Kernel Constrained Covariance for Dependence Measurement

Arthur Gretton*, Alexander Smola†, Olivier Bousquet*, Ralf Herbrich‡, Andrei Belitski*, Mark Augath*, Yusuke Murayama*, Jon Pauls*, Bernhard Schölkopf*, & Nikos Logothetis*

first.last@tuebingen.mpg.de, Alex.Smola@anu.edu.au, rherb@microsoft.com

* MPI for Biological Cybernetics, Tübingen, Germany

† NICTA, Canberra, Australia; ‡ Microsoft Research, Cambridge, UK

Abstract

We discuss reproducing kernel Hilbert space (RKHS)-based measures of statistical dependence, with emphasis on constrained covariance (COCO), a novel criterion to test dependence of random variables. We show that COCO is a test for independence if and only if the associated RKHSs are universal. That said, *no* independence test exists that can distinguish dependent and independent random variables in all circumstances. Dependent random variables can result in a COCO which is arbitrarily close to zero when the source densities are highly non-smooth. All current kernel-based independence tests share this behaviour. We demonstrate exponential convergence between the population and empirical COCO. Finally, we use COCO as a measure of joint neural activity between voxels in MRI recordings of the macaque monkey, and compare the results to the mutual information and the correlation. We also show the effect of removing breathing artefacts from the MRI recording.

1 Introduction

Tests to determine the dependence or independence of random variables are well established in statistical analysis. Some approaches require density estimation as an intermediate step ([13] is a classic study); while others assume a parametric model of how the variables were obtained from independent random variables, as in blind source separation [12].

In this paper we propose a non-parametric independence criterion, which relies on the fact that the random variables[1] x, y are independent if and only if

$$\mathbf{E}_{\mathsf{x}}[f(\mathsf{x})]\mathbf{E}_{\mathsf{y}}[g(\mathsf{y})] = \mathbf{E}_{\mathsf{x},\mathsf{y}}[f(\mathsf{x})g(\mathsf{y})]. \tag{1.1}$$

for bounded, continuous functions f, g (see for instance [14, 19]). The proposed criterion works by maximising the discrepancy between the empirical estimates of the LHS and RHS of (1.1) over pre-specified function classes $f \in \mathcal{F}$ and $g \in \mathcal{G}$, and comparing the discrepancy to the amount of deviation that can be expected from the fact that we are dealing with empirical estimates rather than expectations. We call our criterion the constrained covariance (COCO).[2]

The results presented here build on recent work published on the subject of kernel based dependence measures. In particular, the canonical correlation between functions in a reproducing kernel Hilbert space (KCC), defined in [1] for a variety of kernels and in [15] for splines, can be used as a test of independence. Indeed, in the case of Gaussian kernels, Bach and Jordan show the KCC to be zero if and only if its two arguments are statistically independent. In Section 3, we characterise *all* reproducing kernel Hilbert spaces (RKHSs) for which this property holds (both for COCO and KCC): these are required to be *universal* (the RKHS must be dense in the space of continuous functions [22]). Specifically, the Gaussian and Laplace kernels are universal, as are many exponential-based kernels; polynomial kernels, however, are not universal.

We next demonstrate in Section 4 that for a fixed-size, finite sample of *dependent* random variables, there exists no test that can reliably detect that the random variables are dependent. To clarify how this might affect our criterion, we prove that the population COCO can be made arbitrarily small when certain smoothness assumptions on the density are violated, which makes it difficult to detect dependence on the basis of a finite sample. This is also true of other related kernel dependence measures, including the kernel mutual information (KMI) in [9], and kernel generalised variance (KGV) in [1], both of which were shown in [9] to be upper bounds near independence on the Parzen window estimate of the mutual information. Thus, as in all dependence tests, any inference made is subject to certain assumptions about the underlying generative process - the present work describes these assumptions explicitly for the first time, in the case of kernel-based tests.

[1] We write random variables *sans serif*.

[2] In [9], this was called the kernel covariance (KC).

Next, we give two bounds, based on Rademacher averages, which describe *exponential* convergence between the population and empirical COCO. The first assures us that the population COCO is small when the empirical COCO is small; the second shows that the population COCO is large when the empirical COCO is large (both statements apply with high probability). These results are very interesting, in that they illustrate a broader phenomenon: slow learning rates do not occur in dependence testing, even though they are unavoidable in regression and classification [6, Ch. 7]. This might appear surprising in the specific case of COCO, since this criterion is optimised in the course of kernelised PLS regression (assuming a kernelised output space: see the discussion in [2]). Another important consequence of the bounds is that *any* dependence between the random variables will be detected rapidly *as the sample size increases*, even though perfect dependence detection is impossible for fixed sample size.

Finally, we describe a neuroscience application where our method can be used. A number of groups (e.g. [3]) have begun examining the interactions between neural systems using fMRI in humans. The recent study of BOLD fMRI in the macaque monkey using high field (4.7T & 7T) scanners [17, 16] has resulted in substantial increases in spatial and temporal resolution, when measuring brain activity patterns resulting from various stimuli. In Section 6, we apply COCO to these high resolution data so as to detect dependence between BOLD responses within the visual cortex. In using COCO to detect regions of high dependence, we follow [8], who maximise a kernel-based dependence measure (in their case, the KGV) as a means of variable selection. We also investigate how the measured dependence changes with the removal of breathing artefacts, which is feasible due to the high temporal resolution of our measurements.

2 Definitions and Background

Before presenting our main results, we begin our discussion with some relevant definitions and background theory, covering both classical independence criteria and RKHSs.[3] Let $(\Omega, \mathcal{A}, \mathbf{P}_{\mathsf{x},\mathsf{y}})$ be a probability space. Consider random variables $\mathsf{x} : (\Omega, \mathcal{A}) \to (U, \mathcal{U})$ and $\mathsf{y} : (\Omega, \mathcal{A}) \to (V, \mathcal{V})$, where U and V are complete metric spaces, and $\mathcal{U}$ and $\mathcal{V}$ their respective Borel σ-algebras. The covariance between x and y is defined as follows.

Definition 1 (Covariance). *The covariance of two random variables* x, y *is given as*

$$\mathrm{cov}(\mathsf{x}, \mathsf{y}) := \mathbf{E}_{\mathsf{x},\mathsf{y}}[\mathsf{x}\mathsf{y}] - \mathbf{E}_{\mathsf{x}}[\mathsf{x}]\mathbf{E}_{\mathsf{y}}[\mathsf{y}]. \tag{2.1}$$

For our purposes, the notion of independence of random variables is best expressed using the following characterisation:

[3] This exposition is necessarily dense: see [14] and [21] for more detail.

Theorem 1 (Independence [14]). *The random variables* x *and* y *are independent if and only if* $\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) = 0$ *for each pair* (f, g) *of bounded, continuous functions.*

This theorem suggests the following definition as an independence test.

Definition 2 (Constrained covariance). Given function classes $\mathcal{F}, \mathcal{G}$ containing subspaces $F \in \mathcal{F}$ and $G \in \mathcal{G}$, we define the *constrained covariance* as

$$\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) := \sup_{f \in F, g \in G} [\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y}))]. \tag{2.2}$$

(if F and G are unit balls in their respective spaces, then this is just the norm of the covariance operator mapping $\mathcal{G}$ to $\mathcal{F}$: see [8] and references therein). Given n independent observations $\boldsymbol{z} := ((x_1, y_1), \ldots, (x_n, y_n)) \subset (\mathcal{X} \times \mathcal{Y})^n$, its empirical estimate is defined as

$$\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G) := \sup_{f \in F, g \in G} \left[\frac{1}{n} \sum_{i=1}^{n} f(x_i) g(y_i) - \frac{1}{n^2} \sum_{i=1}^{n} f(x_i) \sum_{j=1}^{n} g(y_j) \right].$$

It follows from Theorem 1 that if $\mathcal{F}, \mathcal{G}$ are the sets of continuous functions bounded by 1 we have $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{F}, \mathcal{G}) = 0$ if and only if x and y are independent.[4] In other words, COCO and $\mathrm{COCO}_{\mathrm{emp}}$ are criteria which can be tested *directly* without the need for an intermediate density estimator (in general, the distributions may not even have densities). It is also clear, however, that unless F, G are restricted in further ways, $\mathrm{COCO}_{\mathrm{emp}}$ will always be large, due to the rich choice of functions available. A *non-trivial dependence measure* is thus obtained using function classes that do not give an everywhere-zero empirical average, yet which still guarantee that COCO is zero if and only if its arguments are independent. A tradeoff between the restrictiveness of these function classes and the convergence of $\mathrm{COCO}_{\mathrm{emp}}$ to COCO can be accomplished using standard tools from uniform convergence theory (see Section 5).

It turns out (Section 3) that unit-radius balls in universal reproducing kernel Hilbert spaces constitute function classes that yield non-trivial dependence estimates. To demonstrate this, we will use certain properties of these spaces [20]. A reproducing kernel Hilbert space is a Hilbert space $\mathcal{F}$ for which at each $x \in \mathcal{X}$, the point evaluation functional, $\delta_x : \mathcal{F} \to \mathbb{R}$, which maps $f \in \mathcal{F}$ to $f(x) \in \mathbb{R}$, is continuous. To each reproducing kernel Hilbert space, there corresponds a unique positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (the reproducing kernel), which constitutes the inner product on this space: this is guaranteed by the Moore-Aronszajn theorem.

[4] Here we set $F = \mathcal{F}$ and $G = \mathcal{G}$.

In RKHSs the representer theorem [21] holds, stating that the solution of an optimisation problem, dependent only on the function evaluations on a set of observations and on RKHS norms, lies in the span of the kernel functions evaluated on the observations. This property is next used to specify an easily computed expression for $\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G)$ where F and G are respectively unit balls in the reproducing kernel Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$. The proof may be found in [9].

Lemma 1 (Value of $\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G)$). *Denote by $\mathcal{F}, \mathcal{G}$ RKHSs on the domains $\mathcal{X}$ and $\mathcal{Y}$ respectively and let F, G be the unit balls in the corresponding RKHS. Then*

$$\mathrm{COCO}_{\mathrm{emp}}(\boldsymbol{z}; F, G) = \frac{1}{n}\sqrt{\|\bar{K}^f \bar{K}^g\|_2} \tag{2.3}$$

where $\bar{K}^f$ is the matrix obtained by the projection $\bar{K}^f = PK^fP$ with projection operator $P_{ij} = \delta_{ij} - \frac{1}{n}$ and Gram matrix $K^f_{ij} = k_f(x_i, x_j)$. $\bar{K}^g$ is defined by analogy using the kernel of $\mathcal{G}$ (which might be different from that of $\mathcal{F}$).

A second theorem which will be crucial in our proofs is Mercer's theorem, which provides a decomposition of the kernel into eigenfunctions and eigenvalues.

Theorem 3 (Mercer's theorem). *Let $k(\cdot,\cdot) \in L_\infty(\mathcal{X}^2)$ be a symmetric real valued function with an associated positive definite integral operator with normalised orthogonal eigenfunctions $\varphi_p \in L_2(\mathcal{X})$, sorted such that the associated eigenvalues $\tilde{k}_p$ do not increase. Then for almost all $x_i \in \mathcal{X}$ and $x_j \in \mathcal{X}$, the series*

$$k(x_i, x_j) := \sum_{p=1}^{\infty} \tilde{k}_p \varphi_p(x_i)\varphi_p(x_j)$$

converges absolutely and uniformly. In addition, the sum $\sum_{i=1}^{p} |\tilde{k}_i|$ converges as $p \to \infty$.

Finally, we give kernel-dependent decay rates for the coefficients used to expand functions in $\mathcal{F}$ in terms of the set of basis functions $\{\varphi_i(\cdot)\}$ from Mercer's theorem.

Lemma 4 (Rate of decay of expansion coefficients). *Let $f \in \mathcal{F}$, where $f(x) := \sum_{i=1}^{\infty} \tilde{f}_i \varphi_i(x)$. Then as long as $(\tilde{k}_i)^{-1}$ increases super-linearly with i, $(|\tilde{f}_i|) \in \ell_1$ and there exists an $l_0 \in \mathbb{N}$ such that for all $\epsilon > 0$ and all $l > l_0$, $|\tilde{f}_l| < \epsilon$.*

Proof. This holds since for any $f \in \mathcal{F}$, $\|f\|^2_{\mathcal{F}} := \sum_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i\right)^{-1} < \infty$. □

The super-linearity requirement in Lemma 4 is satisfied by many kernels, including the Gaussian (for which the $((\tilde{k}_m)^{-1})$ increase as $\exp(m^2)$); see [21]. We assume hereafter that our kernel satisfies the requirements of Lemma 4.

3 A Test for Independence

We now characterise the class of kernels for which COCO is a non-trivial test of dependence. The main result is given in Theorem 6, in which we demonstrate that COCO constitutes such a test when $\mathcal{F}$ and $\mathcal{G}$ are RKHSs with a universal kernel [22].

Definition 5 (Universal kernel). A *continuous kernel $k(\cdot,\cdot)$* on a compact metric space $(\mathcal{X}, d)$ is called universal if and only if the RKHS $\mathcal{F}$ induced by the kernel is dense in $C(\mathcal{X})$ with respect to the topology induced by the infinity norm $\|f - g\|_\infty$.

For instance, [22] shows the following two kernels are universal on compact subsets of $\mathbb{R}^d$:

$$k(x, x') = \exp\left(-\lambda\|x - x'\|^2\right) \text{ and}$$
$$k(x, x') = \exp\left(-\lambda\|x - x'\|\right) \text{ for } \lambda > 0.$$

We now state our main result for this section.

Theorem 6 (COCO is only zero at independence for universal kernels). *Denote by $\mathcal{F}, \mathcal{G}$ RKHSs with universal kernels k_f, k_g on the compact domains $\mathcal{X}$ and $\mathcal{Y}$ respectively and let F, G be the unit balls in the corresponding RKHSs. We assume without loss of generality that $\|f\|_\infty \leq 1$ and $\|g\|_\infty \leq 1$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Then $\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$ if and only if x, y are independent.*

Proof. It is clear that $\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ is zero if x and y are independent. We prove the converse by contradiction, using the starting assumptions $\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; B(\mathcal{X}), B(\mathcal{Y})) = c$ for some $c > 0$ (here $B(\mathcal{X})$ denotes the subset of $C(\mathcal{X})$ of continuous functions bounded by 1 in the $L_\infty(\mathcal{X})$, and $B(\mathcal{Y})$ is defined in an analogous manner) and $\mathrm{COCO}(\mathsf{P}_{\mathsf{x},\mathsf{y}}; F, G) = 0$. There exist two sequences of functions $f_n \in C(\mathcal{X})$ and $g_n \in C(\mathcal{Y})$, satisfying $\|f_n\|_\infty \leq 1, \|g_n\|_\infty \leq 1$, for which

$$\lim_{n\to\infty} \mathrm{cov}(f_n(\mathsf{x}), g_n(\mathsf{y})) = c.$$

More to the point, there exists an n^* for which $\mathrm{cov}(f_{n^*}, (\mathsf{x})g_{n^*}(\mathsf{y})) \geq c/2$. We know that $\mathcal{F}$ and $\mathcal{G}$ are respectively dense in $C(\mathcal{X})$ and $C(\mathcal{Y})$: this means that for all $1/3 > \epsilon > 0$, we can find some $f^* \in \mathcal{F}$ (and an analogous $g^* \in \mathcal{G}$) satisfying $\|f^* - f_{n^*}\|_\infty < \epsilon = \frac{c}{24}$. Writing as $\tilde{f}(\mathsf{x}) := f^*(\mathsf{x}) - f_{n^*}(\mathsf{x}) + f_{n^*}(\mathsf{x})$ (with an analogous $\tilde{g}(\mathsf{y})$ definition), we obtain

$$\begin{aligned}
&\mathrm{cov}(f^*(\mathsf{x}), g^*(\mathsf{y})) \\
&= \mathsf{E}_{\mathsf{x},\mathsf{y}}\left[\tilde{f}(\mathsf{x})\tilde{g}(\mathsf{y})\right] - \mathsf{E}_{\mathsf{x}}(\tilde{f}(\mathsf{x}))\mathsf{E}_{\mathsf{y}}(\tilde{g}(\mathsf{y})) \\
&\geq \mathrm{cov}(f_{n^*}(\mathsf{x}), g_{n^*}(\mathsf{y})) - 2\epsilon\left|\mathsf{E}_{\mathsf{x}}\left(f_{n^*}(\mathsf{x})\right)\right| \\
&\qquad - 2\epsilon\left|\mathsf{E}_{\mathsf{y}}\left(g_{n^*}(\mathsf{y})\right)\right| - 2\epsilon^2 \\
&\geq \frac{c}{2} - 6\frac{c}{24} = \frac{c}{4} > 0.
\end{aligned}$$

This contradicts the assumption that $\mathrm{cov}(f^*(\mathsf{x}), g^*(\mathsf{y})) = 0$, and completes the proof. □

The kernel dependence tests (COCO, KMI, KGV, and KCC) are generalised in [9, 1] to a greater number of random variables, providing tests of pairwise independence.

4 Limitations of Independence Tests

4.1 General independence tests

In this section, we illustrate with a simple example that for a finite sample, there exists no test of independence which can reliably (i.e. with high probability) distinguish dependence from independence. This discussion is intended as a complement to the next section, where we explicitly construct dependent random variables which are difficult for the empirical COCO to distinguish from independence. We illustrate the case where $\mathcal{X}$ is countable, but our reasoning applies equally to continuous spaces.

We begin with some notation. Consider a set $\mathcal{P}$ of probability distributions $\mathbf{P}_{\mathsf{x}}$, where x contains m entries. The set $\mathcal{P}$ is split into two subsets: $\mathcal{P}_i$ contains distributions $\mathbf{P}_{\mathsf{x}}^{(i)}$ of mutually independent random variables $\mathbf{P}_{\mathsf{x}}^{(i)} = \prod_{j=1}^{m} \mathbf{P}_{\mathsf{x}_j}$, and $\mathcal{P}_d$ contains distributions $\mathbf{P}_{\mathsf{x}}^{(d)}$ of dependent random variables. We next introduce a test $\Delta(\boldsymbol{x})$, which takes a data set[5] $\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}}{}^{n}$, and returns

$$\Delta(\boldsymbol{x}) = 1 \ : \ \boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(d)}, \qquad \Delta(\boldsymbol{x}) = 0 \ : \ \boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}$$

Given that the test sees only a finite sample, it cannot determine with complete certainty whether the data are drawn from $\mathbf{P}_{\mathsf{x}^n}^{(d)}$ or $\mathbf{P}_{\mathsf{x}^n}^{(i)}$. We call Δ an α-test when

$$\sup_{\mathbf{P}_{\mathsf{x}}^{(i)} \in \mathcal{P}_i} \mathbf{E}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}} (\Delta(\boldsymbol{x}) = 1) \leq \alpha;$$

in other words α upper bounds the probability of a Type I error. Our theorem is as follows:

Theorem 7 (Universal limit on dependence tests). *For any α-test, some fixed $n \in \mathbb{N}$, and any $1 - \alpha > \epsilon > 0$, there exists $\mathbf{P}_{\mathsf{x}} \notin \mathcal{P}_i$ such that*

$$\mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}}{}^{n}} (\Delta(\boldsymbol{x}) = 0) \geq 1 - \alpha - \epsilon;$$

in other words, a dependence test with a low Type I error can have a severe Type II error.

Proof. We introduce a distribution $\mathbf{P}_{\mathsf{x}}^{(\gamma)} := \gamma \mathbf{P}_{\mathsf{x}}^{(i)} + (1-\gamma)\mathbf{P}_{\mathsf{x}}^{(d)}$, where $0 \leq \gamma < 1$. Clearly, random variables drawn from $\mathbf{P}_{\mathsf{x}}^{(\gamma)}$ are dependent. The probability of a Type II error for this mixture is then

$$\begin{aligned}
&\mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(\gamma)}} (\Delta(\boldsymbol{x}) = 0) && \overset{(a)}{=} \sum_{\boldsymbol{x}} \mathbf{P}_{\mathsf{x}^n}^{(\gamma)}(\boldsymbol{x}) \mathbb{I}_{\Delta(\boldsymbol{x})=0} \\
&= \sum_{\boldsymbol{x}} \prod_{k=1}^{n} \mathbf{P}_{\mathsf{x}}^{(\gamma)}(\mathbf{x}_k) \mathbb{I}_{\Delta(\boldsymbol{x})=0} && = \sum_{\boldsymbol{x}} \prod_{k=1}^{n} \gamma \mathbf{P}_{\mathsf{x}}^{(i)}(\mathbf{x}_k) \mathbb{I}_{\Delta(\boldsymbol{x})=0} \\
&= \gamma^n \mathbf{P}_{\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}^n}^{(i)}} (\Delta(\boldsymbol{x}) = 0) && = \gamma^n (1 - \alpha)
\end{aligned}$$

where the sum following (a) is over all possible draws of $\boldsymbol{x}$ from $\mathbf{P}_{\mathsf{x}^n}^{(\gamma)}$, and $\mathbb{I}_A$ is the indicator function for event A. Taking γ very close to 1 (i.e. making the dependent distribution very unlikely in the mixture) proves the theorem. □

4.2 Kernel independence tests

We prove the existence of a dependent probability distribution for which COCO is small, but with a large covariance between certain functions in $\mathcal{F}$ and $\mathcal{G}$; we then demonstrate that this also holds for the KCC, KMI, and KGV. Although the population COCO is *not* zero for this density, its small size will make this dependence hard to detect unless a large data sample is available. We illustrate this phenomenon by specifying a particular joint density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ (with distribution $\mathbf{P}_{\mathsf{x},\mathsf{y}}$) chosen such that $\mathrm{cov}(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y}))$ is large for some large l (meaning x, y have a non-trivial dependence), but $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G)$ is small. The intuition behind our argument is made clear by re-writing COCO for RKHSs as

$$\mathrm{COCO}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G\right) = \sup_{f \in \mathcal{F},\, g \in \mathcal{G}} \frac{\mathrm{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)}{\|f\|_{\mathcal{F}} \|g\|_{\mathcal{G}}}. \quad (4.1)$$

This will obviously be small when the RKHS norms in the denominator are much larger than the covariance in the numerator: we will see that this motivates our choice of density. More specifically, high order eigenfunctions of the kernel[6] ($\varphi_l(x)$ and $\varphi_l(y)$ for large l) have large RKHS norms, a fact widely exploited in regression as a roughness penalty [21]. Thus, if the high order eigenfunctions are prominent in $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ (i.e., for highly non-smooth densities), we expect COCO to be small even when there exists an l for which $\mathrm{cov}(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y}))$ is large.[7]

Theorem 8 (Dependent random variables can have small COCO). *There exists a density $\mathbf{f}_{\mathsf{x},\mathsf{y}}$ for which* $\mathrm{cov}\left(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y})\right) \geq \beta - \epsilon$ *for non-trivial β and arbitrarily small $\epsilon > 0$, yet for which $\mathrm{COCO}(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G) < \gamma$ for an arbitrarily small $\gamma > 0$.*

Proof. The proof is a sketch only: further detail is given in [10]. We begin by constructing a density for which $\mathrm{cov}\left(\varphi_l(\mathsf{x}), \varphi_l(\mathsf{y})\right) \geq \beta - \epsilon$. This is written

$$\mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y) = \alpha_l + \beta \varphi_l(x) \varphi_l(y) \quad (4.2)$$

where $\mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y) \geq 0$ and $\int \mathbf{f}_{\mathsf{x},\mathsf{y}}(x, y) dx dy = 1$. The first constraint requires $\alpha_l - \beta \min_{x,y}(\varphi_l(x)\varphi_l(y)) \geq 0$,

[5] We denote by $\boldsymbol{x} \sim \mathbf{P}_{\mathsf{x}}{}^{n}$ the drawing of n i.i.d. samples $\boldsymbol{x} := (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ from $\mathbf{P}_{\mathsf{x}}$.

[6] See Theorem 3 for a definition of the eigenfunctions. Note that the kernels in $\mathcal{F}$ and $\mathcal{G}$ may not be identical, and the eigenfunctions $\varphi_i(x)$ and $\varphi_j(y)$ might therefore be different. We use the *arguments* of the eigenfunctions to distinguish between them, since this is unambiguous and avoids messy notation.

[7] This reasoning can be extended to motivate kernel choice for the detection of particular dependencies, although this is beyond the scope of the present study. Note also that an alternative Parzen-window based interpretation of kernel choice is given in [9].

which can be satisfied as long as the $\varphi_l(x)$ and $\varphi_l(y)$ are absolutely bounded.[8] The second constraint affects the covariance between kernel eigenfunctions,

$$\begin{aligned}\widetilde{C}_{i,j} &= \operatorname{cov}(\varphi_i(x), \varphi_j(y)) \\ &:= \mathbf{E}_{\mathsf{x},\mathsf{y}}(\varphi_i(\mathsf{x})\varphi_j(\mathsf{y})) - \mathbf{E}_{\mathsf{x}}(\varphi_i(\mathsf{x}))\mathbf{E}_{\mathsf{y}}(\varphi_j(\mathsf{y})).\end{aligned} \tag{4.3}$$

Indeed, this constraint causes $\widetilde{\mathbf{C}}$ to have i, jth entries

$$\widetilde{C}_{i\neq l, j\neq l} := \epsilon_{ij}, \qquad \widetilde{C}_{l,l} := \beta + \epsilon_{ll}, \tag{4.4}$$

where ϵ_{ij} denotes a quantity with absolute value arbitrarily small for large enough l (the proof requires some tedious algebra, but is not difficult: notably, it makes use of the decay result in Lemma 4).

We next expand the functions f and g which define COCO (i.e. elements of the respective RKHSs at which the supremum is attained) as $f(x) = \sum_{i=1}^{\infty} \tilde{f}_i \varphi_i(x)$ and $g(y) = \sum_{j=1}^{\infty} \tilde{g}_j \varphi_j(y)$ (the expansion coefficients are written as vectors $\tilde{\mathbf{f}}, \tilde{\mathbf{g}}$). Using these expansions, the numerator of (4.1) becomes $\operatorname{cov}(f(x), g(y)) = \tilde{\mathbf{f}}^\top \widetilde{\mathbf{C}} \tilde{\mathbf{g}}$, and

$$\begin{aligned}\tilde{\mathbf{f}}^\top \widetilde{\mathbf{C}} \tilde{\mathbf{g}} &\leq \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} |\tilde{f}_i|\,|\tilde{g}_j|\,\epsilon + |\tilde{f}_l|\,|\tilde{g}_l|\,\beta \\ &= \|\tilde{\mathbf{f}}\|_1 \|\tilde{\mathbf{g}}\|_1 \,\epsilon + |\tilde{f}_l|\,|\tilde{g}_l|\,\beta,\end{aligned}$$

where we replace all entries in $\widetilde{\mathbf{C}}$ with their expressions in (4.4), and ϵ is the ϵ_{ij} with largest absolute value. Lemma 4 ensures that $\|\tilde{\mathbf{f}}\|_1$ and $\|\tilde{\mathbf{g}}\|_1$ both converge. In the case of the remaining term $|\tilde{f}_l|\,|\tilde{g}_l|\,\beta$, we divide through by the norms in the denominator of COCO to get

$$\begin{aligned}&|\tilde{f}_l|\,|\tilde{g}_l|\,\beta \left(\sum\nolimits_{i=1}^{\infty} \tilde{f}_i^2 \left(\tilde{k}_i^f\right)^{-1}\right)^{-\frac{1}{2}} \left(\sum\nolimits_{j=1}^{\infty} \tilde{f}_j^2 \left(\tilde{k}_j^g\right)^{-1}\right)^{-\frac{1}{2}} \\ &\leq \beta\sqrt{\tilde{k}_l^f \tilde{k}_l^g},\end{aligned}$$

and the right hand side approaches zero as $l \to \infty$ thanks to Theorem 3.[9] □

We now address how the KCC [1] has the same limitation, being upper bounded by a constant multiple of COCO. The KCC is defined as

$$\begin{aligned}&\mathcal{J}_\kappa\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; \mathcal{F}, \mathcal{G}\right) \\ &:= \sup_{f\in\mathcal{F}, g\in\mathcal{G}} \frac{\operatorname{cov}\left(f(\mathsf{x}), g(\mathsf{y})\right)}{\sqrt{\operatorname{var}\left(f\left(\mathsf{x}\right)\right) + \kappa\|f\|_{\mathcal{F}}^2}\sqrt{\operatorname{var}\left(g\left(\mathsf{y}\right)\right) + \kappa\|g\|_{\mathcal{G}}^2}} \\ &\leq \kappa^{-1}\|f^*\|_{\mathcal{F}}^{-1}\|g^*\|_{\mathcal{G}}^{-1}\operatorname{cov}\left(f^*(\mathsf{x}), g^*(\mathsf{y})\right) \\ &\leq \kappa^{-1}\mathrm{COCO}\left(\mathbf{P}_{\mathsf{x},\mathsf{y}}; F, G\right),\end{aligned}$$

where f^*, g^* attain the supremum in the first line, and we assume f and g to be bounded.

Finally, we demonstrate that the KMI [9] and KGV [1], which are respectively extensions to COCO and the KCC, have the same property. This follows since the KMI can be written as $-\frac{1}{2}\log(\prod_{i=1}^{n}(1-\rho_i^2))$, where $|\rho_i|$ are upper bounded by COCO, and the KGV as $-\frac{1}{2}\log(\prod_{i=1}^{n}(1-\gamma_i^2))$, where the $|\gamma_i|$ are upper bounded by the KCC. Small COCO will therefore cause small KMI, and small KCC will cause small KGV.

5 Bounds

We give two convergence bounds in this section. The first (and simplest) guarantees small population COCO when the empirical COCO is small; the second, which has a more involved derivation, guarantees that if the empirical COCO is large, then the population COCO is also large. The proofs are given in sketch form only; rigorous derivations are provided in [10]. A consequence of these bounds is that the empirical COCO converges to the population COCO at speed $1/\sqrt{n}$. This means that if we define the independence test $\Delta(\boldsymbol{z})$ (Section 4.1) as the indicator that COCO is larger than a term of the form $C\sqrt{\log(1/\alpha)/n}$ with C a constant, then $\Delta(\boldsymbol{z})$ is an α-test with type II error upper bounded by a term approaching zero as $1/\sqrt{n}$.

Our first bound makes use of the following theorem, which applies to U-statistics of the kind we encounter in calculating covariances.

Theorem 9 (Positive deviation bound for one sample U-statistics [11, p. 25]). *Consider a collection of n i.i.d. random variables $(\mathsf{z}_1, \ldots, \mathsf{z}_n)$. We define the U-statistic*

$$\mathsf{u} := \frac{1}{n(n-1)\ldots(n-r+1)} \sum_{\mathbf{i}_r^n} h_{i_1,\ldots,i_r}(\mathsf{z}_{i_1}, \ldots, \mathsf{z}_{i_r}),$$

where the index set $\mathbf{i}_r^n$ is the set of all r-tuples drawn without replacement from $\{1, \ldots, n\}$, and the function h is called the kernel *of the U-statistic. If $a \leq h \leq b$,*

$$\mathbf{P}_{\mathsf{u}}(\mathsf{u} - \mathbf{E}_{\mathsf{u}}(\mathsf{u}) \geq t) \leq \exp\left(\frac{-2t^2\lceil n/r\rceil}{(b-a)^2}\right).$$

We now state the bound.

Theorem 10 (Upper bound on population COCO). *Assume that functions in F and G are*

[8] This condition is not satisfied for all Mercer kernels: see [21, Exercise 2.24]. The assumption holds in most everyday cases we encounter (e.g. the Fourier basis), however, so it is reasonable in this context.

[9] On the basis of this proof, we might suppose that using an RBF kernel with small width (and thus with a slow decay of the coefficients $\tilde{k}_i^f$ and $\tilde{k}_j^f$) would make COCO larger when dependence takes the form (4.2) above, with high order eigenfunctions $\varphi_l(x), \varphi_l(y)$. While this is true, the empirical estimate of COCO will become inaccurate if the spacing between samples significantly exceeds the kernel width: thus, there is a practical limit on how small we can make the kernel.

bounded a.s. by 1. Then for $n > 1$ and all $\delta > 0$, with probability at least $1 - \delta$,

$$\sup_{f\in F, g\in G} \mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) \leq \sup_{f\in F, g\in G} \widehat{\mathrm{cov}}(f(x), g(y)) + \Lambda.$$

where $\Lambda = \sqrt{\frac{2\log(2/\delta)}{n(\sqrt{2}-1)^2}}$, and we denote the empirical covariance based on the sample $\boldsymbol{z}$ as

$$\widehat{\mathrm{cov}}(f(x), g(y)) := \frac{1}{n(n-1)} \sum_{i\neq j} f_i g_j - \frac{1}{n}\sum_{i=1}^{n} f_i g_i,$$

where $f_i := f(x_i)$ and $g_j := g(y_j)$.

Proof. First, we rearrange

$$\sup_{f\in F, g\in G} \mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) - \sup_{f\in F, g\in G} \widehat{\mathrm{cov}}(f(x), g(y)) \leq \sup_{f\in F, g\in G} \left(\mathrm{cov}(f(\mathsf{x}), g(\mathsf{y})) - \widehat{\mathrm{cov}}(f(x), g(y))\right).$$

We can therefore ignore the suprema, and treat only the random variables $\mathsf{f} := f(\mathsf{x})$ and $\mathsf{g} := g(\mathsf{y})$. To complete the bound, we make the split

$$\mathbf{P}_{\mathsf{f}^n,\mathsf{g}^n}\left(\mathrm{cov}(\mathsf{f},\mathsf{g}) - \widehat{\mathrm{cov}}(f,g) \geq t\right) \leq \mathbf{P}_{\mathsf{f}^n,\mathsf{g}^n}\left(-\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}_i\mathsf{g}_i + \mathbf{E}_{\mathsf{f},\mathsf{g}}(\mathsf{fg}) \geq (1-\alpha)t\right) + \mathbf{P}_{\mathsf{f}^n,\mathsf{g}^n}\left(\frac{1}{n(n-1)}\sum_{i\neq j}\mathsf{f}_i\mathsf{g}_j - \mathbf{E}_{\mathsf{f}}(\mathsf{f})\mathbf{E}_{\mathsf{g}}(\mathsf{g}) \geq \alpha t\right)$$

The first term is bounded straightforwardly using Hoeffding's inequality [11]. To bound the second term in the sum, we define the random vector $\mathsf{z}_i := (\mathsf{f}_i, \mathsf{g}_i)$, and the kernel $h_{i,j}(\mathsf{z}_i, \mathsf{z}_j) := \mathsf{f}_i\mathsf{g}_j$. It is clear that Theorem 9 then applies. We complete the proof by setting $\alpha = 2 - \sqrt{2}$. □

A lower bound on the population COCO is harder to compute, since we have to deal with the suprema.

Theorem 11 (Lower bound on population COCO). *Assume functions in F and G are bounded a.s. by 1, and that the functions $k_f(x,x) \leq 1$ and $k_g(y,y) \leq 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then for $n > 1$ and all $\delta > 0$, with probability at least $1 - \delta$,*

$$\sup_{f\in F, g\in G} \widehat{\mathrm{cov}}(f,g) \leq \sup_{f\in F, g\in G} \mathrm{cov}(f,g) + \frac{134}{\sqrt{n}} + \sqrt{\frac{18\log 2/\delta}{n}}.$$

Proof. We begin with a rearrangement of the suprema;

$$\sup_{f\in F, g\in G} \widehat{\mathrm{cov}}(f(x), g(y)) - \sup_{f\in F, g\in G} \mathrm{cov}_{\mathsf{x},\mathsf{y}}(f(\mathsf{x}), g(\mathsf{y})) \leq \sup_{f\in F, g\in G}\left(\mathbf{E}_{\mathsf{x},\mathsf{y}} f(\mathsf{x})g(\mathsf{y}) - \frac{1}{n}\sum_{i=1}^{n} f(x_i)g(y_i)\right) + \sup_{f\in F, g\in G}\left(\frac{1}{n(n-1)}\sum_{i\neq j} f(x_i)g(y_j) - \mathbf{E}_{\mathsf{x}} f(\mathsf{x})\mathbf{E}_{\mathsf{y}} g(\mathsf{y})\right)$$

The first term is bounded using McDiarmid [18] and symmetrisation in the usual way. In the case of the second term, we begin with McDiarmid to get

$$\mathbf{P}_{\mathsf{x}^n,\mathsf{y}^n}\left(\sup_{f\in F, g\in G}\left(\frac{1}{n(n-1)}\sum_{i\neq j} f_i g_j - \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g\right) \geq \underbrace{\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n}\sup_{f\in F, g\in G}\left[\frac{1}{n(n-1)}\sum_{i\neq j} f_i g_j - \mathbf{E}_{\mathsf{x}} f \mathbf{E}_{\mathsf{y}} g\right]}_{(a)} + t\right) \leq e^{\frac{-nt^2}{8}}.$$

We cannot symmetrise this expression directly: instead, we first apply the Hoeffding decomposition and then decouple, following [5]. This yields an upper bound on the expectation (a) that we can symmetrise. We do not go into detail, but the idea is to replace certain of the random variables by independent copies. After decoupling and symmetrisation, we obtain

$$\begin{aligned}(\mathrm{a}) \leq 32\,\mathbf{E}\sup_{f,g}\Bigg(&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}^{n}\sigma_i\sigma'_j\Big(f(\mathsf{x}_i)g(\mathsf{y}'_j) \\ &- f(\mathsf{x}_i)g(\mathsf{y}''_j) - f(\mathsf{x}''_i)g(\mathsf{y}'_j) + f(\mathsf{x}'_i)g(\mathsf{y}''_j)\Big)\Bigg) \\ &+\frac{4}{n}\mathbf{E}\sup_{f,g}\sum_{i=1}^{n}(\sigma_i f(\mathsf{x}_i)g(\mathsf{y}'_i)) = (b),\end{aligned}$$

where the σ_i are Rademacher random variables that take values in $\{-1, 1\}$ with equal probability, σ'_i are independent copies of σ_i, x'_i, x''_i are independent copies of x_i, and y'_i, y''_i are independent copies of y_i. To conclude the proof, it turns out that we do not need to explicitly deal with these additional copies: instead, we apply a simple additional bound (see [10]) to get

$$\begin{aligned}(b) \leq &\frac{128}{n(n-1)}\mathbf{E}_{\mathsf{x}^n,\mathsf{y}^n}\sqrt{\sum_{i\neq j} k_f(\mathsf{x}_i,\mathsf{x}_i)k_g(\mathsf{y}_i,\mathsf{y}_i)} \\ &+\frac{4}{n}\mathbf{E}_{\mathsf{x}^n,\mathsf{y}'^n}\sqrt{\sum_{i=1}^{n} k_f(\mathsf{x}_i,\mathsf{x}_i)k_g(\mathsf{y}'_i,\mathsf{y}'_i)},\end{aligned}$$

and then substitute $k_f(x,x) \leq 1$ and $k_g(y,y) \leq 1$. □

6 Experiments and discussion

We previously applied COCO in the context of independent component analysis (ICA) [9], where it performed similarly to the kernel canonical correlation [1]. Thus, COCO has been established in practice as a useful test of *independence*. In the present study, we give preliminary results obtained when using COCO to determine regions that show *high* dependence in

the macaque primary visual cortex. We are able to monitor neural activity in three different ways: (a) electrophysiology only with large electrode arrays, (b) fMRI only, and (c) combined electrode and fMRI measurements. The present section deals only with dependence measures on the fMRI signals, but we are currently expanding this analysis to cover a broader range of acquisition techniques [17], so as to compare the dependence found for these different types of measurement. Our fMRI readings were taken using a 4.7T scanner with a sampling frequency of 4Hz and a 96mmx96mm field of view (FOV), with resolution 256x256 and 0.5mm slice thickness, in accordance with the procedure in [16]. The stimulus used was a clip from "Star Wars", which was chosen so as to excite a broad range of activity within the visual cortex. Dependence was investigated for voxels in the primary visual area (V1) for a total of 250 voxels, and the signal duration during stimulus was 250 seconds (1000 samples). The results obtained are an aggregate over 35 such experiments.

The observed fMRI signals were contaminated with a breathing component. Since the macaque monkey was under general anaesthetic during data acquisition, breathing was mechanically assisted, and had a constant frequency of approximately 0.4Hz. We modelled this breathing as being of constant amplitude and linearly superposed on the haemodynamic response. This model is motivated by the narrowness of the spectral peaks at the breathing frequency and harmonics, which suggests that any amplitude modulation of the breathing signal is of very low frequency, and can be assumed effectively constant. Thus, while we could not directly recover the true breathing contamination at each voxel, we were able to use the decrease in the spectral peak at the breathing frequency, averaged across all voxels, as a measure of success in removing the breathing artefacts. Harmonics at integer multiples of the fundamental frequency were modelled in the same way. The exact frequency of the breathing signal was found by averaging the spectrum over all voxels, and the phase at each voxel was chosen to maximise the projection in the time domain of the breathing sinusoid. Only voxels near large blood vessels were contaminated by the breathing signal, and thus a threshold test was applied to the spectrum at each voxel, to test whether a substantial breathing component was visible. Where breathing was present, a sinusoid of corresponding frequency and phase was projected out in the time domain (thus also removing the associated frequency domain sidelobes caused by finite signal duration). The same procedure was used to remove the first two harmonics. We did not band-pass filter the signal to remove the breathing, as this would have eliminated a greater portion of the spectral components due to the haemodynamic activity.

As dependence measures between pairs of voxels, we applied cross correlation between voxels, the mutual information (MI) (computed using the method in [4]), and COCO (using RKHSs with Gaussian kernels). The variation in dependence between voxels was studied with all three methods, as a function of average distance between voxels (in other words, we grouped together all pairs of voxels an equal distance from each other; we then clustered these pairs so as to draw together voxel pairs with similar distances). The regions of interest (ROI) were constrained to be convex sets so as to avoid incorrect distance estimates. Specifically, Euclidean distances at the image level may significantly differ from the actual axonic distances connecting neural sites. We subtracted a baseline dependence from each of the dependence measures, which was obtained by averaging the dependence between the V1 region and a test region of the brain, the latter being unrelated to visual processing. The dependence amplitudes were then divided by the standard deviation in the average dependence between V1 and this test region. Results are plotted in Figure 6.1.

Comparing the dependence measures before and after breathing removal shows significant effects of respiratory artefacts on the high order[10] dependence vs. distance curves (COCO and MI): this finding suggests extreme caution for studies in humans, in which respiration-induced signals cannot easily be modelled due to low temporal sampling rates, as well as variable respiration frequency and amplitude. Prior to breathing removal, COCO and the MI overestimate the dependence between voxels (the breathing artefacts being a source of considerable similarity), as does the correlation, though to a lesser extent. This can be explained by phase shifts between the breathing contamination observed at different voxels, which reduce the correlation but have less effect on more general measures of dependence. The high order dependence curves also flatten out at about 5mm once breathing is removed, but continue to decay with distance when breathing is present. By contrast, the correlation prior to breathing removal is constant (to within observational uncertainty) after about 2mm; following breathing removal, however, the point at which it flattens out is more difficult to determine. Finally, compared with the MI, COCO at short distances is a larger multiple of the standard deviation in test region dependence, which might make COCO a more reliable measure of such short range dependencies. On the other hand, both the MI and the correlation fall to a baseline level of activity greater than that in the test region, which COCO does not detect. Additional experiments on a greater number of subjects and stimuli will be used to verify these observations.

Further work will focus on the construction of dependence-distance functions, using for instance the well developed mathematical framework of flat brain-

[10] The correlation takes into account only second order dependence, whereas COCO and the MI can detect dependence of any order.

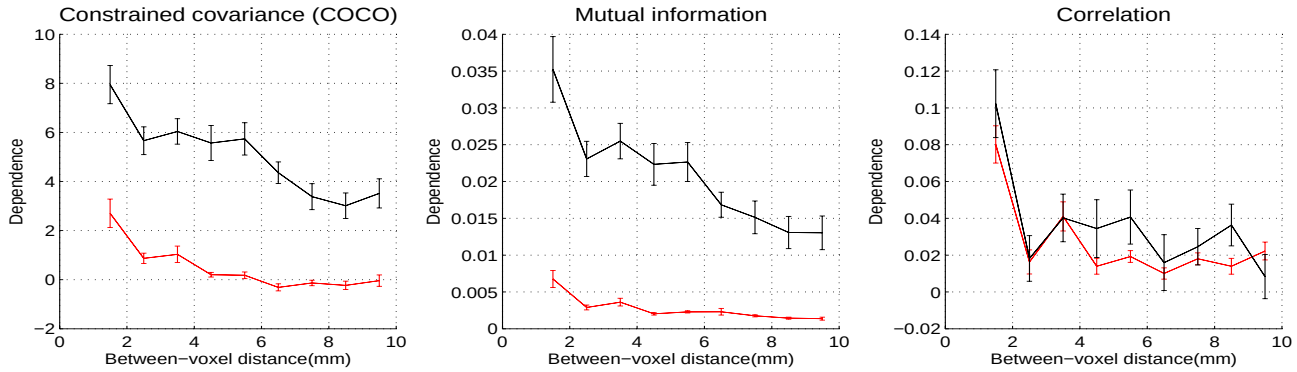

Figure 6.1: Results before (black) and after (red) breathing removal, for COCO, the mutual information (MI), and the correlation. The offset and scaling of the dependence is described in the main body of the text. The horizontal axis displays the average distance between voxel pairs.

map generation [7]. Dependence tests that take into account the fact that the signals are not i.i.d. will also be compared with the present approaches. In addition, it is of interest to develop kernel-based dependence measures for non-i.i.d. time series.

References

[1] F. Bach and M. Jordan. Kernel independent component analysis. *JMLR*, 3:1–48, 2002.

[2] G.H. Bakır, A. Gretton, M. Franz, and B. Schölkopf. Multivariate regression with stiefel constraints. Technical Report 101, Max Planck Institute for Biological Cybernetics, 2004.

[3] C. Buchel and K. Friston. Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. *Cerebral Cortex*, 7:768–778, 1997.

[4] G. A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321, 1999.

[5] V. de la Peña and E. Giné. *Decoupling: from Dependence to Independence*. Springer, New York, 1999.

[6] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of mathematics*. Springer, New York, 1996.

[7] D. Van Essen, H. Drury, S. Joshi, and M. Miller. Functional and structural mapping of human cerebral cortex: solutions are in the surfaces. *Proceedings of Nat. Acad. of Sciences of the USA*, 95:788–795, 1998.

[8] K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *JMLR*, 5:73–99, 2004.

[9] A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. Technical report, MPI for Biological Cybernetics, 2003.

[10] A. Gretton, A. Smola, O. Bousquet, and R. Herbrich. Behaviour and convergence of the constrained covariance. Technical report, MPI for Biological Cybernetics, 2004.

[11] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[12] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, New York, 2001.

[13] Yu. I. Ingster. An asymptotically minimax test of the hypothesis of independence. *J. Soviet Math.*, 44:466–476, 1989.

[14] J. Jacod and P. Protter. *Probability Essentials*. Springer, New York, 2000.

[15] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman. Canonical correlation analysis when the data are curves. *J. R. Stat. Soc. B*, 55(3):725–740, 1993.

[16] N. Logothetis, H. Guggenberger, and J. Pauls S. Peled. Functional imaging of the monkey brain. *Nature Neuroscience*, 2:555–562, 1999.

[17] N. Logothetis, J. Pauls, M. Augath, T. Trinath, and A. Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412:150–157, 2001.

[18] C. McDiarmid. On the method of bounded differences. *Surveys in Combinatorics*, pages 148–188, 1989. Cambridge University Press.

[19] A. Rényi. On measures of dependence. *Acta Math. Acad. Sci. Hungar.*, 10:441–451, 1959.

[20] S. Saitoh. *Theory of Reproducing Kernels and its Applications*. Longman Scientific and Technical, Harlow, UK, 1988.

[21] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT press, Cambridge, MA, 2002.

[22] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2, 2001.