Testing Hypotheses by Regularized Maximum Mean Discrepancy

Somayeh Danafar
IDSIA/SUPSI
Università della Svizzera Italiana
Lugano, Switzerland
Email: somayeh {at} idsia.ch

Paola M.V. Rancoita
CUSSB, Vita-Salute San Raffaele University
Milan, Italy

Tobias Glasmachers
Institut für Neuroinformatik, Ruhr-Universität
Bochum, Germany

Kevin Whittinstall
Dept. of Diagnostic Radiology, Université de Sherbrooke
Sherbrooke Molecular Imaging Center, Université de Sherbrooke
QC, Canada

Jürgen Schmidhuber
IDSIA/SUPSI
Università della Svizzera Italiana
Lugano, Switzerland

Abstract— **Regularized Maximum Mean Discrepancy (RMMD), our novel measure for kernel-based hypothesis testing, excels at hypothesis tests involving multiple comparisons with power control even when sample sizes are small. We derive asymptotic distributions under the null and alternative hypotheses, and assess power control. Outstanding results are obtained on challenging benchmark datasets.**

Keywords- kernel-based hypothesis testing, Homogeneity testing, Multiple comparisons, Power

I. Introduction

Homogeneity testing is an important problem in statistics and machine learning. It tests whether two samples are drawn from different distributions. This is relevant for many applications, for instance, schema matching in databases [1] and speaker identification [2]. Popular two-sample tests like Kolmogorov-Smirnov [3] and Cramer-von-Mises [4] are not capable of capturing statistical information of densities with high frequency features. Nonparametric kernel-based statistical tests such as Maximum Mean Discrepancy (MMD) [1], [5] enable one to obtain greater power than such density based methods. MMD is applicable not only to Euclidean spaces $\mathbb{R}^n$, but also to groups and semigroups [6], and to structures such as strings or graphs in bioinformatics, and robotics problems, etc. [7]. Here we consider a regularized version of MMD to address hypothesis testing.

With more than two distributions to be compared simultaneously, we face the multiple comparisons setting, for which statistical methods exist to deal with the issue of multiple test correction [8]. Given a prescribed global significance threshold α (type I error) for the set of all comparisons, however, the corresponding threshold per comparison becomes small, which greatly reduces the power of the test. In situations where one wants to retain the null hypothesis, tests with small α are not conservative. Our main contribution is the definition of a regularized MMD (RMMD) method.

The regularization term in RMMD allows controlling the power of the test statistic. The regularizer is set **provably optimal** for maximal power; there is no need for fine-tuning by the user. RMMD improves on MMD through higher power, especially for small sample sizes, while preserving the advantages of MMD. **Power control** enables us to look for true sets of null distributions among the significant ones in challenging multiple comparison tasks.

We provide experimental evidence of good performance on a challenging Electroencephalography (EEG) dataset, artificially generated periodic and Gaussian data, the CIFAR10, the MNIST and Covertype datasets. We also assess power control with the Asymptotic Relative Efficiency (ARE) test.

The paper is organized as follows. In section 2, we elaborate on hypothesis testing and define maximum mean discrepancy (MMD) as a metric. We describe how to use MMD for homogeneity testing, and how to extend it to multiple comparisons. In section 3, we define RMMD for hypothesis testing and compare it to MMD and Kernel Fisher Discriminant Analysis (KFDA), and assess power control through ARE. Additional empirical justification of our test on various datasets is presented in section 4.

II. STATISTICAL HYPOTHESIS TESTING

A statistical hypothesis test is a method which, based on experimental data, aims to decide whether a hypothesis (called null or H_0) is true or false, against an alternative hypothesis (H_1). The level of significance α of the test represents the probability of rejecting H_0 under the assumption that H_0 is true (type I error). A type II error (β) occurs when we reject H_1 although it holds.

The **power** of the statistical test is usually defined as 1-β. A desirable property of a statistical test is that for a prescribed global significance level α, the power equals one in the population limit. We divide the discussion of hypothesis testing into two topics: homogeneity testing and multiple comparisons.

A. Maximum Mean Discrepancy (MMD)

Embedding probability distributions into Reproducing Kernel Hilbert Spaces (RKHS) yields a linear method that takes information of higher order statistics into account [1], [9], [10]. Characteristic kernels [6], [10], [11], injectively map the probability distribution onto its mean element in the corresponding RKHSs. The distance between **mean elements** (μ) in the RKHS is known as MMD [1], [5]. The definition of MMD [1] is given in the following theorem:

Theorem 1. Let $(\mathcal{X}, \mathcal{B})$ be a metric space, and let P, Q be two Borel probability measures defined on $\mathcal{X}$. The kernel function $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ embeds the points $x \in \mathcal{X}$ into the corresponding reproducing kernel Hilbert space $\mathcal{H}$. Then $P = Q$ if and only if *MMD(P,Q) = 0*, where

$$MMD(P,Q) \coloneqq \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}} = \left\|E_P[k(x,.)] - E_Q[k(x,.)]\right\|_{\mathcal{H}} = \left(E_{x,x'\sim P}[k(x,x')] + E_{y,y'\sim Q}[k(y,y')] - 2E_{x\sim P, y\sim Q}[k(x,y)]\right)^{\frac{1}{2}} \qquad (1)$$

B. Homogeneity Testing

A two-sample test investigates whether two samples are generated by the same distribution. To do testing, MMD can be used to measure the distance between embedded probability distributions in RKHS. Besides calculating the distance measure, we need to check whether this distance is significantly different from zero. For this, the asymptotic distribution of this distance measure is used to obtain a threshold on MMD values, and to extract the statistically significant cases. We perform a hypothesis test with null hypothesis H_0: $P = Q$ and alternative H_1: $P \neq Q$ on samples drawn from two distributions P and Q. If the result of MMD is close enough to zero, we accept H_0, which indicates that the distributions P and Q coincide; otherwise the alternative is assumed to hold. With α as a threshold on the asymptotic distribution of the empirical MMD (when$P = Q$), the $(1 - \alpha) -$ quantile of this distribution is statistically significant. Our MMD test determines it by means of a bootstrap procedure.

C. Multiple Comparisons

Statistical analysis of a data set typically needs testing many hypotheses. The multiple comparisons or multiple testing problem arises when we evaluate several statistical hypotheses simultaneously. Let α be the overall type I error, and let $\bar{\alpha}$ denote the type I error of a single comparison in the multiple testing scenario. Maintaining the prescribed significance level of α in multiple comparisons yields $\bar{\alpha}$ to be more stringent than α. Nevertheless, in many studies $\alpha = \bar{\alpha}$ is used without correction. Several statistical techniques have been developed to control α [8]. We use the Dunn-Ŝidák method for *n* independent comparisons in multiple testing, the significance level α is obtained by$\alpha = 1 - (1 - \bar{\alpha})^n$. As α decreases, the probability of type II error (β) increases and the power of the test decreases. This requires controlling β while correctingα. To tackle this problem, and to control β, we define a new hypothesis test based on RMMD, which has higher power than the MMD-based test, in the next section. To compare the distributions in the multiple testing problem we use two approaches: one-vs-all and pairwise comparisons. In the one-vs-all case each distribution is compared to all other distributions in the family, thus *M* distributions require *M-1* comparisons. In the pairwise case each pair of distributions is compared at the cost of $\frac{M(M-1)}{2}$ comparisons.

III. REGULARIZED MAXIMUM MEAN DISCREPANCY (RMMD)

The main contribution of this paper is a novel regularization of MMD measure called RMMD. This regularization aims to provide a test statistics with greater power (power closer to 1 with a prescribed type I error α). Erdogmus and Principe [12] showed that $-\log\|\mu_P\|_{\mathcal{H}}^2$ is the Parzen window estimation of the Renyi entropy [13]. With RMMD we obtain a statistical test with greater power by penalizing the term $\|\mu_P\|_{\mathcal{H}}^2 + \left\|\mu_Q\right\|_{\mathcal{H}}^2$. We formulate RMMD and its empirical estimator as follows:

$$RMMD(P,Q) \coloneqq MMD(P,Q)^2 - \kappa_P\|\mu_P\|_{\mathcal{H}}^2 - \kappa_Q\left\|\mu_Q\right\|_{\mathcal{H}}^2 \quad (2)$$

$$\widehat{RMMD}(P,Q) \coloneqq \left\|\hat{\mu}_P - \hat{\mu}_Q\right\|_{\mathcal{H}}^2 - \kappa_P\|\hat{\mu}_P\|_{\mathcal{H}}^2 - \kappa_Q\left\|\hat{\mu}_Q\right\|_{\mathcal{H}}^2 \quad (3)$$

where κ_P, and κ_Q are nonnegative regularization constants. For simplicity we consider $\kappa_P = \kappa_Q = \kappa$ in many application, however, we can introduce prior knowledge about the complexity of distributions by choosing $\kappa_P \neq \kappa_Q$. The modified Jensen-Shanon divergence (JS) [14] corresponding to RMMD is defined as:

$$D(P,Q) \coloneqq H_S(P,Q) - (\kappa + 1)\big(H_S(P) + H_S(Q)\big) \qquad (4)$$

where H$_s$ denotes the (cross) entropy. Since κ is positive, the absolute value of second term on the right-hand side of (4) increases, leading to a higher weight for the mutual information than for the entropy (vice versa if κ would be lower than -1). [1]

Here we summarize the notation needed in the next section. Given samples $\{x_i\}_{i=1}^{n_1}$and $\{y_i\}_{i=1}^{n_2}$drawn from distributions P and Q, respectively, the mean element, the cross-covariance operator and the covariance operator are defined as follows [1], [15]: $\hat{\mu}_P = \frac{1}{n_1}\sum_{i=1}^{n_1} k(x_i,.)$, $\hat{\Sigma}_{PQ} = \frac{n_1 n_2}{n_1+n_2}\left(\hat{\mu}_P - \hat{\mu}_Q\right) \otimes \left(\hat{\mu}_P - \hat{\mu}_Q\right)$, and $\hat{\Sigma}_P = \frac{1}{n_1}\sum_{i=1}^{n_1}(k(x_i,.) \otimes k(x_i,.)) - (\hat{\mu}_P \otimes \hat{\mu}_Q)$, where $u \otimes v$ for $u, v \in \mathcal{H}$ is defined for all $f \in \mathcal{H}$ as $(u \otimes v)f = \langle v, f\rangle_{\mathcal{H}} u$. The quantities $\hat{\mu}_Q$ and $\hat{\Sigma}_Q$are defined analogously for the second sample $\{y_i\}_{i=1}^{n_2}$. The population counterparts, i.e., the population mean element and the population covariance operator are defined for any probability measure P as $\langle \mu_P, f\rangle_{\mathcal{H}} = E[f(x)]$ for all $f \in \mathcal{H}$, and $\langle f, \Sigma_P g\rangle_{\mathcal{H}} = cov_P[f(x), g(y)]$ for $f, g \in \mathcal{H}$. From now on we call $\Sigma_B = \Sigma_{PQ}$ the *between-distribution covariance*. The pooled covariance operator (which we call also the *within-distribution covariance*) is denoted by: $\Sigma_W = \frac{n_1}{n_1+n_2}\Sigma_P + \frac{n_2}{n_1+n_2}\Sigma_Q$.

A. *Limit distribution of RMMD Under Null and Fixed Alternative Hypotheses*

Now we derive the distribution of the test statistics under the null hypothesis of homogeneity H$_0$: $P = Q$ (Theorem 2), which implies $\mu_P = \mu_Q$ and $\Sigma_P = \Sigma_Q = \Sigma_W$. Consistency of the test is guaranteed by the form of the distribution under H$_1$: $P \neq Q$ (Theorem 2). Assume that $\{x_i\}_{i=1}^{n_1}$ and $\{y_i\}_{i=1}^{n_2}$ are independent samples from P and Q, respectively (a priori they are not equally distributed). Let $z_i \coloneqq (x_i, y_i)$, $h(z_i, z_j) \coloneqq k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i) - h'(z_i, z_j)$, and $h'(z_i, z_j) = \kappa_P k(x_i, x_j) + \kappa_Q k(y_i, y_j)$, and $\overset{D}{\rightarrow}$ denotes convergence in distribution. Without loss of generality we assume $n_1 = n_2 = n$, and $\kappa_P = \kappa_Q = \kappa$. The proofs hold even when $\kappa_P \neq \kappa_Q$. Based on Hoeffding [16], Theorem A (p. 192) and Theorem B (p. 193) by Serfling [17], we can prove the following theorem:

Theorem 2. If $E[h^2] < \infty$, under *H$_1$*, $\widehat{RMMD}$ is asymptotically normally distributed

$$\sqrt{n}\left(\widehat{RMMD} - RMMD\right) \overset{D}{\rightarrow} \mathcal{N}(0, \hat{\sigma}^2)$$

with variance $\hat{\sigma}^2 = 4\left(E_z\left[E_{z'}[h(z,z')^2]\right] - E^2{}_{z,z'}[h(z,z')]\right)$, uniformly at rate $1/\sqrt{n}$. Under *H$_0$*, the same convergence holds with $\hat{\sigma}^2 = 4\left(E_z\left[E_{z'}[h'(z,z')^2]\right] - E^2{}_{z,z'}[h'(z,z')]\right) > 0$.

To increase the power of our RMMD-based test we need **to decrease the variance under H$_1$** in Theorem 2. The following Theorem can be used to obtain maximal power by setting $\kappa = 1$. This will give us a fixed hyper-parameter – no need for user tuning. The optimal value of κ decreases both the variance of H$_1$ and H$_0$ simultaneously and the fixed α is defined over the changed variance of H$_0$.

Theorem 3. The **highest power** of RMMD is obtained for κ =1.

Proof. Let denote $A = k(x_i, x_j) + k(y_i, y_j)$, and $B = k(x_i, y_j) + k(x_j, y_i)$. Based on Theorem 2, the variance under H$_1$ is obtained by:

$$\begin{aligned} \hat{\sigma}^2 &= 4\left(E_z\left[E_{z'}[h(z,z')^2]\right] - E^2{}_{z,z'}[h(z,z')]\right) \\ &= 4\left(E\left[\left((1-\kappa)A - B\right)^2\right] - (E^2[(1-\kappa)A - B])\right) \\ &= 4\left((1-\kappa)^2(E[A^2] - E^2[A]) + E[B^2] - E^2[B]\right) \\ &= 4\left((1-\kappa)^2 \mathrm{var}(A) + \mathrm{var}(B)\right), \end{aligned} \tag{5}$$

where $\mathrm{var}(A)$, and $\mathrm{var}(B)$ denote the variances. To get maximal power, we set

$$\frac{\partial\left((1-\kappa)^2\mathrm{var}(A)+\mathrm{var}(B)\right)}{\partial\kappa} = 0, \tag{6}$$

which yields $\kappa = 1$.

B. *Comparisn between RMMD, MMD, and KFDA*

According to Theorem 8 by Gretton et al. [1], under the null hypothesis the test statistics of MMD degenerates. This corresponds to $\hat{\sigma}^2 = 0$ in our Theorem 2. For large sample sizes the null distribution of MMD approaches in distribution as an infinite weighted sum of independent χ_1^2 random variables, with weights equal to the eigenvalues of the within-distribution covariance operator Σ_W . If we denote the test statistics based on MMD by $\hat{T}_n^{MMD}$, then $\hat{T}_n^{MMD} \overset{D}{\rightarrow} C\sum_{l=1}^{\infty}\lambda_l(z_l^2 - 1)$, where $z_l \sim \mathcal{N}(0,2)$ are i.i.d. random variables, and C is a scaling factor. Harchaoui et al. [2], [18] introduced Kernel Fisher Discriminant Analysis (KFDA) as a homogeneity test by regularizing MMD with the within-distribution covariance operator. The maximum Fisher discriminant ratio defines this test statistic. The empirical KFDA test statistic is denoted as $\widehat{KFDA}(P,Q) = \frac{n_1 n_2}{n_1+n_2}\left\|\frac{\hat{\mu}_P - \hat{\mu}_Q}{(\hat{\Sigma}_P + \gamma_n I)^{1/2}}\right\|_{\mathcal{H}}^2$. To analyze the asymptotic behaviour of this statistics under the null hypothesis, Harchaoui et al. [2] consider two situations regarding the regularization parameter γ_n : 1) one where γ_n is held fixed, obtaining the limit distribution similar to MMD under H$_0$; 2) one where γ_n tends to zero slower than $n^{-1/2}$. In the first situation the test statistic converges to $\hat{T}_n^{KFDA(\gamma_n)} \overset{D}{\rightarrow} C\sum_{l=1}^{\infty}(\lambda_l + \gamma_n)^{-1}\lambda_l(z_l^2 - 1)$. Thus, the test statistics based on KFDA normalizes the weights of χ_1^2 random variables by using the covariance operator as the regularizer. In comparison MMD is more sensitive to the information of higher order moments because of their bigger weights (larger eigenvalues of the covariance

[1] RMMD with negative-valued κ can be used in clustering as a divergence to compare clusters. We achieve greater entropy with broader clusters. The resulting clustering method avoids overfitting with narrow clusters.

operator). In the second situation (applicable in practice only for very large sample sizes) the test statistics converges to $\hat{T}_n^{KFDA(\gamma_n)} \xrightarrow{D} \mathcal{N}(C,1)$, where C is a constant.

The asymptotic convergence of the test statistic based on RMMD is $\hat{T}_n^{RMMD} \xrightarrow{D} \mathcal{N}(0,\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the variance of the function h in Theorem 2. The precise analytical normal distribution obtains higher power in RMMD. Because of the divergence ($\sigma^2 = 0$ in the asymptotic distribution) for MMD and KFDA, they use an estimation of the distribution under the null hypothesis which loses the accuracy and affect the power. In contrast to MMD and KFDA, RMMD is consistent since the divergence under the null hypothesis does not happen anymore. RMMD is the generalized form of the test statistics based on MMD, which we obtain for $\kappa = 0$. Moreover, by minimizing the variance of the normal distribution, we obtain the best power for $\kappa = 1$ and thus the hyper-parameter κ is fixed without requiring tuning by the user.

In comparison to KFDA, RMMD does not require restrictive constraints to obtain high power. It also results in higher power than MMD and KFDA in cases with small sample size. The speed of power convergence in KFDA is $O_p(1)$ which is slower than $O_p(n^{-1/2})$ in RMMD when $n \to \infty$.

Regarding the computational complexity, for MMD a parametric model with lower order moments of the test statistics is used to estimate the value of MMD which degenerates under H$_0$, and which has no consistency or accuracy guarantee. In comparison, the bootstrap resampling and the eigen-spectrum of the gram matrix are more consistent estimates with computational cost of $O(n^2)$, where n is the number of samples [19]. For RMMD, the convergence of the test statistic to a Normal distribution enables a fast, consistent and straightforward estimation of the null distribution within $O(n^2)$ time without the need of using an estimation method. The results of power comparison between these tests are reported in section 4.

C. *Assyptotic Relative Efficiency of Statistical Tests*

To assess the power control we use the asymptotic relative efficiency. This criterion shows that RMMD is a better test statistic and obtains higher power rather than KFDA and MMD with smaller sample size. Relative efficiency enables one to select the most effective statistical test quantitatively [20]. Let *T* and *V* be test statistics to be compared. The necessary sample size for the test statistics *T* to achieve the power *1-* β with the significance level α is denoted by $N_T(\alpha, 1-\beta)$. The relative efficiency of the statistical test *T* with respect to the statistical test *V* is given by:

$$e_{T,V}(\alpha,1-\beta) = N_V(\alpha,1-\beta)/N_T(\alpha,1-\beta)\,. \qquad (7)$$

Since calculating $N_T(\alpha,1-\beta)$ is hard even for the simplest test statistics, the limit value $e_{T,V}(\alpha,1-\beta)$, as $1-\beta \to 1$, is used. The limiting value is called the Bahadur Asymptotic Relative Efficiency (ARE) denoted by $e^B_{T,V}$,

$$e^B_{T,V} := \lim_{1-\beta\to 1} e_{T,V}(\alpha,1-\beta), \qquad (8)$$

The test statistic *V* is considered better than *T*, if $e_{T,V}$ is smaller than 1, because it means that *V* needs a lower sample size to obtain a power of $1-\beta$, for the given α. In [2], the authors assessed the power control by means of analysis of local alternatives which work when we have very large sample size or when n tends to infinity. In this article, we focus our attention on the small sample size case, which is more challenging. In section 4, we compute $e^B_{MMD,RMMD} = \frac{N_{RMMD}}{N_{MMD}}$, $e^B_{MMD,KFDA} = \frac{N_{KFDA}}{N_{MMD}}$ and $e^B_{KFDA,RMMD} = \frac{N_{RMMD}}{N_{KFDA}}$ using artificial datasets and two types of kernels, and we obtain smaller ARE for RMMD rather than KFDA and MMD. This means RMMD gives higher power with much smaller sample size. Results for different data sets are reported in Table 2, Figure 2, and Figure 3.

IV. EXPERIMENTS

MMD [1] was experimentally shown to outperform many traditional two-sample tests such as the generalized Wald-Wolfowitz test, the generalized Kolm-ogorov-Smirnov (KS) test [3], the Hall-Tajvidi (Hall) test [21], and the Biau-Györf test. It was shown [2] that KFDA outperforms the Hall-Tajvidi test. We select KS and Hall as traditional baseline methods, on top of which we compare RMMD, KFDA, and MMD. To experimentally evaluate the utility of the proposed hypothesis testing method, we present results on various artificial and real-world benchmark datasets.

A. *Artificial Benchmarks with Periodic and Gaussian Distributions*

Our proposed method can be used for testing the homogeneity of structured data, which is an advantage over traditional two-sample tests. We artificially generated distributions from Locally Compact Abelian Groups (periodic data) and applied our RMMD-test to decide whether the samples come from the same distributions or not. Suppose the first sample is drawn from a uniform distribution *P* on the unit interval. The other sample is drawn from a perturbed uniform distribution Q_ω with density $1+\sin(\omega x)$. For higher perturbation frequencies ω it becomes harder to discriminate Q_ω from *P*. Since the distributions have a periodic nature, we use a characteristic kernel tailored to the periodic domain, $k(x,y) = \cosh(\pi - (x-y)_{mod 2\pi})$. For 200 samples from each distribution, the type II error is computed by comparing the prediction to the ground truth over 1000 repetition. We average the results over 10 runs. The significance level is set to α = *0.05*. We perform the same experiment with MMD, KFDA, KS and Hall. The powers of the homogeneity test for comparing P and Q_6 with the above mentioned methods are reported in Table 1 as Periodic1. The best power is achieved by RMMD, and as expected, the results of kernel methods are better than traditional ones.

Since the selection of the kernel is a critical choice in kernel-based methods, we also investigated the usage of a

TABLE 1. THE POWER OBTAINED ON THE PERIODIC DATA, THE GAUSSIAN, THE MNIST, COVERTYPE, AND FLARE SOLAR DATASETS, BY APPLYING RMMD WITH $\kappa = 0.8$ FOR THE PERIODIC DATA AND $\kappa = 1$ FOR THE OTHERS, AND KFDA WITH $\gamma = 10^{-1}$.

	RMMD	***KFDA***	***MMD***	***KS***	***Hall***
Periodic1	**0.40±0.02**	0.24±0.01	0.23±0.02	0.11±0.02	0.19±0.04
Periodic2	**0.83±0.03**	0.66±0.05	0.56±0.05	0.11±0.02	0.19±0.04
Gaussian	**1.00**	0.89±0.03	0.88±0.03	0.04±0.02	**1.00**
MNIST	**0.99±0.01**	0.97±0.01	0.95±0.01	0.12±0.04	0.77±0.04
Covertype	**1.00**	**1.00**	**1.00**	0.98±0.02	0.00
Flare-Solar	**0.93**	0.91	0.89	0.00	0.00
CIFAR10	**0.99±0.01**	**0.99±0.01**	**0.99±0.01**	0.64±0.07	0.00

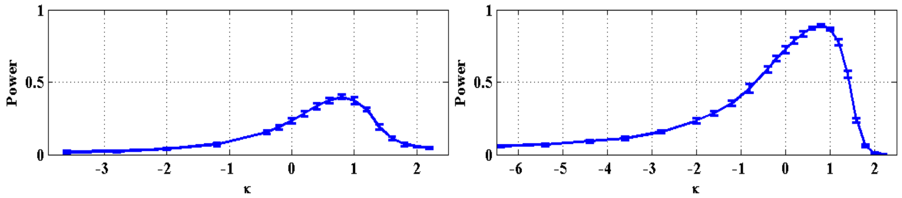

Figure 1. Effect of κ on the power of the test. The alternatives are Q_6 in the left and Q_4 in the right figure.

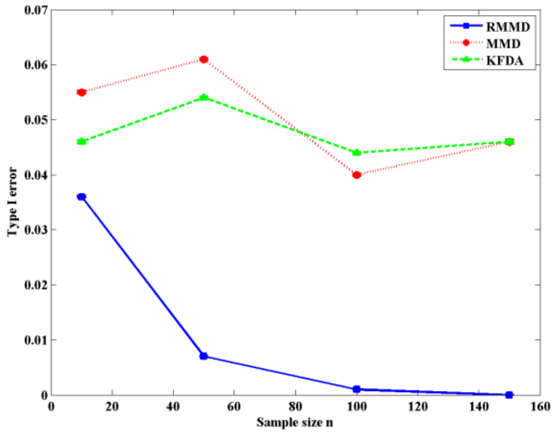

Figure 2. Type I error changed based on different sample size *n*.

different kernel and replaced the previous kernel with $k(x,y) = -\log(1 - 2\theta\cos(x-y) + \theta^2)$, where θ is a hyperparameter. We report the best results achieved by $\theta = 0.9$ as Periodic2 in Table 1. The reader is referred to [6], [22] for a detailed study on these kernels.

We also report the results on the toy problem of comparing two *25*-dimensional Gaussian distributions with *250* samples, both with zero mean vector but with covariance matrix *1.5 I* and *1.8 I*, respectively. This dataset is referred as Gaussian in Table 1.

An investigation of the effect of kernel selection and tuning parameters [23] showed that best results for MMD can be achieved by those kernels and parameters that obtain supreme value for MMD. Our reported results agree. The results of kernel-based test statistics (RMMD, KFDA, and MMD) are improved by kernel justification and parameter tuning, and in all cases RMMD outperform KFDA and MMD. For instance, the result of periodic kernel with tuned hyper-parameter θ is better than the one of the first periodic kernel without hyper-parameter (reported in Table 1 as Periodic2 and Periodic1, respectively). For Gaussian kernel-processed datasets, the median distance between data points provided the best results. We used the 5-fold cross validation procedure to tune the parameters in our experiment.

The effect of changing κ on the power is simulated in two tests: first, by testing the similarity between the uniform distribution and Q_4, and second with Q_6. In both cases, the best power is obtained for $\kappa = 0.8$. The results slightly differ from the theoretical value $\kappa = 1$) because of the relatively small sample sizes ($n_1 = n_2 = 200$) used for the tests. For samples with larger sizes we obtained maximal power with $\kappa = 1$. The results are depicted in Figure 1.

To assure that the statistical test is not aggressive for rejecting the null hypothesis, we reported the results of type I error for RMMD, KFDA, and MMD with different sample sizes in Figure 2. Both samples are supposed to be drawn from Q_6. We used Gaussian kernel with a variance equals to medium distance of data points. The results were averaged over 100 runs and the confidence interval obtained by 10 replicates (notice that the intervals are not visible in Figure 2 since their magnitude is less than 0.001). RMMD obtains zero type I error with smaller sample sizes, and the results of KFDA and MMD are comparable.

To assess the power control of the test statistics we also compared, $e^B_{MMD,RMMD}$, $e^B_{MMD,KFDA}$ and $e^B_{KFDA,RMMD}$ under H_1 when P is a uniform distribution and the alternative is Q_6. We obtained smaller ARE for RMMD rather than for KFDA and MMD. This means RMMD gives higher power with fewer samples. Table 2 shows the results, averaged over 1000 runs, for periodic data (Periodic1 and Periodic2). Figure 3 depicts

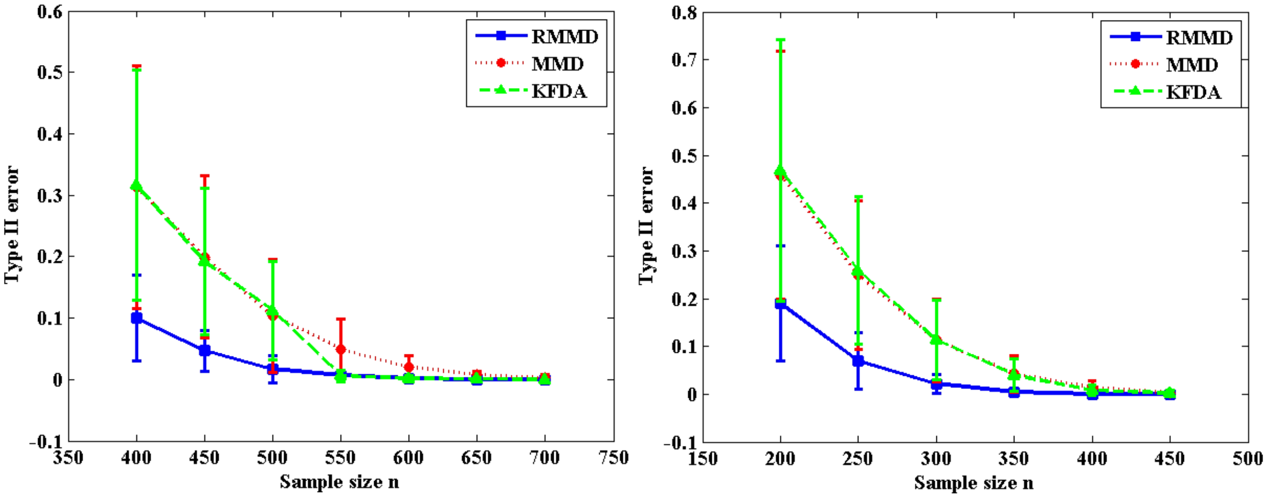

Figure 3. Type II error change based on different sample size n. On the left, the results with Periodic kernel 1 and on the right, the results with Periodic kernel 2.

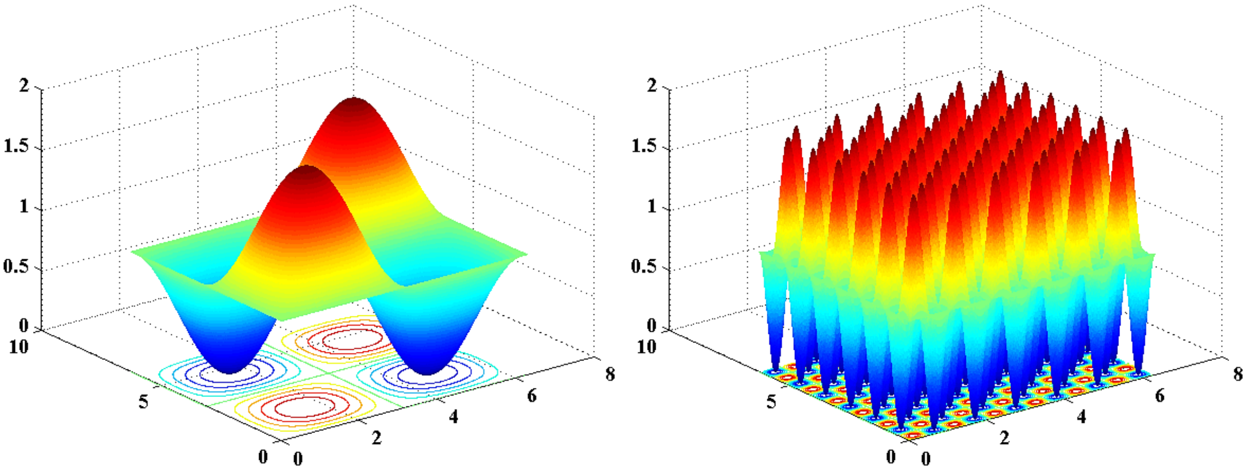

Figure 4. The probability density function of Puni1 with ω = 1 on the left and the probability density function of Puni6 with ω = 6 on the right. As ω increases the probability density function looks more similar to the uniform distribution and the discrimination of P and Q_ω becomes more difficult for the test statistics.

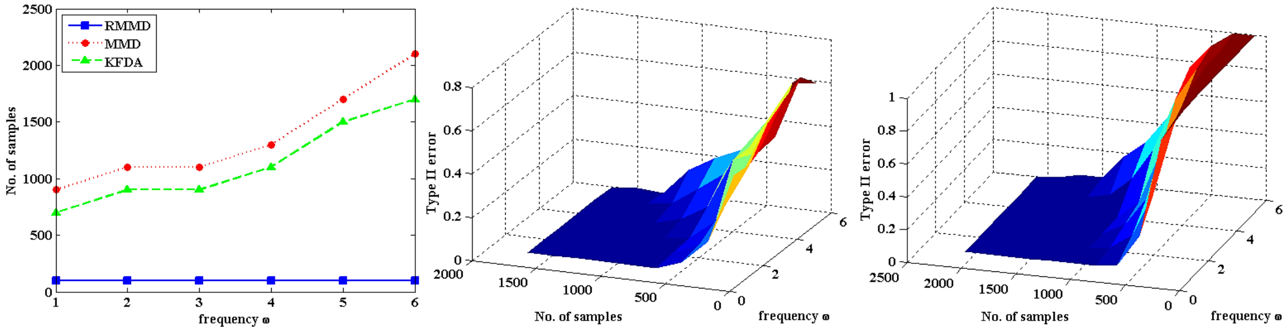

Figure 5. On the left, different sample sizes *n* for different frequencies ω are shown. The type II error changes based on different sample sizes *n* and different frequencies ω, in the middle for the KFDA-based test, and on the right for the MMD-based test.

the detailed results of the type II error for RMMD, MMD, and KFDA based on different sample sizes *n*. AREs are also calculated for more complex tasks. Consider the first sample is drawn from a uniform distribution P on the unit area. The other sample is drawn from the perturbed uniform distribution

of ω, the discrimination of Q_ω from P becomes harder (Figure 4). The range of ω changes between 1 to 6. We call these problems Puni1 to Puni6, respectively. The best results for all statistical kernel-based methods are achieved by using a characteristic kernel tailored to the periodic domain, $k(x,y) = \prod_{i=1}^{2} 1/(1-2\theta\cos(x_i - y_i) + \theta^2)$, with $\theta = 0.9$ tuned using the 5-fold cross validation procedure. The results reported in Table 2 show much smaller values of ARE for RMMD rather than for KFDA and MMD. Figure 5 shows the detailed results of type II error for RMMD, MMD, and KFDA based on different sample sizes n and different frequencies ω. As displayed in Figure 5, RMMD obtains the robust result of zero type II error for 100 samples over all different frequencies. Instead KFDA and MMD need much larger samples for the more difficult cases with larger ω to obtain a power of one.

TABLE I. THE ARE OBTAINED ON THE PERIODIC DATA, BY APPLYING RMMD WITH $\kappa = 1$, AND $\theta = 0.9$ IN PERIODIC KERNELS, AND KFDA WITH $\gamma = 10^{-1}$.

	$e^{B}_{MMD,RMMD}$	$e^{B}_{MMD,KFDA}$	$e^{B}_{KFDA,RMMD}$
Periodic1	**0.71**	0.75	0.93
Periodic2	**0.75**	1	**0.75**
Puni1	**0.11**	0.78	0.14
Puni2	**0.09**	0.82	0.11
Puni3	**0.09**	0.82	0.11
Puni4	**0.08**	0.85	0.09
Puni5	**0.07**	0.88	0.06
Puni6	**0.05**	0.81	0.06

B. *Performance on Benchmark Datasets*

Moving from synthetic data to standard benchmarks, we tested our method on three datasets: 1) the MNIST dataset of handwritten digits (LibSVM library: 10 classes, 500 sample size, and 784 dimensions); 2) the Covertype dataset of forest cover types (LibSVM library: 7 classes, 200 sample size, and 54 dimensions); 3) the Flare-Solar dataset (mldata.org: 2 classes, 50 sample size, 10 dimensions); 4) the CIFAR10 dataset of tiny object images (10 classes, 200 sample size, 3072 dimensions (raw features)). We compare the performance of RMMD with $\kappa = 1$, KFDA with $\gamma = 10^{-1}$and MMD, using the pairwise approach and testing for differences between the distributions of the classes, see Table 1. We average the results over 10 runs. The family wide level is set to $\alpha = 0.05$ (resulting in $\bar{\alpha} = 0.0011$, $\bar{\alpha} = 0.0024$, $\bar{\alpha} = 0.05$ and $\bar{\alpha} = 0.0011$ for each individual comparison for MNIST, Covertype, Flare-Solar, and CIFAR10 datasets, respectively). The RMMD-based test achieves higher power than the other methods (see Table 1).

C. *Electroencephalography Data*

We recorded EEG from four subjects performing a visual task. A checkerboard was presented in the subject's left visual field. We refer to [24] for details on data collection and preprocessing. In our learning task, for each subject we have 64 signal distributions assigned to 64 electrodes. The data contain 360 instances of a 200 dimensional feature vector for each distribution. The goal of hypothesis testing is to disambiguate signals recorded from electrodes corresponding to early visual cortex from the rest. This is difficult because of low signal-to-noise ratio and the similarity of the patterns of all electrodes. Moreover, the high number of electrodes makes this experiment a good candidate to assess the multiple comparison part of our method. In the one-vs-all approach the normalized distribution of each electrode is compared to the normalized combined distribution of the other 63 electrodes. RMMD with $\kappa = 1$ with Gaussian kernel is used as our hypothesis test. The parameter σ of the Gaussian kernel is set to the median distance of data points. The results of our hypothesis test reject the null hypothesis and confirm the dissimilarity of distributions in 63 electrodes. The results of the pairwise approach with RMMD and MMD are depicted in Figure 6.

Neuroscientists usually subjectively assess the results obtained from imaging techniques and inferred from machine learning. For instance, in the current experiment the expectation is that electrodes in region A_1 (see Figure 7) are categorized together by means of EEG imaging techniques and multiple comparisons. But electrodes of other area (such as A_2 and A_3, see Figure 7) can be confused as belonging to A_1 due to the high noise. Figure 7 describes the categorization of the electrodes.

We assess our results quantitatively by means of False Discovery Rates (FDR), using the following FDRs to compare the results of RMMD to those of MMD:

$$FDR_0 = \frac{\text{no.of electrodes categorized for the visual task in } A_2 \cup A_3 \cup B}{U},$$

$$FDR_1 = \frac{\text{no.of electrodes categorized for the visual task in } A_3 \cup B}{U},$$

$$FDR_2 = \frac{\text{no.of electrodes categorized for the visual task in } B}{U},$$

where U is the total number of electrodes categorized for the task. The results are depicted in Figure 7. RMMD obtained more robust and better results than MMD with smaller FDRs.

I. CONCLUSION

Our novel regularized maximum mean discrepancy (RMMD) is a kernel-based test statistic generalizing the MMD test. We proved that RMMD overpowers MMD and KFDA; power consistency is obtained with higher rate. Power control makes RMMD a good hypothesis test for multiple comparisons, especially for the crucial case of small sample sizes. In contrast to KFDA and MMD, the convergence of RMMD-based test statistics to the normal distribution under null and alternative hypotheses yields fast and straightforward RMMD estimates. Experiments with goldstandard benchmarks (CIFAR10, MNIST, Covertype and Flare-Solar dataset) and with EEG data yield state of the art results.

II. ACKNOWLEDGMENT

This work was partially funded by the State representation in reward based learning from spiking neuron models to psychophysics (SINERGIA: CRSIKO_122697/1) grant, and kernel-based dimensionality reduction (SNSF: P1TIP2_148352) grant.

REFERENCES

[1] A. Gretton, B.K. Borgwadt, M. Rasch, B. Schölkopf, and A. Smola, " A kernel method for the two-sample-problem," In Advances in Neural

Information Processing Systems. Schölkopf, B., J. Platt, T. Hofmann (eds.), MIT Press, Cambridge, MA, USA, vol.19, pp. 513-520, 2007.

[6] K. Fukumizu, B. Sriperumbudur, A. Gretton, and B. Schölkopf, " Characteristic kernels on groups and semigroups," In Advances in Neural Information Processing Systems. Koller, D., D. Schuurmans, Y. Bengio, L. Bottou (eds.), vol. 21, pp. 473-480, Curran, Red Hook, NY,

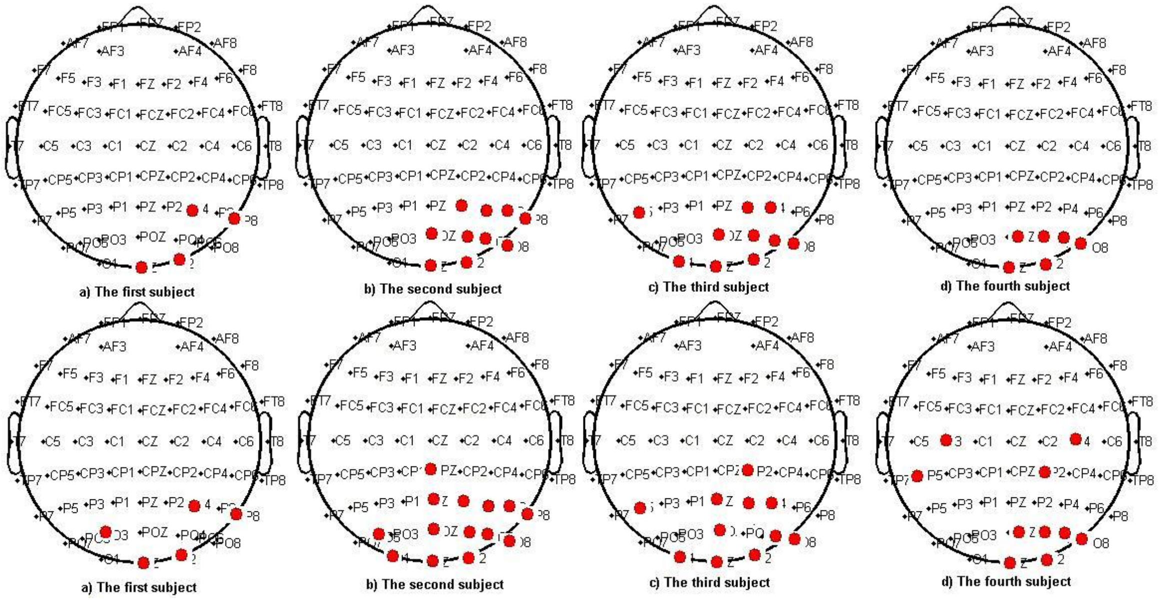
a) The first subject
b) The second subject
c) The third subject
d) The fourth subject
a) The first subject
b) The second subject
c) The third subject
d) The fourth subject

Figure 6. The results of RMMD and MMD as hypothesis tests on the EEG data recorded from 64 electrodes per subject in the top row and the bottom row, respectively. Categorized electrodes recognized by the two methods as related to the visual task are colored.

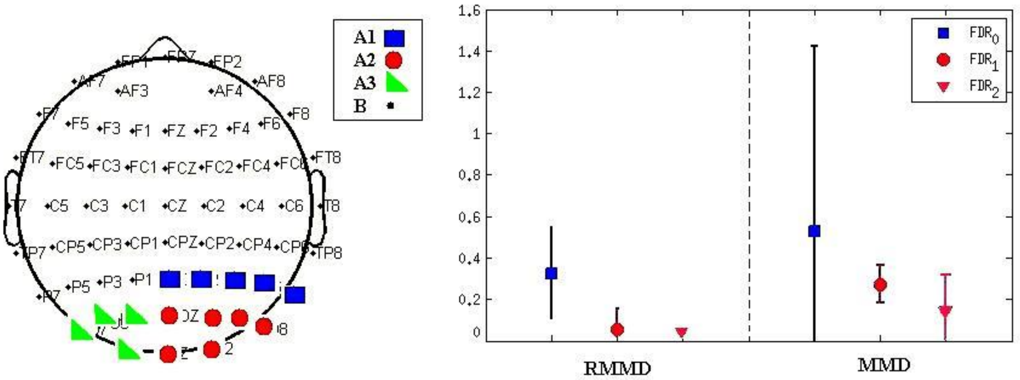

Figure 7. The reference image of the EEG electrodes is shown on the left. We categorized electrodes into four groups as follows: A1, the electrodes corresponding to visual cortex in the region of interest, A2, the peripheral electrodes that can be wrongly detected due to noise, A3, the electrodes in the left visual cortex often detected due to noise or interrelation between brain areas, and B, all the remaining electrodes. On the right, the results of RMMD and MMD are quantitatively compared based of the FDRs defined in the text. The smallest and most robust FDRs are obtained by RMMD.

[2] Z. Harchaoui, F.R. Bach, and E. Moulines, "Testing for homogeneity with kernel fisher discriminant analysis," In Advances in Neural Information Processing Systems, vol.20, pp. 609-616, 2008.

[3] J. Freidman, and L. Rafsky,"Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," The Annals of statistics, vol.7, pp. 697-717, 1979.

[4] Y. Rubner, C. Tomasi, and L.J. Gubias, "The earth movers distance as a metric for image retrieval," International Journal of Computer Vision, vol.40(2), pp. 99-121, 2000.

[5] A. Gretton, K. Fukumizu, C.H. Teo, L. Song, B. Schölkopf, and A. Smola,"A Kernel Statistical Test of Independence," In Advances in Neural Information Processing System. Platt, J. C.,D. Koller, Y. Singer, S. Roweis (eds.), vol.20, pp. 585-592, MIT Press, Cambridge, MA, USA, 2008.

[7] K. Borgwardt, A. Gretton, M. Rash, H.P. Kriegel, B. Schölkopf, and A.J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," Bioinformatics, vol.22(14), pp. 49-57, 2006.

[8] B. Walsh, "Multiple Comparisons: Bonferroni Corrections and False Discovery Rates," Lecture Notes for EEB, vol.581, 2004.

[9] A.J. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert Space Embedding for Distributions," In Algorithmic Learning Theory. Hutter, M., R. A. Servedio, E. Takimoto (eds.), vol.18, pp. 13-31, Springer, Berlin, Germany, 2007.

[10] B.K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckeriet, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," In proceedings of the 21st Annual Conference on Learning Theory. Servedio, R. A., T. Zhang (eds.), pp. 111-122, Omnipress, Madison, WI, USA, 2008.

[11] K. Fukumizu, F.R. Bach, and M.I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert Spaces," Journal of Machine Learning, vol.5, pp. 73-99, 2004.

[12] D. Erdogmus, and J.C. Principe, "From Linear adaptive filtering to nonlinear information processing," IEEE Signal Processing Magazine, vol.23(6), pp. 14-23, 2006.

[13] A. Renyi, "On measures of entropy and information," In Proc. Fourth Berkeley Symp, Math, Statistics and Probability, pp. 547-561, 1960.

[14] B. Fuglede, and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," In Proceedings of the International Symposium on Information Theory (ISIT), 2004.

[15] K. Fukumizu, F.R. Bach, A. Gretton, "Statistical consistency of kernel canonical correlation analysis," Journal of Machine Learning, vol.8, pp. 361-383, 2007.

[16] W. Hoeffding,"A class of statistics with asymptotically normal distributions," The Annals of mathematical Statistics, vol.19(3), pp. 293-325, 1948.

[17] B. Serfling, "Approximation Theorems of Mathematical Statistics," Wiley, New York, 1980.

[18] Z. Harchaoui, F. Vallet, and A. Lung-Yut-Fong, "A regularized kernel0based approach to unsupervised audio segmentation", In proceedings of International Conference on Acoustics, Speech, and Signal processing (ICASSP), 2009.

[19] A. Gretton, K. Fukumizu, Z. Harchaoui, and B. Sriperumbudur, "A Fast, Consistent Kernel Two-Sample Test", In Advances in Neural Information Processing Systems, vol.22, 2009.

[20] Y. Nikitin, "Asymptotic efficiency of non-parametric tests", Cambridge University Press, 1995.

[21] P. Hall, and N. Tajvidi, "Permutation tests for equality of distributions in high-dimensional settings," Biometrika, vol.89, pp. 359-374, 2002.

[22] S. Danafar, A. Gretton, and J. Schmidhuber, "Characteristic kernels on structured domains excel in robotics and human action recognition," ECML: Machine Learning and Knowledge Discovery in Databases, pp. 264-279, LCNS, Springer, 2010.

[23] B.K. Sriperumbudur, K. Fukumizu, A. Gretton, G. Lanckeriet, and B. Schölkopf, "Kernel choice and classifiability for RKHS embeddings of probability distributions," In Advances in Neural Information Processing Systems 22, 2009.

[24] K. Whittingstall, W. Dough, S. Matthias, and S. Gerhard, "Correspondence of visual evoked potentials with fMRI signals in human visual cortex, " Brain Topogr, Servedio, R. A., T. Zhang (eds.), vol.21, pp. 86-92, Omnipress, Madison, WI, USA, 2008.