A Kernel Independence Test for Random Processes

Kacper Chwialkowski KACPER.CHWIALKOWSKI@GMAIL.COM
University College London, Computer Science Department

Arthur Gretton ARTHUR.GRETTON@GMAIL.COM
University College London, Gatsby Computational Neuroscience Unit

Abstract

A non-parametric approach to the problem of testing the independence of two random processes is developed. The test statistic is the Hilbert-Schmidt Independence Criterion (HSIC), which was used previously in testing independence for i.i.d. pairs of variables. The asymptotic behaviour of HSIC is established when computed from samples drawn from random processes. It is shown that earlier bootstrap procedures which worked in the i.i.d. case will fail for random processes, and an alternative consistent estimate of the p-values is proposed. Tests on artificial data and real-world forex data indicate that the new test procedure discovers dependence which is missed by linear approaches, while the earlier bootstrap procedure returns an elevated number of false positives.

1. Introduction

Measures of statistical dependence between pairs of random variables (X, Y) are well established, and have been applied in a wide variety of areas, including fitting causal networks (Pearl, 2000), discovering features which have significant dependence on a label set (Song et al., 2012), and independent component analysis (Hyvärinen et al., 2004). Where pairs of observations are independent and identically distributed, a number of non-parametric tests of independence have been developed (Feuerverger, 1993; Gretton et al., 2007; Székely et al., 2009; Gretton & Györfi, 2010), which determine whether the dependence measure value is statistically significant. These non-parametric tests are consistent against any fixed alternative - they make no assumptions as to the nature of the dependence.

For many data analysis tasks, however, the observations be-

ing tested are drawn from a time series: each observation is dependent on its past values. Examples include audio signals, financial data, and brain activity. Given two such random processes, we propose a hypothesis test of instantaneous dependence, of whether the two signals are dependent at a particular time t. Our test satisfies two important properties: it is consistent against any fixed alternatives, and it is non-parametric - we do not assume the dependence takes a particular form (such as linear correlation), nor do we require parametric models of the time series. We further avoid making use of a density estimate as an intermediate step, so as to avoid the assumption that the distributions have densities (for instance, when dealing with text or other structured data).

We use as our test statistic the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005; 2007), which can be interpreted as the distance between embeddings of the joint distribution and the product of the marginals in a reproducing kernel Hilbert space (RKHS) (Gretton et al., 2012, Section 7). When characteristic RKHSs are used, the HSIC is zero iff the variables are independent (Sriperumbudur et al., 2010). Under the null hypothesis of independence, $P_{XY} = P_X P_Y$, the minimum variance estimate of HSIC is a degenerate U-statistic. The distribution of the empirical HSIC under the null is an infinite sum of independent χ^2 variables (Gretton et al., 2007), which follows directly from e.g. (Serfling, 2009, Ch. 5). In practice, given a sample $(x_i, y_i)_{i=1}^n$ of pairs of variables drawn from P_{XY}, the null distribution is approximated by a bootstrap procedure, where a histogram is obtained by computing the test statistic on many different permutations $\{x_i, y_{\pi(i)}\}_{i=1}^n$, to decouple X and Y.

In the case where the samples $Z_t = (X_t, Y_t)$ are drawn from a random process, the analysis of the asymptotic behaviour of HSIC requires substantially more effort than in the i.i.d. case. As our main contribution, we obtain both the null and alternative distributions of HSIC for random processes, where the null distribution is defined as X_t being independent of Y_t at time t. Such a test may be used for re-

jecting causal effects (i.e., whether one signal is not dependent on the values of another signal at a particular delay) or instant coupling (see our first experiment in Section 4.2).[1] The null distribution is again an infinite weighted sum of χ^2 variables, however these are now correlated, rather than independent. Under the alternative hypothesis, the statistic has an asymptotically normal distribution.

For the test to be used in practice, we require an empirical estimate of the null distribution, which gives the correct test threshold when $Z_t = (X_t, Y_t)$ is a random process. Evidently, the bootstrap procedure used in the i.i.d. case is incorrect, as the temporal dependence structure within the Y_t will be removed. This turns out to cause severe problems in practice, since the permutation procedure will give an increasing rate of false positives as the temporal dependence of the Y_t increases (i.e., dependence will be detected between X_t and Y_t, even though none exists, this is also known as a Type I error). Instead, our null estimate is obtained by making shifts of one signal relative to the other, so as to retain the dependence structure within each signal. Consequently, we are able to keep the Type I error at the designed level $\alpha = 0.05$. In our experiments, we address three examples: one artificial case consisting of two signals which are dependent but have no correlation, and two real-world examples on forex data. HSIC for random processes reveals dependencies that classical approaches fail to detect. Moreover, our new approach gives the correct Type I error rate, whereas a bootstrap-based approach designed for i.i.d. signals returns too many false positives.

Related work Prior work on testing independence in time series may be categorized in two branches: testing serial dependence within a single time series, and testing dependence between one time series and another. The case of serial dependence turns out to be relatively straightforward, as under the null hypothesis, the samples become independent: thus, the analysis reduces to the i.i.d. case. Pinkse (1998); Diks & Panchenko (2005) provide a quadratic forms function-based serial dependence test which employs the same statistic as HSIC. Due to the simple form of the null hypothesis, the analysis of (Serfling, 2009, Ch. 5) applies. Further work in the context of the serial dependency testing includes simple approaches based on rank statistics e.g. Spearman's correlation or Kendall's tau, correlation integrals e.g. (Brock et al., 1996); criteria based on integrated squared distance between densities e.g (Rosenblatt & Wahlen, 1992); KL-divergence based criteria e.g. (Robinson, 1991; Hong & White, 2005); and generalizations of KL-divergence to so called q-class entropies e.g. (Granger et al., 2004; Racine & Maasoumi, 2007).

In most of the tests of independence of two time series, specific conditions have been enforced, e.g that processes follow a moving average specification or the dependence is linear. Prior work in the context of dependency tests of two time series includes cross covariance based tests e.g. (Haugh, 1976; Hong, 1996; Shao, 2009); and a Generalized Association Measure based criterion (Fadlallah et al., 2012). Some work has been undertaken in the non-parametric case, however. A non-parametric measure of independence for time series, based on the Hilbert-Schmidt Independence Criterion, was proposed by Zhang et al. (2008). While this work established the convergence in probability of the statistic to its population value, no asymptotic distributions were obtained, and the statistic was not used in hypothesis testing. To our knowledge, the only non-parametric independence test for pairs of time series is due to Besserve et al. (2013), which addresses the harder problem of testing independence across all time lags simultaneously. [2] The procedure is to compute the Hilbert-Schmidt norm of a cross-spectral density operator (the Fourier transform of the covariance operator at each time lag). The resulting statistic is a function of frequency, and must be zero at all frequencies for independence, so a correction for multiple hypothesis testing is required. It is not clear how the asymptotic analysis used in the present work would apply to this statistic, and this remains an interesting topic of future study.

The remaining material is organized as follows. In Section 2 we provide a brief introduction to random processes and various mixing conditions, and an expression for our independence statistic, HSIC. In Section 3, we characterize the asymptotic behaviour of HSIC for random variables with temporal dependence, under the null and alternative hypotheses, and establish the test consistency. We propose an empirical procedure for constructing a statistical test, and demonstrate that the earlier bootstrap approach will not work for our case. Section 4 provides experiments on synthetic and real data.

2. Background

In this section we introduce necessary definitions referring to random processes. We then go on to define a V-statistic estimate of the Hilbert-Schmidt Independence Criterion, which applies in the i.i.d. case.

Random process. First, we introduce the probabilistic tools needed for pairs of time series. Let $(Z_t, \mathcal{F}_t)_{t \in \mathbb{N}}$ be a strictly stationary sequence of random variables defined on a probability space Ω with a probability measure P and natural filtration $\mathcal{F}_t$. Assume that Z_t denotes a pair of ran-

---

[1]We distinguish our case from the problem of ensuring time series are independent simultaneously across all time lags, e.g the null will hold even if $X_t = Y_{t-1}$ where Y_t is white noise.

[2] Let X_t follow a MA(2) model and put $Y_t = X_{t-20}$. This is a case addressed by Besserve et al. (2013), who will reject their null hypothesis, whereas our null is accepted

dom variables i.e. $Z_t = (X_t, Y_t)$, where X_t is defined on $\mathcal{X}$, and Y_t on $\mathcal{Y}$. Each Z_t takes values in a measurable Polish space $(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$. The space $\mathbf{Z}$ is a Cartesian product of two Polish spaces $\mathbf{X}$ and $\mathbf{Y}$, endowed with a natural Borel sigma field and a probability measure.

We introduce a sequence of independent copies of Z_0, i.e., $(Z_t^*)_{t \in \mathbb{N}}$. Since Z_t is stationary, Z_t^* retains the dependence between random variables X_t and Y_t, but breaks the temporal dependence.

Next, we formalize a concept of memory of a process. A process is called absolutely regular (β-mixing) if $\beta(m) \to 0$, where

$$\beta(m) = \frac{1}{2} \sup_n \sup \sum_{i=1}^{I} \sum_{j=1}^{J} |P(A_i \cap B_j) - P(A_i)P(B_j)|.$$

The second supremum in the $\beta(m)$ definition is taken over all pairs of finite partitions $\{A_1, \cdots, A_I\}$ and $\{B_1, \cdots, B_J\}$ of the sample space such that $A_i \in \mathcal{A}_1^n$ and $B_j \in \mathcal{A}_{n+m}^\infty$, and $\mathcal{A}_b^c$ is a sigma field spanned by a subsequence, $\mathcal{A}_b^c = \sigma(Z_b, Z_{b+1}, ..., Z_c)$. A process is called uniform mixing (ϕ-mixing) if $\phi(m) \to 0$, where

$$\phi(m) = \sup_n \sup_{A \in \mathcal{A}_1^n} \sup_{B \in \mathcal{A}_{n+m}^\infty} |P(B|A) - P(B)|.$$

Uniform mixing implies absolute regularity, i.e. $\beta(m) \le \phi(m)$ (Bradley et al., 2005). Under technical assumptions, Autoregressive Moving Average processes — or more generally Markov Chains — are absolutely regular or uniformly mixing (Doukhan, 1994).

Hilbert-Schmidt Independence Criterion Let k, l be positive definite kernels associated with respective reproducing kernel Hilbert spaces $\mathcal{H}_\mathcal{X}$ on $\mathcal{X}$, and $\mathcal{H}_\mathcal{Y}$ on $\mathcal{Y}$. We assume that k and l are bounded and continuous. We associate to the random variable X a mean embedding $\mu_X(x) := \mathcal{E}_X k(X, x)$, such that $\forall f \in \mathcal{H}_\mathcal{X}$, $\langle f, \mu_X \rangle_{\mathcal{H}_\mathcal{X}} = \mathcal{E}_X(f(X))$ (Berlinet & Thomas-Agnan, 2004; Smola et al., 2007). We assume k, l are characteristic kernels, meaning the mappings μ_X and $\mu_Y(y) := \mathcal{E}_Y l(Y, y)$ are injective embeddings of the probability measures to the corresponding RKHSs; i.e., distributions have unique embeddings (Fukumizu et al., 2008; Sriperumbudur et al., 2010).

We next recall a measure of statistical dependence, the Hilbert-Schmidt Independence Criterion (HSIC), which can be expressed in terms of expectations of RKHS kernels (Gretton et al., 2005; 2007). Denote a group of permutations over 4 elements by S_4, with π one of its elements, i.e., a permutation of four elements. We define

$$h(z_1, z_2, z_3, z_4) = \frac{1}{4!} \sum_{\pi \in S_4} k(x_{\pi(1)}, x_{\pi(2)})[l(y_{\pi(1)}, y_{\pi(2)}) +$$
$$+ l(y_{\pi(3)}, y_{\pi(4)}) - 2l(y_{\pi(2)}, y_{\pi(3)})].$$

Lemma 1. *Let γ be an expected value of the function h, $\gamma = \mathcal{E}h(Z_1^*, Z_2^*, Z_3^*, Z_4^*)$. This expectation corresponds to HSIC, computed using a function symmetric in its arguments. For k and l characteristic, continuous, translation invariant, and vanishing at infinity, γ is equal to zero if and only if the null hypothesis holds (see (Lyons, 2013, Lemma 3.8), applying (Sriperumbudur et al., 2011, Proposition 2), and the note at the end of Section 5).*

The value of γ corresponds to a distance between embeddings of (X_1^*, Y_2^*) and (X_1^*, Y_1^*) to an RKHS with the product kernel $\kappa = k \cdot l$ (Gretton et al., 2012, Section 7). A biased empirical estimate of the Hilbert-Schmidt Independence Criterion can be expressed as a V-statistic (the unbiased estimate is a U-statistic, however the difference will be accounted for when constructing a hypothesis test, through an appropriate null distribution).

V statistics. A V-statistic of a k-argument, symmetric function f is written

$$V(f, Z) = \frac{1}{n^k} \sum_{1 \le i_1, \cdots, i_k \le n} f(Z_{i_1}, ..., Z_{i_k}). \quad (1)$$

Gretton et al. (2005) show that the biased estimator of γ is $V(h, Z)$. The asymptotic behaviour of this statistic depends on the degeneracy of the function that defines it. We say that a k-argument, symmetric function f is j-degenerate ($j < k$) if for each $z_1, \cdots, z_j \in \mathbb{Z}$,

$$\mathcal{E}f(z_1, \cdots, z_j, Z_{j+1}^*, \cdots, Z_k^*) = 0.$$

If $j = k - 1$ we say that the function is canonical. We refer to a normalized V statistic as a V-statistic multiplied by the sample size, $n \cdot V$.

3. HSIC for random processes

In this section we construct the Hilbert-Schmidt Independence Criterion for random processes, and define its asymptotic behaviour. We then introduce an independence testing procedure for time series.

We introduce two hypotheses: the null hypothesis $\mathbf{H_0}$ that X_t and Y_t are independent, and the alternative hypothesis $\mathbf{H_1}$ that they are dependent. To build a statistical test based on $n \cdot V(h, Z)$ we need two main results. First, if null hypothesis holds, we show $n \cdot V(h, Z)$ converges to a random variable. Second, if the null hypothesis does not hold, the $n \cdot V(h, Z)$ estimator diverges to infinity. Following these results, the Type I error (the probability of mistakenly rejecting the null hypothesis) will stabilize at the design parameter α, and the Type II error (the probability of mistakenly accepting the null hypothesis when the variables are dependent) will drop to zero, as the sample size increases.

We begin by introducing an auxiliary kernel function s, and characterize the normalized V-statistic distribution of s us-

ing a CLT introduced by (Borisov & Volodko, 2008). We then show that the normalized V-statistic associated with the function s has the same asymptotic distribution as the $n \cdot V(h, Z)$ distribution.

Let s be an auxiliary function $s(z_1, z_2) = \tilde{k}(x_1, x_2)\tilde{l}(y_1, y_2)$, where

$$\tilde{k}(x_1, x_2) = k(x_1, x_2) - \mathcal{E}k(x_1, X_2) - \mathcal{E}k(X_1^*, x_2) + \mathcal{E}k(X_1^*, X_2^*),$$

and $\tilde{l}$ is defined similarly.

Both $\tilde{k}$ and $\tilde{l}$ are kernels, meaning that they are dot products between features centred in their respective RKHSs (Berlinet & Thomas-Agnan, 2004). Therefore $s = \tilde{k} \cdot \tilde{l}$ defines a kernel on a product space of pairs Z_t. Using Mercer's Theorem we obtain an expansion for s.

Statement 1. *By Steinwart & Scovel (2012) Corollary 3.5, the bounded, continuous kernel s has a representation*[3]

$$s(z_a, z_b) = \sum_{i=1}^{\infty} \lambda_i e_i(z_a) e_i(z_b), \qquad (2)$$

where $(e_i)_{i \in \mathbb{N}+}$ denotes an orthonormal basis of $L^2(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$. The series $(\sum_{i=1}^{N} \lambda_i \ e_i(z_a) e_i(z_b))$ converges absolutely and uniformly. e_i are eigenfunctions of s and λ_i are eigenvalues of s.

We will henceforth assume that for every collection of pairwise distinct subscripts (t_1, t_2), the distribution of (Z_{t_1}, Z_{t_2}) is absolutely continuous with respect to the $(Z_{t_1}^*, Z_{t_2}^*)$ distribution. This assumption prevents the occurrence of degenerate cases, such that all Z_t being the same. The following results are proved in Section 5.1.

Lemma 2. *Let the process Z_t have a mixing coefficient smaller than m^{-3} $(\beta(m), \phi(m) \leq m^{-3})$ and satisfy either of the following conditions:*

A Z_t *is ϕ-mixing.*

B Z_t *is β-mixing. For some $\epsilon > 0$ and for an even number $c \geq 2$, the following holds*

    1. $\sup_i \mathcal{E}|e_i(X_1)|^{2+\epsilon} \leq \infty$, *where e_i is the basis introduced in the Statement 1 and $|\cdot|$ denotes an absolute value.*

    2. $\sum_{m=1}^{\infty} \beta^{\epsilon/(2+\epsilon)}(m) < \infty$.

If the null hypothesis holds, then s is a canonical function and a kernel. What is more,

$$\lim_{n \to \infty} n \cdot V(s, Z) \stackrel{D}{=} \sum_{j}^{\infty} \lambda_j \tau_j^2,$$

---

[3]A bounded kernel is compactly embedded into $L^2(\mathbf{Z}, \mathcal{B}(\mathbf{Z}), P_{\mathbf{Z}})$ (Steinwart & Scovel, 2012).

where τ_j is a centred Gaussian sequence with the covariance matrix

$$\mathcal{E}\tau_a \tau_b = \mathcal{E}e_a(Z_1)e_b(Z_1) + \\ + \sum_{j=1}^{\infty} [\mathcal{E}e_a(Z_1)e_b(Z_{j+1}) + \mathcal{E}e_b(Z_1)e_a(Z_{j+1})].$$

We now characterize the asymptotics of $V(h, Z)$.

Theorem 1. *Under assumptions of Lemma 2, if $\mathbf{H_0}$ holds, then the asymptotic distribution of the empirical HSIC (with scaling n) is the same as that of $n \cdot V(s, Z)$,*

$$\lim_{n \to \infty} n \cdot V(h, Z) \stackrel{D}{=} \lim_{n \to \infty} n \cdot V(s, Z).$$

Theorem 2. *Under assumptions of the Lemma 2, if $\mathbf{H_1}$ holds, then $\gamma > 0$ and $\sqrt{n}(V(h, Z) - \gamma)$ has asymptotically normal distribution with mean zero and finite variance.*

Consequently, if the null hypothesis does not hold then $P(n \cdot V(h, Z) > C) = P(V(h, Z) > \frac{C}{n}) \to 1$ for any fixed C. Finally, we show that the γ estimator is easy to compute. According to Gretton et al. (2007, equation 4), $V(h, Z) = n^{-2}trHKHL$, where $K_{ab} = k(X_a, X_b)$, $L_{ab} = l(Y_a, Y_b)$, $H_{ij} = \delta_{ij} - n^{-1}$ and n is a sample size.

Testing procedure We begin by showing that the H_0 distribution of the γ estimator obtained via the bootstrap approach of (Diks & Panchenko, 2005; Gretton et al., 2007) gives an incorrect p-value estimate when used with independent random processes. In fact, the null hypothesis obtained by permutation corresponds to the processes being *both* i.i.d. *and* independent from each other. Recall the covariance structure of the γ estimator from Theorem 1,

$$\mathcal{E}\tau_a \tau_b = \mathcal{E}e_a(Z_1)e_b(Z_1) + \\ + \sum_{j=1}^{\infty} [\mathcal{E}e_a(Z_1)e_b(Z_{j+1}) + \mathcal{E}e_b(Z_1)e_a(Z_{j+1})]. \tag{3}$$

We can represent e_a and e_b as $e_a(z) = e_u^X(x)e_o^Y(y)$, $e_b(z) = e_i^X(x)e_p^Y(y)$, as a decomposition of the $\mathbf{Z}$ basis into bases of $\mathbf{X}$ and $\mathbf{Y}$, respectively. Consider a partial sum T_n of series from the above equation (3), with X_t replaced with its permutation $X_{\pi(t)}$,

$$T_n = \sum_{j=1}^{n} \mathcal{E}e_u^X(X_{\pi(1)})e_i^X(X_{\pi(j+1)})\mathcal{E}e_o^Y(Y_1)e_p^Y(Y_{j+1}). \tag{4}$$

Using covariance inequalities from (Doukhan, 1994, Section 1.2.2) we conclude that $\mathcal{E}e_o^Y(Y_1)e_p^Y(Y_{j+1}) = O(\Lambda(j)^{\frac{1}{2}})$ and $\mathcal{E}e_u^X(X_{\pi(1)})e_i^X(X_{\pi(j+1)}) = O(\Lambda(|\pi(j) - $

$\pi(1)|)^{\frac{1}{2}})$ where Λ is an appropriate mixing coefficient (β or ϕ). Recall that $0 < \Lambda(j) < Cj^{-3}$.

We can therefore reduce the problem to the convergence of a random variable

$$S_n = \sum_{j=1}^{n} \Lambda(j)^{\frac{1}{2}} \Lambda(|\pi(j) - \pi(1)|)^{\frac{1}{2}}, \qquad (5)$$

where π is a random permutation drawn from the uniform distribution over the set of n-element permutations. In the supplementary material we show that this sum converges in probability to zero.

Since $S_n > T_n > 0$, then T_n converges to zero in probability, and consequently the covariance matrix entry $\mathcal{E}\tau_a\tau_b$ converges to unity for $a = b$, and to zero otherwise. Indeed, the expected value $\mathcal{E}e_a((X_{\pi(1)}, Y_1))e_b((X_{\pi(1)}, Y_1)) = 0$ if $a \neq b$ and is equal to one otherwise. Note that this is the covariance matrix described by Gretton et al. (2007).

A correct approach to approximating the asymptotic null distribution of $n \cdot V(h, Z)$ under $\mathbf{H}_0$ is by *shifting* of one time series relative to the other. Define the shifted process $S_t^c = Y_{t+c \bmod n}$ for an integer c, $0 \leq c \leq n$ and $0 \leq t \leq n$. If we let c vary over $0 \leq A \leq B \leq n$ for A such that the dependence between Y_{t+A} and X_t is negligible, then we can approximate the null distribution with an empirical distribution calculated on points $(V(h, Z^k))_{A \leq k \leq B}$, where $Z_t^k = (X_t, S_t^k)$. This is due to the fact that the shifted process S_t^c retains most of the dependence, since it does not scramble the time index.[4] We call this method Shift HSIC. In the supplementary material we show that Shift HSIC samples from the correct null distribution.

4. Experiments

In the experiments we compare Shift HSIC with the Bootstrap HSIC of Gretton et al. (2007). We investigate three cases: an artificial dataset, where two time series are coupled non-linearly; and two forex datasets, where in one case we seek residual dependence after one time series has been used to linearly predict another, and in the other case, we reveal strong dependencies between signals that are not seen via linear correlation.

4.1. Artificial data

Non-linear dependence. We investigate two dependent, autoregressive random processes X_t, Y_t, specified by

$$X_t = aX_{t-1} + \epsilon_t \quad Y_t = aY_{t-1} + \eta_t, \qquad (6)$$

with an autoregressive component a. The coupling of the processes is a result of the dependence in the innovations

---

Algorithm 1 Generate innovations
> **Input:** extinction rate $0 \leq p \leq 1$, radius r.
> **repeat**
>> Initialize η_t, ϵ_t to $N(0, 1)$ and d to a number uniformly distributed on $[0, 1]$.
>> **if** $\eta_t^2 + \epsilon_t^2 > r^2$ **or** $d > p$ **then**
>>> return η_t, ϵ_t
>> **end if**
> **until** true

ϵ_t, η_t. These ϵ_t, η_t are drawn from an Extinct Gaussian distribution, defined in Algorithm 1. The parameter p (called extinction rate) controls how often a point drawn form a ball $B(0, r)$ dies off. According to Algorithm 1, the probability of seeing a point inside the ball $B(0, r)$ is different than for a two dimensional Gaussian $N(\mathbf{0}, Id)$. On the other hand, as p goes to zero, the Extinct Gaussian converges in distribution to $N(\mathbf{0}, Id)$. Figure 1 illustrates the joint distribution of X_t, Y_t. The left scatter plot in Figure 1 presents X_t and Y_t generated with an extinction rate of 50%, while the right hand plot is generated with an extinction rate of 99.87%. Processes used in this experiment had an autoregressive component of 0.2, and the radius of the innovation process was 1.

Figure 2 compares the power of the Shift HSIC test and the correlation test. The X axis represents an extinction rate, while the Y axis shows the true positive rate. Shift HSIC is capable of detecting non-linear dependence between X_t and Y_t, which is missed by linear correlation. The red star depicts performance of the KCSD algorithm developed by Besserve et al. (2013), with parameters tuned by its authors: note that this result required using four times as many data points as HSIC.

False positive rates. We next investigate the rate of false positives for Shift HSIC and Bootstrap HSIC on independent copies of the $AR(1)$ processes used in the previous experiment. To generate independent processes, we first sampled two pairs (X_t, Y_t), (X_t', Y_t') of time series using (6), and then constructed Z by taking X from the first pair and Y from the second, i.e., $Z_t = (X_t, Y_t')$. We set an extinction rate to 50%. [5] The AR component a in the model (6) controls the memory of a processes - the larger this component, the longer the memory. We performed the Shift HSIC and the Bootstrap HSIC tests on Z_t generated under $\mathbf{H_0}$ with different AR components. Figure 3 illustrates the results of this experiment. The X axis is indexed by the AR component and Y axis shows the FP rate. As the temporal dependence increases, the Bootstrap HSIC incorrectly gives an increasing number of false positives: thus,

---

[4]As a illustration, consider $W_t = Y_{t-10}$. If Y_t is stationary then the dependence structure of (W_{t_1}, W_{t_2}) and (Y_{t_1}, Y_{t_2}) is the same. If we set $W_t = Y_{\pi(t)}$ this property does not hold.

[5]As a reviewer pointed out, the example for the FP rates can be simplified, however we decided to be consistent with the marginal distribution of X_t, Y_t across the experiments.
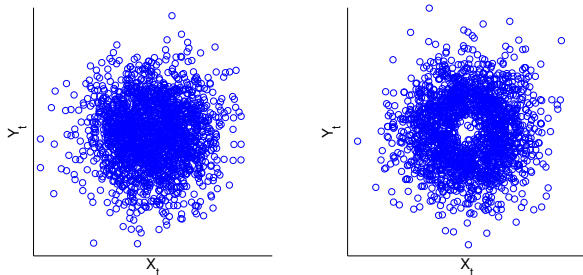
Figure 1. X_t and Y_t, described in the Experiment 4.1, with extinction rates 50% (left) and 99.8% (right), respectively.
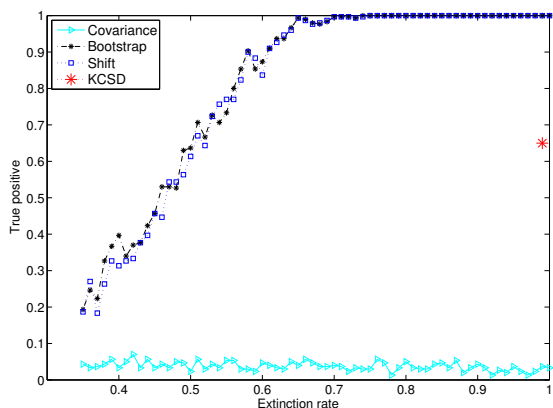

Figure 2. True positive rate for the Shift HSIC, the Bootstrap HSIC and correlation based test: sample size 1200, results averaged over 300 repetitions. The red star shows KCSD performance at $4\times$ the HSIC sample size; see Section 4.1 for details.

it cannot be relied on to detect dependence in time series. The Shift HSIC false positive rate remains at the targeted 5% p-value level.

4.2. Forex data

We use Foreign Exchange Market quotes to evaluate Shift HSIC performance on the real life data. Practitioners point out that forex time series are noisy and hard to handle, especially at low granulations (smaller then 15 minutes). We decided to work with forex time series to show that Shift HSIC can detect dependence even on such a difficult dataset. The forex time series were granulated to obtain two minute sampling (the granulation function returned the last price in the two minute window). Using the test of Diks & Panchenko (2005), we checked that serial dependence of the differentiated time series decays fast enough to satisfy the assumed mixing conditions (by a differentiated time series, we refer to $(X_t - X_{t-1})_{t\in\mathbf{N}}$). The choice of the pairs and trading day (21st January 2013) were arbitrary.

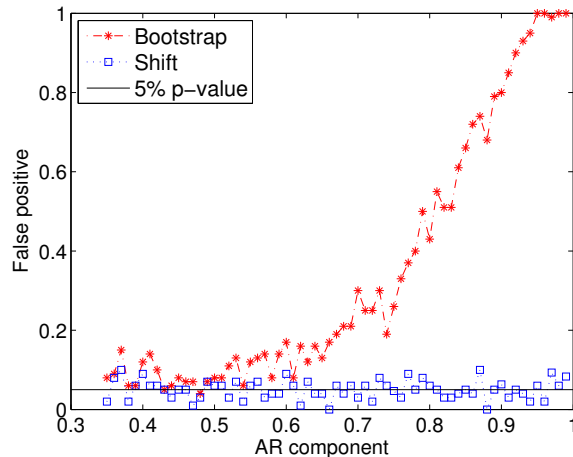
Figure 3. False positive rate for the Shift HSIC and the Bootstrap HSIC. The sample size was 1200, and results were averaged over 300 repetitions.

Instantaneous coupling and causal effect. Having one Australian dollar we may obtain a quantity of Yen in two ways, either by using AUD/JPY exchange rate explicitly or by buying Canadian dollars and then selling them at the CAD/JPY rate. Let X_t be a differentiated AUD/JPY exchange rate and Y_t be a differentiated product of exchange rates AUD/CAD$\times$CAD/JPY. We will investigate the relation between these two. Common sense dictates that Y_t should behave similarly to X_t. After examining the cross-correlation of X_t and Y_t, we propose a simple regression model to describe the interaction between the signals,

$$\hat{Y}_t = a_0 X_t + a_1 X_{t-1} + \cdots + a_6 X_{t-6}.$$

We fit the model and see that $a_0 = 0.97$, and the remaining coefficients are not bigger then 0.06 in absolute value. This suggest that most of the dependence is explained by an instantaneous coupling. We further investigate the cross-correlation between residuals $R_t = Y_t - \hat{Y}_t$ and X_t. We observe no significant correlations in the first 30 lags.

Next we investigate dependence of residuals with lagged values of the explanatory variables, i.e., R_t with X_{t-k} for $k \in (0, \cdots, 30)$. After calculating p-values using the Bootstrap HSIC and the Shift HSIC, we discover dependence only at lags 4, 5, 9, 13 and 29, as presented in the Figure 4. Lack of the dependence at lag zero suggests that the linear model for coupling is reasonable. However, both the Bootstrap HSIC and the Shift HSIC support the hypothesis that there is a strong relation at lag 5, which is not explained well by the linear model.

The questions remains whether test statistics at lags 4, 9, 13 and 29 indicate further model misspecification. Under $\mathbf{H_0}$, at a significance level 94%, we expect 1.8 out of 30 statis-
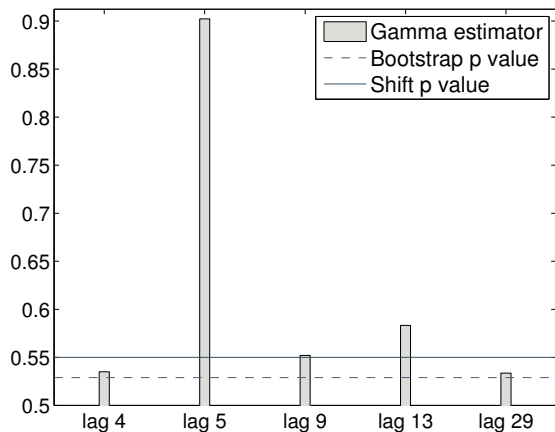
Figure 4. Instantaneous coupling. Results for 720 samples, null threshold of Shift HSIC used 300 lags in range $100 - 400$.

tics to be higher than the 94th percentile. Excluding the statistic at lag 5, the Shift HSIC test reports two statistics above this percentile, while Bootstrap HSIC reports four. Should the statistics at the different lags be independent from each other, the probabilities of seeing two and four statistics above the percentile are respectively 25% and 6%. Shift HSIC indicates that the model fits the data well, while the Bootstrap HSIC suggests that some non-linear dependencies remain unexplained.

Dependence structure. The data are five currency pairs. A correlation based independence test, and the Shift HSIC test, were performed on each pair of currencies. The dependencies revealed by these tests are depicted in Figure 5 - nodes represent the time series and edges represent dependence. Shift HSIC reveals a strong coupling between EUR/RUB and USD/JPY, HKD/JPY and XAU/USD that was not found by simple correlation. All edges revealed by Shift HSIC have p-values at most at level 0.03 - clearly, the Shift HSIC managed to find a strong non-linear dependence. Note that the obtained graphs are cliques.

5. Proofs

A U-statistic of a k-argument, symmetric function f, is written

$$U(f, Z) = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} f(Z_{i_1}, ..., Z_{i_k}).$$

A decomposition due to Hoeffding allows us to decompose this U-statistic into a sum of U-statistics of canonical functions, $U(h, Z) = \sum_{k=1}^{l} \binom{l}{k} U(h_k, Z)$, where $h_k(z_1, ..., z_l)$ are components of the decomposition. According to Serfling (2009, section 5.1.5), each of h_1, h_2, h_3, h_4 is symmetric and canonical. Note that h_k is defined using indepen-
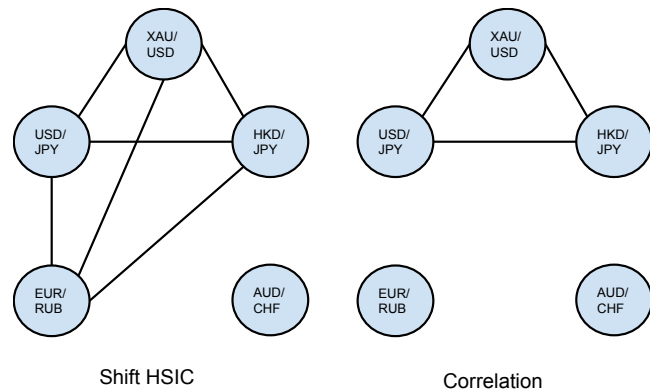

Figure 5. Differences between the dependence structure on the forex revealed by the Shift HSIC and covariance. Parameter settings are as in Figure 4.

dent samples Z^* - this is because the CLT or LLN state that U-statistics or V-statistics of mixing processes converge to their expected value taken with respect to independent copies, i.e., Z^*. Under $\mathbf{H}_0$, h_1 is equal to zero everywhere and $h_2 = \frac{1}{6} s$, where these results were obtained by Gretton et al. (2007).[6] See supplementary material for details.

In order to characterize $U(h, Z)$, we show that under null hypothesis $U(h_2, Z)$ converges to a random variable, and both $U(h_3, Z), U(h_4, Z)$ converge to zero in a probability (the latter proof can be found in the supplementary material). Bellow we characterise $U(h_2, Z)$ convergence.

Lemma 3. *Under assumptions of Lemma 2,*

$$\lim_{n \to \infty} n \cdot U(h_2, Z) \overset{D}{=} \frac{1}{6} \sum_{i_1}^{\infty} \lambda_{i_1} (\tau_{i_1}^2 - 1).$$

Proof. First recall that under null hypothesis $h_2 = \frac{1}{6} s$. We will check the conditions of (Borisov & Volodko, 2008, Theorem 1) (also available in the supplementary).
First, from Mercer's Theorem (Steinwart & Scovel, 2012, Corollary 3.5), we deduce that the h_2 coefficients in $L_2(\mathbf{Z}, \mathcal{B}_\mathbf{Z}, P_\mathbf{Z})$ are absolutely summable. In the supplementary material we show that $\mathcal{E} e_i(Z_1^*) = 0$.
Recall the assumptions of Lemma 2. If **A** holds then $\sum_{k=1}^{\infty} \phi(k)^{\frac{1}{2}} < \infty$ and $\sup_i \mathcal{E} |e_i(X_1)|^2 = 1 < \infty$ (e_i is an orthonormal eigenfunction). Finally, if **B** holds then the process Z_t is α-mixing. The remaining assumptions concerning uniform mixing in Borisov & Volodko (2008) are exactly the same as in this lemma.
□

5.1. Main body proofs

Proof. (Lemma 2) We use the fact that h_2 is equal to s up to scaling ($6U(h_2, Z) = U(s, Z)$), and Lemma 3, to

---

[6]The second result is hard to locate - it is in appendix A.2, text between equations 12 and 13

see that $nU(s,Z) \xrightarrow{D} \sum_i^\infty \lambda_i(\tau_i^2 - 1)$. Since $\mathcal{E}s(Z_t, Z_t) = \mathcal{E}\sum_{i=1}^\infty \lambda_i e_i(Z_t)^2 = \sum_{i=1}^\infty \lambda_i$, then by the LLN for mixing processes,

$$\lim_{n\to\infty} \frac{1}{n}\sum_i^n s(Z_i, Z_i) \overset{P}{=} \sum_{i=1}^\infty \lambda_i. \tag{7}$$

We use a relationship between U and V statistics,

$$\lim_{n\to\infty} nV(s,Z) \overset{D}{=} \lim_{n\to\infty} nU(s,Z) + \lim_{n\to\infty} \frac{1}{n}\sum_i^n s(Z_i, Z_i)$$
$$\overset{D}{=} \sum_{i=1}^\infty \lambda_i + \sum_i^\infty \lambda_i(\tau_i^2 - 1) \overset{D}{=} \sum_i^\infty \lambda_i \tau_i^2.$$

$\square$

Proof. (Theorem 1) We operate under the null hypothesis. Recall that $U(h,Z)$ can be decomposed as $U(h,Z) = \sum_{k=1}^4 \binom{4}{k} U(h_k, Z)$. Here $h_1 \equiv 0$. We show in the supplementary material that $U(h_3, Z)$ and $U(h_4, Z)$ tend to zero in probability. From Lemma 3,

$$\lim_{n\to\infty} nU(h,Z) \overset{D}{=} \lim_{n\to\infty} nU(s,Z) \overset{D}{=} \sum_i^\infty \lambda_i(\tau_i^2 - 1). \tag{8}$$

We define an auxiliary symmetric function w,

$$w(z_1, z_2, z_3) = h(z_1, z_1, z_2, z_3) + h(z_1, z_2, z_2, z_3)$$
$$+ h(z_1, z_2, z_3, z_3) + h(z_1, z_1, z_3, z_2)$$
$$+ h(z_3, z_2, z_2, z_1) + h(z_2, z_1, z_3, z_3).$$

It is obvious that $\mathcal{E}w(Z_1^*, Z_2^*, Z_3^*) = 6\mathcal{E}h(Z_1^*, Z_1^*, Z_2^*, Z_3^*)$. We consider the difference between the unnormalized V and U statistics,

$$S_n = \sum_{1 \le i_1, i_2, i_3, i_4 \le n} h(Z_{i_1}, ..., Z_{i_4}) - \sum_{i \in C_4} h(Z_{i_1}, ..., Z_{i_4}),$$

where $\sum_{i \in C_m}$ denotes summation over all $\binom{n}{m}$ combinations of m distinct elements $\{i_1, \cdots, i_m\}$ from $\{1, \cdots, n\}$. The difference is equal to the sum over 4-tuples with at least one pair of equal elements. We can choose such tuples in $\binom{4}{2} = 6$ ways. Observe that w covers the choice of all these six tuples. Since for any $z_1, z_2 \in \mathbf{Z}$, $h(z_1, z_1, z_1, z_2) = 0$, then w is zero whenever more than two indices are equal. Therefore we can sum w over distinct indices z_1, z_2, z_3,

$$S_n = \sum_{i \in C_3} w(Z_{i_1}, Z_{i_2}, Z_{i_3}).$$

We see that S_n is almost a U-statistic ($U(w, Z)$). By the CLT for U-statistics from Denker & Keller (1983), Theorem 1(c), we obtain

$$\lim_{n\to\infty} \frac{1}{n(n-1)(n-2)} S_n \overset{P}{=} 6\mathcal{E}h(Z_1^*, Z_1^*, Z_2^*, Z_3^*).$$

On the other hand, via the relation $h_2 = \frac{1}{6}s$ and the h_2 definition, we get $\mathcal{E}s(Z_1^*, Z_1^*) = 6\mathcal{E}h(Z_1^*, Z_1^*, Z_2^*, Z_3^*)$, and therefore

$$\lim_{n\to\infty} \frac{1}{n(n-1)(n-2)} S_n \overset{P}{=} \lim_{n\to\infty} \frac{1}{n}\sum_i^n s(Z_i, Z_i). \tag{9}$$

Finally, we rewrite S_n as

$$\sum_{1 \le i_1, i_2, i_3, i_4 \le n} h(Z_{i_1}, ..., Z_{i_4}) = S_n + \sum_{i \in C_4} h(Z_{i_1}, ..., Z_{i_4}).$$

We normalize by $\frac{1}{n(n-1)(n-2)}$, and take the limit in n,

$$\lim_{n\to\infty} \frac{n^4}{n(n-1)(n-2)} V(h,Z) \overset{D}{=}$$
$$= \lim_{n\to\infty} \left(\frac{1}{n(n-1)(n-2)} S_n + (n-4)U(h,Z) \right).$$

We substitute (9) and (8) on the right hand side, and use equation (7) from Lemma 2 to replace $\lim_{n\to\infty} \frac{1}{n}\sum_i^n s(Z_i, Z_i)$ with $\sum_{i=1}^\infty \lambda_i$, yielding

$$\lim_{n\to\infty} n \cdot V(h,Z) \overset{D}{=}$$
$$\overset{D}{=} \lim_{n\to\infty} \frac{1}{n}\sum_i^n s(Z_i, Z_i) + \lim_{n\to\infty} \frac{1}{n}\sum_{i,j}^n s(Z_i, Z_j) \overset{D}{=}$$
$$\overset{D}{=} \sum_{i=1}^\infty \lambda_i + \sum_i^\infty \lambda_i(\tau_i^2 - 1) \overset{D}{=} \sum_i^\infty \lambda_i \tau_i^2.$$

$\square$

Proof. (Theorem 2) If the null hypothesis does not hold, then $\gamma > 0$ (Gretton et al., 2005). In this case h is nondegenerate, and we can use Denker & Keller (1983, Theorem 1(c)) to see that $\frac{\sqrt{n}}{4} \frac{\sqrt{n}}{\sqrt{\sigma}}(V(h,Z) - \gamma) \sim N(0,1)$, where σ is finite (see the note below Theorem 1 of (Denker & Keller, 1983), stating that in case (c) σ^2 is finite, and the note above Theorem 1 stating that $\sigma^2 = \lim_{n\to\infty} n^{-1}\sigma_n^2$). $\square$

Proof. (Lemma 1) We use Lemma 1 and Theorem 4 from Gretton et al. (2005) to show that $\mathcal{E}h(Z_1^*, Z_2^*, Z_3^*, Z_4^*) = 0$ iff (X_1^*, Y_1^*) has a product distribution. Since $Z_1^* \overset{D}{=} Z_1$ and $Z_t \overset{D}{=} Z_1$, we infer that X_t is independent from Y_t iff $\mathcal{E}h(Z_1^*, Z_2^*, Z_3^*, Z_4^*) = 0$. $\square$

References

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Besserve, M., Logothetis, N., and Schölkopf, B. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*, pp. 2535–2543, 2013.

Borisov, I. and Volodko, N. Orthogonal series and limit theorems for canonical U- and V-statistics of stationary connected observations. *Siberian Advances in Mathematics*, 18(4):242–257, 2008.

Bradley, R. et al. Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys*, 2(107-44):37, 2005.

Broock, W., Scheinkman, J., Dechert, D., and LeBaron, B. A test for independence based on the correlation dimension. *Econometric Reviews*, 15(3):197–235, 1996.

Denker, M. and Keller, G. On U-statistics and v. Mises statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64(4):505–522, 1983.

Diks, C. and Panchenko, V. Nonparametric tests for serial independence based on quadratic forms. Technical report, Tinbergen Institute Discussion Paper, 2005.

Doukhan, P. Mixing. properties and examples. In *Mixing*, number 85 in Lect. Notes in Stat., pp. 87–109. Springer, January 1994.

Fadlallah, B., Brockmeier, A., Seth, S., Li, L., Keil, A., and Principe, J. An association framework to analyze dependence structure in time series. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pp. 6176–6179. IEEE, 2012.

Feuerverger, A. A consistent test for bivariate dependence. *International Statistical Review/Revue Internationale de Statistique*, pp. 419–433, 1993.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS 20*, pp. 489–496, 2008.

Granger, C., Maasoumi, E., and Racine, J. A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*, 25(5):649–669, 2004.

Gretton, A. and Györfi, L. Consistent nonparametric tests of independence. *JMLR*, 11:1391–1423, 2010.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic learning theory*, pp. 63–77. Springer, 2005.

Gretton, A., Fukumizu, K., Teo, C, Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. In *NIPS*, volume 20, pp. 585–592, 2007.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. A kernel two-sample test. *JMLR*, 13:723–773, 2012.

Haugh, L. Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71 (354):378–385, 1976.

Hong, Y. Testing for independence between two covariance stationary time series. *Biometrika*, 83(3):615–625, 1996.

Hong, Y. and White, H. Asymptotic distribution theory for nonparametric entropy measures of serial dependence. *Econometrica*, 73(3):837–901, 2005.

Hyvärinen, A., Karhunen, J., and Oja, E. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Lyons, R. Distance covariance in metric spaces. *The Annals of Probability*, 41(5):3051–3696, 2013.

Pearl, J. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

Pinkse, J. A consistent nonparametric test for serial independence. *Journal of Econometrics*, 84(2):205–231, 1998.

Racine, J. and Maasoumi, E. A versatile and robust metric entropy test of time-reversibility, and other hypotheses. *Journal of Econometrics*, 138(2):547–567, 2007.

Robinson, P. Consistent nonparametric entropy-based testing. *The Review of Economic Studies*, 58(3):437–453, 1991.

Rosenblatt, M. and Wahlen, B. A nonparametric measure of independence under a hypothesis of independent components. *Statistics & probability letters*, 15(3):245–252, 1992.

Serfling, R. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Shao, X. A generalized portmanteau test for independence between two stationary time series. *Econometric Theory*, 25(01): 195–210, 2009.

Smola, A. J, Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory*, volume LNAI4754, pp. 13–31, Berlin/Heidelberg, 2007. Springer-Verlag.

Song, L, Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *JMLR*, 13: 1393–1434, 2012.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. Universality, characteristic kernels and RKHS embedding of measures. *JMLR*, 12:2389–2410, 2011.

Steinwart, I. and Scovel, C. Mercers theorem on general domains: on the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35(3):363–417, 2012.

Székely, G. J, Rizzo, M. L, et al. Brownian distance covariance. *The annals of applied statistics*, 3(4):1236–1265, 2009.

Zhang, X., Song, L., Gretton, A., and Smola, A. Kernel measures of independence for non-iid data. In *NIPS*, volume 22, 2008.