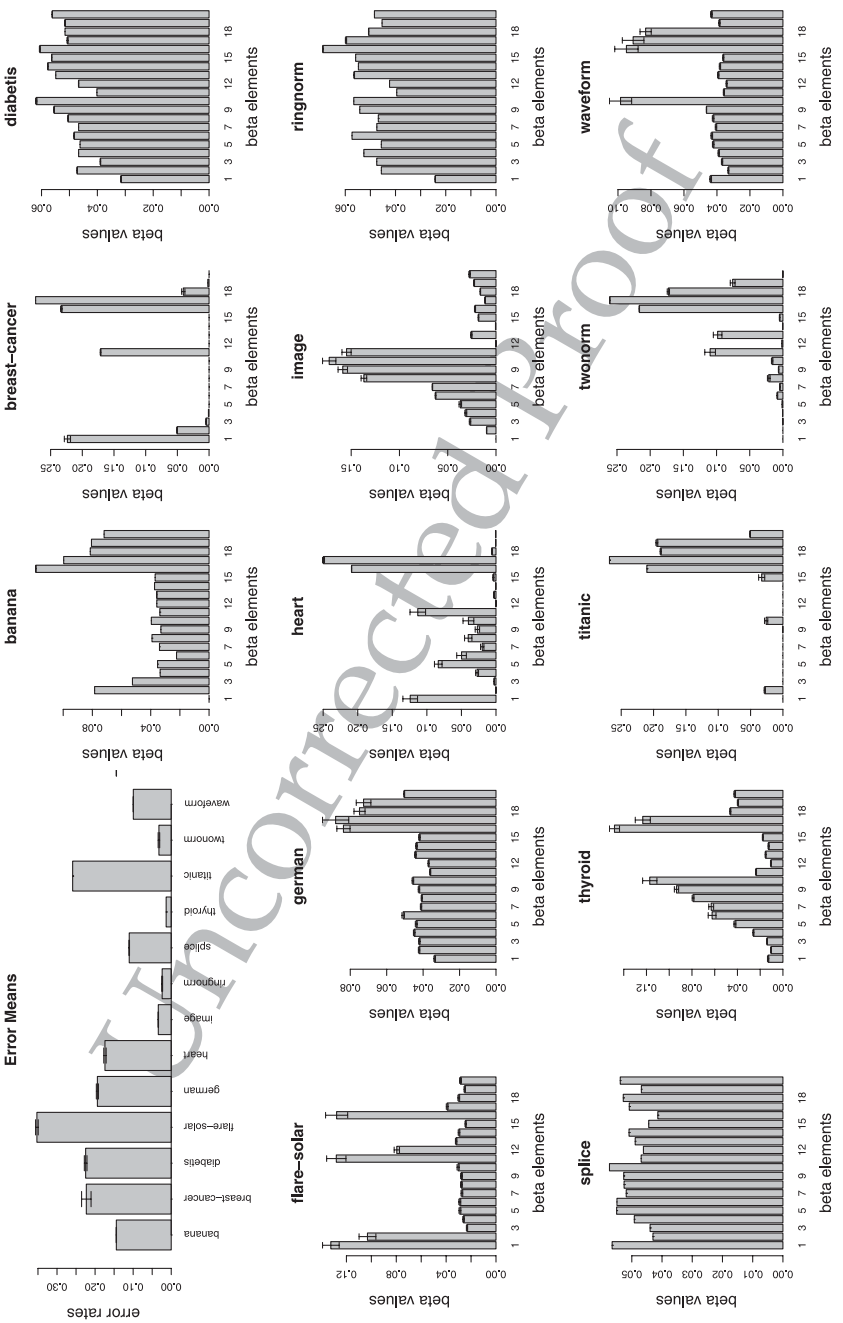


Figure 6: Means and standard deviations of classification accuracy and kernel combination coefficients with various evaluation points $\phi_\beta(t)$.

Uncorrected Proof

one standard deviation error bars in other panels of Figure 6. From this result, we conclude that the selection of $\phi_{\boldsymbol{\beta}}(\boldsymbol{t})$ from the given data does not affect the result much.

Appendix D: Gradient and Hessian of Objective Function

Gradient vectors of $J(\boldsymbol{\beta})$ and $\tilde{J}(\boldsymbol{\beta})$ and the Hessian matrix for $\tilde{J}(\boldsymbol{\beta})$ are explicitly given. For notational simplicity, we denote $S^*_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})$, $S_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})$, $|c^*_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})|^2$, and $|c_{\mathcal{D}_l}(\boldsymbol{t};\boldsymbol{\beta})|^2$ as S^*_l, S_l, $|c^*_l|^2$, and $|c_l|^2$, respectively. The gradient vector of

$$\begin{aligned} J(\boldsymbol{\beta}) &= \sum_{l\in\mathcal{Y}} M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) + \eta M_V(\phi_{\boldsymbol{\beta}}(\mathcal{D}_{+1}), \phi_{\boldsymbol{\beta}}(\mathcal{D}_{-1})) \\ &= \sum_{l\in\mathcal{Y}} \frac{|\mathcal{D}_l|}{n}(S^*_l - S_l)^2 + \eta(S^*_{+1} - S^*_{-1})^2 \end{aligned}$$

is given by

$$\nabla J(\boldsymbol{\beta}) = \sum_{l\in\mathcal{Y}} \nabla M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) + 4\eta\{\boldsymbol{\beta}^\top (V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})\boldsymbol{\beta}\}(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})\boldsymbol{\beta},$$

where

$$\nabla M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) = 2\frac{|\mathcal{D}_l|}{n}(S^*_l - S_l)\nabla(S^*_l + S_l),$$

and

$$\nabla S^*_l = V_{\mathcal{D}_l}\boldsymbol{\beta}, \quad \nabla S_l = \frac{\nabla|c_l|^2}{|c_l|^2}.$$

In the above formula, the gradient vectors of $|c_l|^2$ are given by

$$\nabla|c_l|^2 = -\frac{2}{|\mathcal{D}_l|^2} \sum_{\substack{i>j \\ \boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{D}_l}} \sin(\boldsymbol{d}(\boldsymbol{t},\boldsymbol{x}_i,\boldsymbol{x}_j)^\top\boldsymbol{\beta}) \cdot \boldsymbol{d}(\boldsymbol{t},\boldsymbol{x}_i,\boldsymbol{x}_j),$$

where

$$\boldsymbol{d}(\boldsymbol{t},\boldsymbol{x}_i,\boldsymbol{x}_j) = \begin{pmatrix} k_1(\boldsymbol{t},\boldsymbol{x}_i) - k_1(\boldsymbol{t},\boldsymbol{x}_j) \\ \vdots \\ k_S(\boldsymbol{t},\boldsymbol{x}_i) - k_S(\boldsymbol{t},\boldsymbol{x}_j) \end{pmatrix} \in \mathbb{R}^S.$$

Uncorrected Proof

We next consider a vector $\boldsymbol{c}_t$ and a matrix H_t in equation 4.3, which are a gradient vector and a Hessian matrix of

$$\tilde{J}(\boldsymbol{\beta}) = \sum_{l \in \mathcal{Y}} M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) + \eta \boldsymbol{\beta}^\top \{(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top + \lambda I_S\}\boldsymbol{\beta},$$

evaluated at $\boldsymbol{\beta}_t$. Then $\boldsymbol{c}_t$ and H_t are explicitly written as

$$\boldsymbol{c}_t = \sum_{l \in \mathcal{Y}} \nabla M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_t} + 2\eta((V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top + \lambda I_S)\boldsymbol{\beta}_t,$$

$$H_t = \sum_{l \in \mathcal{Y}} \nabla^2 M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) \Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_t} + 2\eta((V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})(V_{\mathcal{D}_{+1}} - V_{\mathcal{D}_{-1}})^\top + \lambda I_S).$$

In the above formula,

$$\begin{aligned}
&\frac{1}{2}\nabla^2 M_G(\phi_{\boldsymbol{\beta}}(\mathcal{D}_l)) \\
&= \frac{|\mathcal{D}_l|}{n}\nabla^2 (S_l^* - S_l)^2 \\
&= \frac{|\mathcal{D}_l|}{n}\left\{(\nabla^2 S_l^* - \nabla^2 S_l)(S_l^* - S_l) + (\nabla S_l^* - \nabla S_l)(\nabla S_l^* - \nabla S_l)^\top\right\},
\end{aligned}$$

where

$$\nabla^2 S_l^* = V_{\mathcal{D}_l}, \quad \nabla^2 S_l = -\frac{(\nabla^2 |c_l|^2)|c_l|^2 - (\nabla |c_l|^2)(\nabla |c_l|^2)^\top}{(|c_l|^2)^2}$$

and

$$\nabla^2 |c_l|^2 = \frac{2}{|\mathcal{D}_l|^2} \sum_{\substack{i>j \\ \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}_l}} \cos(\boldsymbol{d}(\boldsymbol{t}, \boldsymbol{x}_i, \boldsymbol{x}_j)^\top \boldsymbol{\beta}) \cdot \boldsymbol{d}(\boldsymbol{t}, \boldsymbol{x}_i, \boldsymbol{x}_j)\boldsymbol{d}(\boldsymbol{t}, \boldsymbol{x}_i, \boldsymbol{x}_j)^\top.$$

Acknowledgments

We are grateful to A. Rakotomamonjy, Y. Grandvalet, F. Bach, and S. Canu for providing simple MKL Matlab implementation and to T. Suzuki and R. Tomioka for providing SpicyMKL Matlab implementation. We are also grateful to F. Yan for providing l_p MK-FDA Matlab implementation. Parts of experiments were done with the help of T. Aritake. We express our special thanks to the editor and reviewers, whose comments led to valuable improvements of the manuscript. Part of this work was supported by JSPS Grant-in-Aid for Research Activity Start-up No. 22800067.

Uncorrected Proof

References

Amari, S., & Wu, S. (1999). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*(6), 783–789.

Bach, F., Lanckriet, G., & Jordan, M. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the Twenty-First International Conference on Machine Learning*. New York: ACM.

Bennett, K. P., Momma, M., & Embrechts, M. J. (2002). Mark: A boosting algorithm for heterogeneous kernel models. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 24–31). New York: ACM.

Bi, J., Zhang, T., & Bennett, K. P. (2004). Column-generation boosting methods for mixture of kernels. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 521–526). New York: ACM.

Bousquet, O., & Herrmann, D.J.L. (2002). On the complexity of learning the kernel matrix. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 399–406). Cambridge, MA: MIT Press.

Crammer, K., Keshet, J., & Singer, Y. (2002). Kernel design using boosting. In S. Becker, S. Thrün, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15* (pp. 537–544). Cambridge, MA: MIT Press.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Cristianini, N., Shawe-Taylor, J., & Kandola, J. (2001). On kernel target alignment. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14* (pp. 367–373). Cambridge, MA: MIT Press.

Do, H., Kalousis, A., Woznica, A., & Hilario, M. (2009). Margin and radius based multiple kernel learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Datasets* (pp. 330–343). Berlin: Springer.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. New York: Wiley-Interscience.

Eriksson, J., & Koivunen, V. (2003). Characteristic-function-based independent component analysis. *Signal Processing*, *83*(10), 2195–2208.

Feuerverger, A., & Mureika, R. (1977). The empirical characteristic function and its applications. *Annals of Statistics*, *5*, 88–97.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals Eugenics*, *7*, 179–188.

Hettich, R., & Kortanek, K. O. (1993). Semi-infinite programming: Theory, methods, and applications. *SIAM Review*, *35*, 380–429.

Karatzoglou, A., Smola, A. J., Hornik, K., & Zeileis, A. (2004). Kernlab: An S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20.

Kim, S., Magnani, A., & Boyd, S. (2006). Optimal kernel selection in kernel Fisher discriminant analysis. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 465–472). New York: ACM Press.

Kloft, M., Brefeld, U., Laskov, P., & Sonnenburg, S. (2008). Non-sparse multiple kernel learning. In *Proceedings of the NIPS Workshop on Kernel Learning: Automatic Selection of Kernels* (pp. 1179–1225). Cambridge, MA: MIT Press.

Uncorrected Proof

Koutrouvelis, I. (1980). A goodness-of-fit test of simple hypotheses based on the empirical characteristic function. *Biometrika*, *67*(1), 238–240.

Lanckriet, G., Cristianini, N., Bartlett, P., Ghaoui, L., & Jordan, M. (2004). Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, *5*, 27–72.

Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces*. New York: Springer.

Lukacs, E. (1960). *Characteristic functions*. London: Charles Griffin & Company.

Micchelli, C., & Pontil, M. (2005). Learning the kernel function via regularization. *Journal of Machine Learning Research*, *6*, 1099–1125.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., & Müllers, K. (1999). Fisher discriminant analysis with kernels. In *Proceedings of Neural Networks for Signal Processing IX, 1999* (pp. 41–48). Piscataway, NJ: IEEE.

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A. J., & Müller, K. R. (2000). Invariant feature extraction and classification in kernel spaces. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12* (pp. 526–532). Cambridge, MA: MIT Press.

Murata, N. (2001). Properties of the empirical characteristic function and its application to testing for independence. In *Proceedings of 3rd International Conference on Independent Component Analysis and Signal Separation* (pp. 295–300). San Diego: University of California, San Diego, Institute for Neural Computation.

Murota, K., & Takeuchi, K. (1981). The studentized empirical characteristic function and its application to test for the shape of distribution. *Biometrika*, *68*(1), 55–65.

Rakotomamonjy, A., Bach, F., Canu, S., & Grandvalet, Y. (2008). SimpleMKL. *Journal of Machine Learning Research*, *9*, 2491–2521.

Rätsch, G., Onoda, T., & Müller, K. R. (2001). Soft margins for Adaboost. *Machine Learning*, *42*(3), 287–320.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Reed, M., & Simon, B. (1981). *Functional analysis*. San Diego: Academic Press.

Schölkopf, B., Mika, S., Burges, C.J.C., Knirsch, P., Müller, K. R., Rätsch, G., et al. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks*, *10*, 1000–1017.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, *27*, *379–423*, 623–656.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Sonnenburg, S., Rätsch, G., Schäfer, C., & Schölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, *7*, 1531–1565

Stam, A. J. (1959). Some inequalities satisfied by the quantities of information of Fisher and Shannon. *Information and Control*, *2*(2), 101–112.

Suzuki, T., & Tomioka, R. (2011). SpicyMKL: A fast algorithm for multiple kernel learning with thousands of kernels. *Machine Learning*, *85*(1), 77–108.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley-Interscience.

Xiong, H., Swamy, M.N.S., & Ahmad, M. O. (2005). Optimizing the kernel in the empirical feature space. *IEEE Transactions on Neural Networks*, *16*(2), 460–474.

Uncorrected Proof

Yan, F., Kittler, J, Mikolajczyk, K., & Tahir, A. (2009). Non-sparse multiple kernel learning for Fisher discriminant analysis. In *Proceedings of Ninth IEEE International Conference on Data Mining* (pp. 1064–1069). Piscataway, NJ: IEEE.

Yuille, A. L., & Rangarajan, A. (2003). The concave-convex procedure. *Neural Computation*, *15*(4), 915–936.

Received September 18, 2011; accepted January 8, 2012.

Uncorrected Proof

This article has been cited by: