REU Final Report

Conflict Dynamics in Pigtailed Macaques

Author:
Edward D. Lee

Advisors:
Dr. Bryan Daniels
Dr. David Krakauer
Dr. Jessica Flack

Abstract

We elaborate on previous Inductive Game Theoretic methods on pigtailed macaque conflict behavior. We investigate alternative measures of dependence to enhance previous measurements by (1) capturing other features of the data set potentially neglected by the original measures and (2) resolving the time anomalous observations made with the original measure with the new measure. We compare ΔP, Pearson's correlation coefficient and distance covariance. We find some evidence for temporal variations in these measurements, but these interpretations are confounded by inherent variations from the current structure of the data. We also find evidence for inhibitory connections as a relationship that exists only between higher order groups.

Santa Fe Institute
1399 Hyde Park
Santa Fe, NM 87501

Contents

1 Introduction

By virtue of containing many interacting components, complex systems can show misalignment between these components, leading to frustration. Some have gone so far as to argue that lack of alignment, or "frustration," in many body systems is the defining feature of complex systems [28]. A simple example would be the spin glass model where frustration between triplets of spins is one of the important characteristics that explain the dynamics underlying the system. We might see this frustration manifest in various ways such as stress or conflict. In abiotic systems, we can conceive of conflict occurring in boundaries straddling discontinuous energy gradients such as between tectonic plates. How stress has builds and how it is released is crucial in explaining the evolution of the landscape. Analogously, the creation and dissipation of conflict in social systems is fundamentally linked to the stability and evolution of the system. Nevertheless, the relationship between stress and conflict is not as clear because conflict may not just be a dissipative mechanism; it may build stress [3]. Conflict inherently destabilizes the configuration of the system, and it is of utmost interest to know what the causes of such instability are. Especially in social systems, conflict seems to play a significant role. How conflicts occur, how they propagate and other questions about its underlying dynamics will lead to important insights about the costs and functions of conflict [9][10].

With our social system of interest, we observe conflict as acts of aggression between individuals. We are interested in the how individual strategic decision-making contributes to the pattern of conflict. As Dedeo, Krakauer and Flack (DKF) stated in their paper,

> Our aim is to determine the strategies individuals use when deciding to join fights and their implications for collective conflict dynamics, including statistical patterns at the aggregate level like the distribution of fight sizes. In previous work we have developed Monte Carlo methods for extracting conflict strategies and payoffs, and their implications for robustness and other organizational properties, directly from time series data. We call this approach Inductive Game Theory.[1] In this paper, we further develop this technique, asking whether different measures of dependence and a different representation of data lead to new observations" [8].

Following these investigations, there are several interesting avenues for enhancing the analysis of the time series data. In [8], DKF assumed temporal invariance and aggregated the data set over all distinct observational periods in the four month period to calculate their proxy for individual strategies, ΔP's.[2] If we calculate the ΔP's separately over all periods, however, we find large variance relative to the mean (See Figure 1). We hope to address this question about temporal variance in the individual or coalition strategies in this system.

[1] See Section 7.3 for further explanation.

[2] We define ΔP in Section 3.

When we address this question of whether the strategies are a function of time (or $\Delta P(t)$?), we must address several parts of this question.

1. Are the ΔP's an adequate measure for capturing the temporal dependence?[3] We know that the ΔP's are sufficient to reproduce the conflict size distribution, but we do not know whether different measures might provide us with more stability between days. This question is specific to advantages and disadvantages of the measure we use.

2. Are the temporal variations artifacts of the format of the data? Since we are dealing with sparse and binary conflict data, we may expect that our measures behave in undesired ways.

3. Related to our second question is our third: can we answer the question about temporal stability in the format that DKF used? In other words, do our measures return genuine dependencies in the data? Implicitly, we are also asking about the level of aggregation required such that results are independent of the length of the observational period when generated from a stationary process.

To be able to figure out what rules macaques implement that result in the observed pattern will be a step to understanding the mechanics of conflict and the best methods of investigating them.

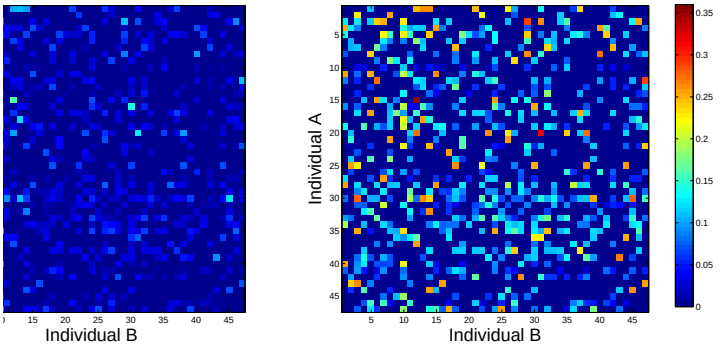
2 Format of data

Our data set records the conflict behavior of a group of 47 pigtailed macaques, Old World primates in a compound at Yerkes. Although there are 84 individuals, we only consider the 48 that are socially mature (subadults or adults), but remove one from our analysis because she died during the course of the study. DKF considered all 48 [8].

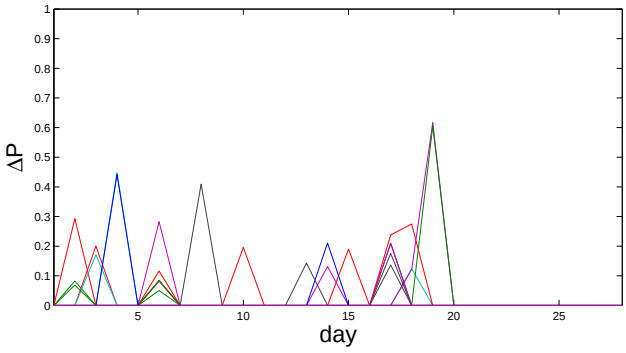
The data consists of a binary array labeling participation or non-participation with ones and zeros associated with individual identity as well as location of the fight in the time series. Therefore, we measure time in bouts and not real time. A fight consists of any interaction in which one individual threatens or aggresses a second individual. A conflict was considered terminated if no aggression or withdrawal responses (fleeing, crouching, screaming, running away, submission signals) was exhibited by any of the conflict participants for two minutes from the last such event [7].

For the time series analysis, we split the data set into days and excluded the participation in fights where the membership of a previous fight could not be determined because of observational breaks. For some of the extended analyses, we excluded observational days that were intended for control purposes to acquire observations in a period of time that was under-represented in the rest of the data set.

[3]Here, we define dependency to be some relationship between two random variables such that the behavior of one can be predicted from the behavior of the other.



(a)



(b)

Figure 1: **(a)** ΔP array between all 47 individuals with means over all observational periods with more than 45 fights on the left. There are 14 such periods out of the 28. On the right, standard deviation over the same periods. Standard deviation is large relative to the mean. **(b)** Changes in Eo→Eo ΔP values over all observational periods. Nulls for both plots are 10,000 time series permutations.

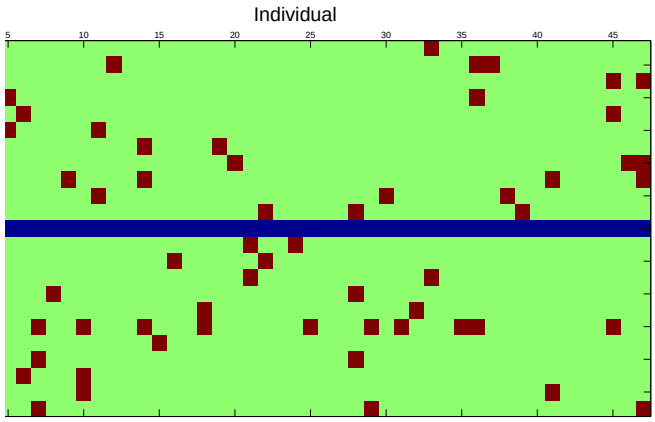


Figure 2: Example of binary data. Green marks non-participation, brown participation and blue a break in data collection.

3 Statistical Measures

Here, we redefine the ΔP measure as

$$\Delta P = \frac{N(B|A)}{N(A)} - \frac{N_{\text{null}}(B|A)}{N_{\text{null}}(A)}. \tag{1}$$

This redefinition from DKF is necessary because the number of null fights is variable according to the alignment of the fights with respect to the location of the breaks. When we shuffle the data for the null measurements, we fix the locations of the breaks so it is possible for some fights, which were not previously adjacent, to appear adjacent to breaks. As we mentioned earlier, we must discount participation in a fight in our analysis if there is no previous fight from which to calculate a ΔP. Thus, the total number of fights fluctuates with the shuffles such that we must normalize the null differently to remain consistent. See appendix for more details (Section 7.2).

As we mentioned above, we find the ΔP's are limited to correlations among frequencies of individuals or coalitions. Naturally, we look for a measure that may be more general and provide some comparison. Although ΔP's are easily computable and useful measures in that they distinguish between positive and negative correlations—making them particularly apt for the excitatory/inhibitory model—it is probable that there are more complicated, nonlinear relationships in the data.[4] Since they are tuned to measure a specific kind of relation, they will fail to detect a significant relationship such as a parabola

[4]Measure is a specific mathematical term to refer to a mapping of a set of points to the real line. In our case, measure is closely related to probability because we are trying to assign a real number between 0 and 1 to how well an individual predicts another individual's behavior. Mathematically, we take two sets and assign them a real number $f : M \times N \to \mathbb{R}$.

(that is half increasing and half decreasing). As a more general measure, we worked with Distance Covariance (DCOV). Before addressing this issue, we discuss some measure theory.

3.1 Dependence

Predicting the behavior of a random variable from a black box system is a difficult problem. If we are looking for a simple pattern such as a correlation, the measures and methods that we use are quite established: as X decreases or increases, does Y increase or decrease? In more complicated systems, however, we may be looking for a more general kind of relation that we call dependence—some relationship between X and Y that may be nonlinear. In a more general context, X and Y may not be of the same dimension.[5] Since the nature of this dependence could be quite general (whether it be quadratic, exponential or some other complicated function), one approach is to test for independence. Probability theory contains the statement that for two random variables to be independent, we must find that their joint probability is equal to the product of their marginals

$$P(X, Y) = P(X)P(Y).$$

If this equality does not hold, then the variables are dependent in some way [12].[6] This statement declares that looking at the joint distribution of X and Y does not tell you more about the distributions of X and Y as would looking at their separate distributions. This statement leads to the trivial restatement that we want our distance measure to somehow be a function such that

$$\delta(X, Y) = 0 \quad \text{iff } P(X, Y) - P(X)P(Y) = 0. \tag{2}$$

If we find function that follows this rule, then we can claim to have found some reasonable mapping of dependence to the real line.

In the early 20th century, Rényi, a mathematician declared eight axioms for marking a good dependence measure [18].[7] Let the dependence measure be δ and the set of inputs and their corresponding outputs be $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$,

[5] For example, if for every defined input, we obtain realizations from a probability distribution of the output such that each one-dimensional X corresponds to an N-dimensional Y.

[6] Notation can be a real trickster, so we better clear up some possible sources of confusion. The italicized and capital letters (X, Y) refer to the true distribution of some random variable whereas the bold $(\mathbf{X}, \mathbf{Y})$ to the set of finite realizations of the underlying distributions. When denoted with a subscript, (X_i, Y_i), we refer the i^{th} realization of $(\mathbf{X}, \mathbf{Y})$. An apostrophe such as X' or $\mathbf{X}'$ refers to an IID sample of the same size as X or $\mathbf{X}$.

[7] The following axioms have been generalized to a multiple variables and not just two as he originally proposed. Rényi's axioms assume that our observations of the random variable consist of samples that are independent and from an identical underlying distribution. We refer to this assumption as IID, which is a specific form of exchangeability. Exchangeability refers to the case that all samples come from the same underlying distribution such that any sequence is equally likely. Since they are all equally likely, we may exchange any two realization in the sequence without changing the probability of obtaining that sequence.

1. $\delta(Z)$ is defined for any random vector Z when it is not a constant with probability 1.

2. $\delta(X, Y) = \delta(Y, X)$. The more general case is as follows: For any permutation $\sigma = (i_1, i_2, ..., i_n)$ of the indices, we have equality such that $\delta(Z) = \delta(X_1, Y_1, X_2, Y_2, ..., Y_n) = \delta(X_{i_1}, X_{i_1}, X_{i_2}, Y_{i_2}, ..., Y_{i_n})$.

3. $0 \leq \delta(Z) \leq \gamma$, where $\gamma \geq 0$ could be $+\infty$. A stricter condition that Rényi originally postulated was that $0 \leq \delta(Z) \leq 1$.

4. $\delta(Z) = 0$ iff X and Y are independent.

5. $\delta(Z) = \gamma$ iff there is strict dependence between the elements of X and Y. So, for any i, we must have that $X_i = g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$ with probability 1 and where g_i is a real, measurable function. This is also saying that the probability measures $P(X, Y)$ and $P(X)P(Y)$ are singular.

6. If we map the random variable onto the real axis in a one-to-one way, the measure should not change. So, for the one-to-one transformation $T = (T_1, T_2, ..., T_n)$ of Z onto $\mathbb{R}^n$, we have

$$\delta[T(Z)] = \delta(T_1(Z_1), T_2(Z_2), ..., T_n(Z_n)) = \delta(Z).$$

7. If Z has a bivariate normal distribution, then $\delta(Z)$ is a strictly increasing function of $|\rho(Z)|$, where ρ is the Pearson correlation coefficient.

8. Micheas et al. [18] also talk about this further axiom related to the strict sub-additivity of Shannon entropy.

These rules seem reasonable methods for deciding on a dependence measure, but hold for few measures used today. The product-moment correlation coefficient fails to satisfy Axiom 4, and it is one of the most deeply investigated measures. Only its absolute value remains within the $[0, 1]$ bounds according to Axiom 3, but a distinction between positive and negative correlation may be important. Many other measures fail on one or more rules, but their failure to follow these rules does not necessarily diminish their utility. Truly general measures of dependence can be powerful in eliminating the independence hypothesis, but it will not specify the kind of dependence that exists in the data. The one measure that fulfilled all these criteria during Rényi's time was mean-square contingency.

Rényi's axioms say nothing about the interpretation of a measure of dependence between the bounds and for informing us about the nature of dependence. In general, both kinds of interpretation are unavoidable difficulties with general dependence measures because they must account for many different kinds of dependence no matter how the relation may be defined. After all, we compressing

an entire functional relationship between two random variables onto half of a one dimensional real line.[8]

For a good dependence measure, we could specify more conditions that would be useful [27].

1. $\hat{\delta}(\mathbf{Z})$ should not depend on any free parameters, a logical extension enforcing consistence between different instantiations of the same underlying distribution.

2. We should be able to establish an order of dependence amongst the variables. Although Rényi's axioms state that $\delta(X,Y) = \delta(Y,X)$, we may encounter a situation where the relationship is non-invertible so that $\delta(X,Y) \geq \delta(Y,X)$ where the equality holds only if f where $Y = f(X)$ is invertible. Such non-invertibility may be a proxy for the direction of causation. For example, if from a pair of two distributions, one can be recovered from the other, but not the converse, we can make some argument about unidirectional information flow such as what happens when moving from the sine function to its inverse. One measure with this asymmetry is mutual information.

3. Furthermore, we should restrict the functional relationship to an appropriate set of functions to avoid the problem with overfitting a finite number of realizations. We can do this by making our measure represent closeness to some appropriate set of functions such as a line, monotonic curve or some other smooth, continuous curve.

Since we will use the correlation coefficient and distance covariance in this paper, we define these two.

3.2 Correlation

Correlation, ρ, is a linear relationship between two distributions. The Pearson correlation coefficient is defined as

$$\rho = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

$$= \frac{\langle (x-\bar{x})(y-\bar{y}) \rangle}{(\langle x^2 \rangle - \bar{x}^2)(\langle y^2 \rangle - \bar{y}^2)} \tag{4}$$

[8] Another problem with these postulates is that Rényi worked with random variables but not their realizations. Since we never have an infinite number of measurements to work with, we face the problem that we can always construct a function that perfectly fits the finite data set. We have no regular method for assessing estimators, but we probably can rely on intuition and previous knowledge for restricting the class of estimators we can use for the data at hand. This note also brings up the thorny issues of estimators. We denote an estimator to $\delta(Z)$ as $\hat{\delta}(\mathbf{Z})$ to show explicitly how the estimator is a function of the finite realizations of Z and not a perfect measure of the underlying distribution.

where both the bar and brackets denote means. It is the same as the measure of autocorrelation when X and Y are the same data at different time points. We can also re-express the numerator as

$$\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle , \tag{5}$$

which makes clear its relationships with independence although the joint probability and marginal probabilities are weighted by the values of X and Y. We can interpret this measure in several different ways as noted in [25]. See this reference for many interesting interpretations of the correlation coefficient. However, this measure is not robust to outliers, and it is also possible to remove correlation with a linear transformation. Note that if we have two distributions that we scale with some real positive number ϵ,

$$\lim_{\epsilon \to 0} \rho(\epsilon X, \epsilon Y) = \frac{\mathrm{cov}(\epsilon X, \epsilon Y)}{\sqrt{\mathrm{var}(\epsilon X)\mathrm{var}(\epsilon Y)}} = 0. \tag{6}$$

We define

$$\Delta\rho = \rho - \rho_{\mathrm{null}}, \tag{7}$$

where the null is the same time series shuffle as used for ΔP.

We introduce the correlation coefficient for comparison with the ΔP's. Since the next measure that we introduce, distance covariance, is not signed and is general, we do not know what kind of functional relationships it captures and so the $\Delta\rho$'s provide a method of comparison.

3.3 Distance covariance

Our goal here is to find a measure that will demonstrate to us in an easily interpretable result and using a computationally simple process whether two random variables are independent or related by some function. Ideally, this measure should not be specific to monotonic functions (such as Spearman's rank coefficient) or linear functions (Pearson's product-moment correlation coefficient), but should generalize to any kind of dependence [33]. Furthermore, abiding to some of the principles laid out in the above section, this measure should be scale-free according to some subset of transformations that preserve the subspaces of X and Y.

As we mentioned above, dependence is some function of the difference between the joint probabilities and product of marginals of two random vectors.

$$\delta(X, Y) = F(P(X, Y) - P(X)P(Y)).$$

Instead of working with these distributions directly, however, we move to the complex plane of these functions for mathematical convenience. We define

the characteristic function of some distribution as

$$f_X : \mathbb{R}^p \to \mathbb{C}^p;$$
$$f_X(t) = E[e^{i\langle t,x\rangle}] = \int_{\mathbb{R}^p} e^{i\langle t,x\rangle} dF_X(x) \left(= \int_{\mathbb{R}^p} e^{i\langle t,x\rangle} f_X(x)dx \right). \quad (8)$$

This series of equations states several things. First, the characteristic function maps the distribution from the $\mathbb{R}^p$, p-dimensional real space, to the complex space. Since frequencies exist in the complex space, we come to the third equation that shows the characteristic function to be the Fourier transform of the density function. Then, t is analogous to what we think of as the frequencies that underly the distribution of X. However, the probability density function does not exist if the cumulative distribution function is not differentiable (thus, the parentheses). The characteristic function contains the same information as the CDF.

Szekély, Rizzo and Bakirov (2007) proposed a measure of the form

$$\mathcal{V}^2(Z) = \int_{\mathbb{R}^{p+q}} |f_{XY}(t,s) - f_X(t)f_Y(s)|^2 \omega(t,s)dtds; \quad (9)$$

that is, some weighted norm of the distance of the joint characteristic function from the product of the marginal characteristic functions. The variables t and s are real-valued vectors of the same dimensions of X and Y, which are p and q, and $\omega(t,s)$ is some weight function.

Eq 9 is one way of measuring dependence that looks at the squared distance between the joint and marginal characteristic functions. The innovation that SRB discovered was a weight function that returns a value of 0 iff the squared distance is zero as well as remaining invariant to transformations in X and Y. SRB discuss the discovery of such a weight function by drawing on the following lemma as cited in Section 7.4. Other weight functions may lead to different measures, but only weight functions of this type lead to distance covariance type measures [31]. This weight function guarantees the following scale invariance property. This measurement is analogous to covariance.

SRB define the measure normalized to 1 as distance correlation:

$$\mathcal{R}^2 = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0; \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \quad (10)$$

It can be shown that this equation will reduce to a deterministic function of $|\rho|$ in the bivariate normal distribution.

An alternative formulation of this measure for computational purposes is

$$\mathcal{V}^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}B_{kl}, \quad (11)$$

where A_{kl} is the centered pairwise distance matrices (along columns and rows) of X and B_{kl} for Y. Then, $\mathcal{V}^2$ is the elementwise product of these two centered, distance matrices. This is surprisingly simple to compute [19]. Another formulation of distance covariance from an intermediate step is

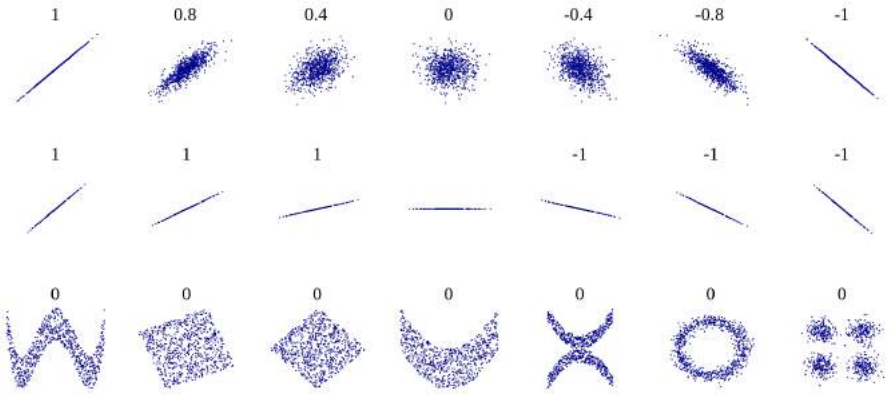


Figure 3: Examples of when Pearson's product-moment correlation coefficient fails and return 0 despite some obvious bivariate relationship. Distance covariance will not return 0 except for the bivariate normal distribution in the middle of the first line. Image from Wikipedia page "Pearson product-moment correlation coefficient."

$$\mathcal{V}^2 = S_1 + S_2 - 2S_3. \tag{12}$$

where

$$S_1 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q, \tag{13}$$

$$S_2 = \frac{1}{n^2} \sum_{k,l=1}^{n} |X_k - X_l|_p \sum_{k,l=1}^{n} |Y_k - Y_l|_q, \tag{14}$$

$$S_3 = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q, \tag{15}$$

This alternative formulation makes explicit the role of the joint distribution which is not apparent in the first formulation in Eq 11.

In summary, the properties of this measure are

1. If $E(|X|_p + |Y|_q) < \infty$, then $0 \leq \mathcal{R} \leq 1$, and $\mathcal{R}(X, Y) = 0$ iff X and Y are independent.
2. $0 \leq \mathcal{R} \leq 1$.
3. If $\mathcal{R}(\mathbf{X}, \mathbf{Y}) = 1$, then there exists a vector a, a nonzero real number b and an orthogonal matrix C such that $\mathbf{Y} = a + b\mathbf{X}C$.
4. For the bivariate normal case, $\mathcal{R}^2$ reduces to a deterministic function of $|\rho|$.

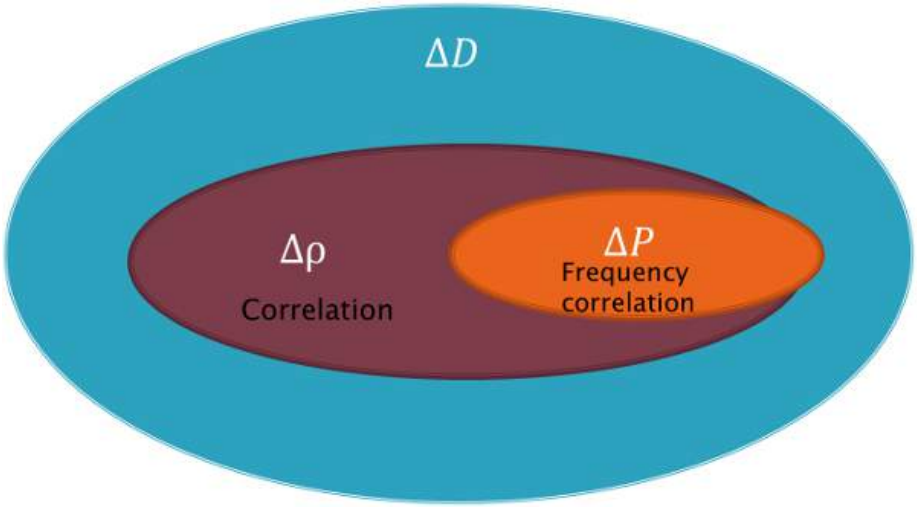


Figure 4: Our expected relationships between the measures. ΔD should be the most general, and thus should capture the most number of patterns. The other measures are subsets of what we should expect to see from the ΔD. Since the ΔP's are similar to frequency correlations, they should mostly agree with $\Delta\rho$'s, but not completely. This graph does not say anything about the expected relationships between the values of these three measures, but rather the number of significant patterns detected by each.

5. $\mathcal{R}^2$ is defined for X and Y of arbitrary dimension.

SRB find that "[i]n Monte Carlo studies, the distance covariance test exhibited superior power relative to parametric or rank-based likelihood ratio tests against nonmonotone types of dependence" [32]. For more examples, refer to [33], [32], and the reviews of [33] published in the same issue.

As with the previous measures, let us define

$$\Delta D = \mathcal{R}^2 - \mathcal{R}^2_{\text{null}}. \tag{16}$$

Again, the null refers to the same time series shuffle as used with the ΔP's.

3.4 Stability Comparisons

We first check to see if these two measures are more stable between the different days. In Figure 5, we see no such improvement. For both $\Delta\rho$ and ΔD, one standard deviation is quite large compared to the mean such that for the $\Delta\rho$, a positive average correlation could be negative within one standard deviation. For ΔD, a significant measure could be insignificant within one standard deviation. These statistics reflect the fact that there are large fluctuations within the data set between observational periods since some monkeys do not appear in some of the observed days such that most pairs have measures of 0 for a majority of the days.

We assume that ΔD is a better measurement of dependence than ΔP because it is not specific to a certain kind of dependence as we represent in Figure 4. However, when we plot the intersections of the different measures solely based on whether they are significant or insignificant for the specific (A, B) pair of individuals, we find a very different picture (Figure 6). To obtain more fine-grained picture, we also show all the measures plotted against each other so that we can visualize the relationships. From the comparisons in Figure 7, it seems that $\Delta\rho$ is a proper reference to compare with ΔD given the nearly linear relationship between them. Given this relationship and our knowledge of the expected relationships between ΔD and $\Delta\rho$, we might be tempted to conclude that ΔD is detecting significant relationships that encompass those made by $\Delta\rho$. We can see, however, from the previous Figure 6 that there is a large body of $\Delta\rho$'s that do not fall under the ΔD umbrella.

Although these initial results are not aligned with the expected behavior, note the shift in intersections between the measures in Figure 6 that occurs when we aggregate or disaggregate the data set into its observational periods. We believe this hints at the true underlying explanation: the current instance of the data causes our measures to behave strangely.

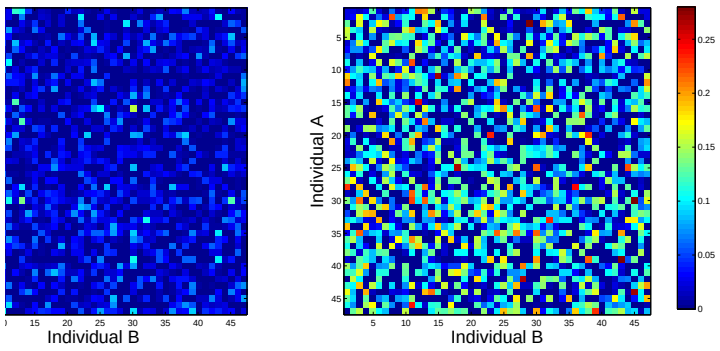
One way we can see how this sparseness and binary representation introduces a large variance is to calculate same measures on a simulated data set. We will take this step in future research.

3.5 Variance in measures

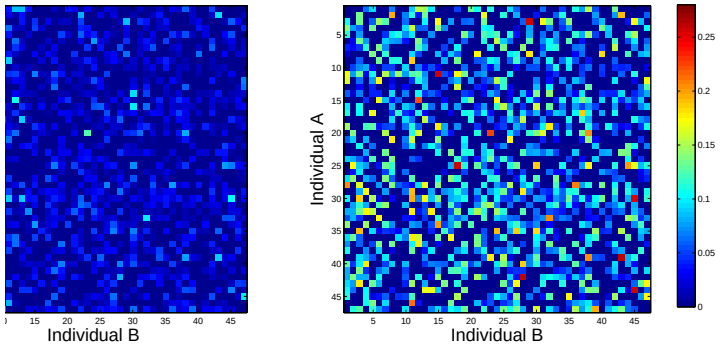
With all these measures, we find that the normalization to one results in an odd scale that gives a strong weight to a rare agreement between adjacent time sequences for which the null does not account (with respect to our intuition for what the measures should tell us). To make this statement explicit, we present an example time series that looks like the following,

$$\begin{array}{cc} A & B \\ & \begin{pmatrix} \vdots & \vdots \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ \vdots & \end{pmatrix} \\ \begin{array}{l} \\ t_{i-1} \\ t_i \\ t_{i+1} \\ \vdots \end{array} & \end{array}$$

such that B appears in fight at t_{i+1} after A, but they do not participate in any other of the n fights that day. To be concrete, let us imagine that the length of the data set is $N = 100$ and $N(A) = N(B) = 3$. Since N is large compared to the amount of participation, we will find that this series returns values of $\Delta P \approx \Delta D \approx \Delta\rho \approx 1$. With the ΔP and $\Delta\rho$'s, we can show that this result is



(a)

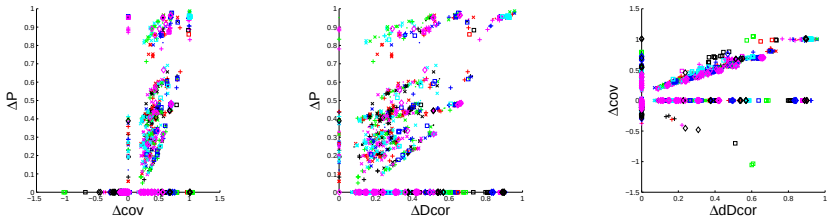


(b)

Figure 5: **(a)** $\Delta\rho$ array between all 47 individuals with absolute values of means over all observational periods with more than 45 fights on the left. There are 14 such periods out of the 28. On the right, standard deviation over the same periods. Standard deviation is large relative to the mean. **(b)** Similar graphs for ΔD.

(a) (b)

Figure 6: **(a)** Proportional representation of significant measures when data is divided into days. Overlaps show how many significant measures they share out of their respective populations. The black circle represents all possible relationships between individuals A and B over all 28 observational periods ($47 \cdot 47 \cdot 28 = 61852$.). The percentage sizes of each of the circles are as follows: $\Delta P = 1.65\%$, $\Delta\rho = 5.37\%$, $\Delta D = 3.46\%$. **(b)** Proportional representation of significant measures when data is aggregated over all observational periods. The black circle represents only $47^2 = 2209$ possible relationships here. The percentage sizes of each of the circles are as follows: $\Delta P = 8.51\%$, $\Delta\rho = 10.41\%$, $\Delta D = 9.64\%$. Surprisingly, these relationships do not match our expected relationships between the measures. This peculiar result may have to do with the structure of our instance of the data set.



(a) (b) (c)

Figure 7: Comparison of measures when data is divided into days. For each measure, there are $47^2 \times 28 = 61582$ data points (since we have 47 individuals and 28 observational periods). However, we do not plot the points where measures of both axes are 0.

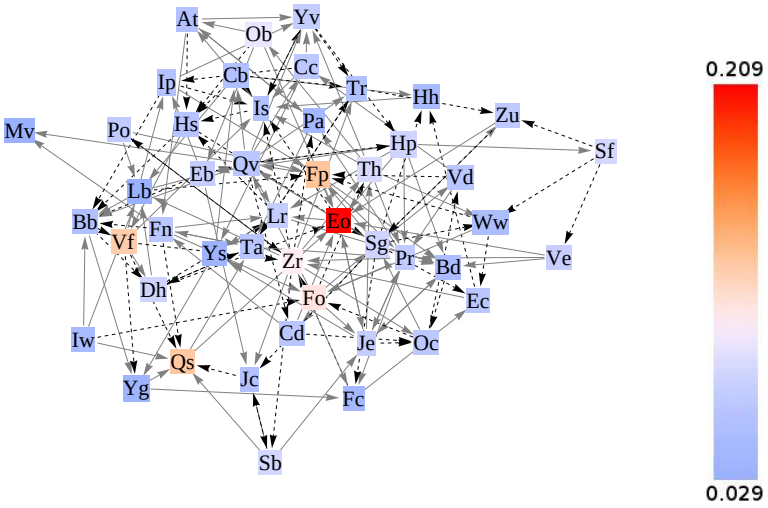


Figure 8: ΔP networks over aggregated time series. Individuals' colors represent fight participation frequency. Solid gray lines represent positive measures and dashed black lines negative measures. The individuals Eo, Ec, Fp and Qs are all policers, macaques with dominant power roles in this group. Since they intervene in many fights, their fight participation rates tend to be unusually high. However, this discrepancy between high participation and low participation individuals does not affect our comparison of measures because the nulls account for high rates.

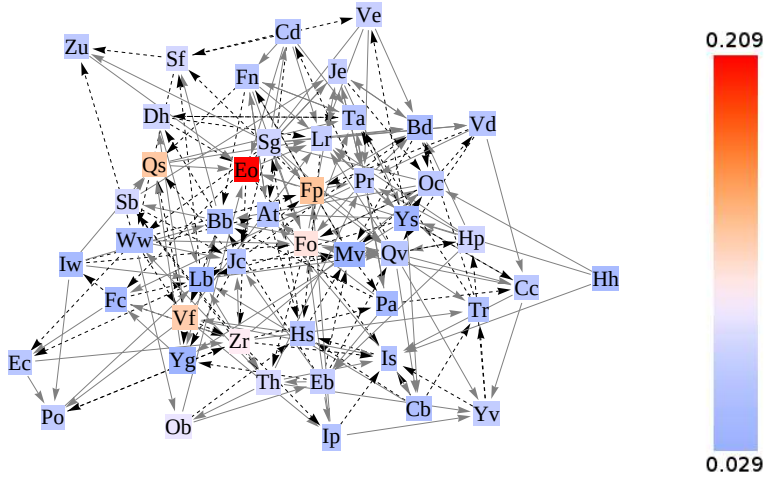


Figure 9: $\Delta\rho$ network over aggregated time series. Refer to 8 for legend.

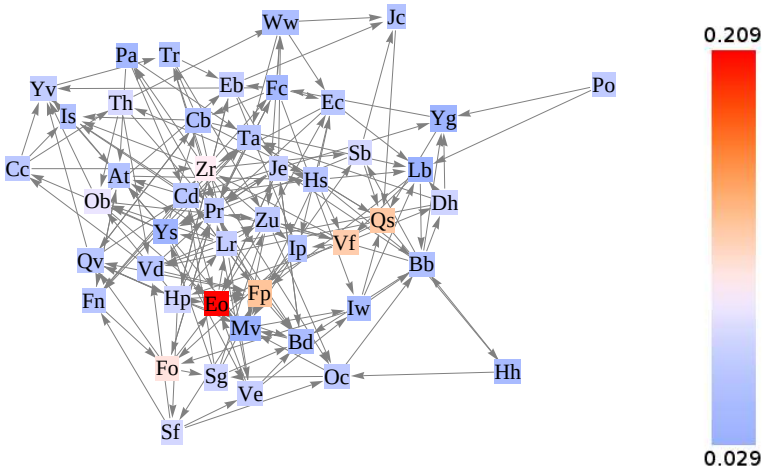


Figure 10: ΔD network over aggregated time series. Refer to 8 for legend.

quite obvious:

$$\Delta P \approx \frac{N(B|A)}{N(A)} - 0 = \frac{1}{1} = 1; \tag{17}$$

$$\Delta \rho \approx \frac{\sum_i (a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_i (a_i - \bar{a})^2 \sum_j (b_j - \bar{b})^2}} - 0 = \frac{(1-0)(1-0)}{\sqrt{(1-0)^2(1-0)^2}} = 1. \tag{18}$$

When n is large enough that the possible number of shuffles that could align the two fights becomes insignificant (remember, the number of possible pairwise combinations grows with N^2), the nulls play a small role is reducing the values of these alignments. Similar logic applies to the case of $\mathcal{R}^2$. This result makes sense because a coinciding signal when all signals are rare may be strong evidence for some relationship between the two. However, this interpretation stems from the assumption that all observations are IID, and that assumption is a strong one to make in this system. This is a problem in the sense that rare alignments will carry strong weight in rarely populated fight vectors, which may not necessarily be a sign that these should be weighted strongly. What we need is a good prior that signals whether a rare alignment should be counted strongly or whether it should not be.

For example, even if we only have a fight sequence that is 20 fights long where $N(A) = N(B) = 1$ and they align, then we have 20^2 possible nulls where only 19 of them will result in alignment, or $19/20^2 \approx 5\%$. With a fight sequence of this length, the statistic will be a shade above the p-value cutoff of 5%. For larger values, our p-value significance level should scale with the ratio of fight participation and total number of fights that day. It is an important question to ask how our significance cutoff should scale with these numbers.

Another issue is that the nulls treat the same signal across two shorter fight vectors differently from longer fight vectors. The reason for this difference is that the number of possible null shuffles does not scale the same as the measures. If we have two aligned series and shuffle them at random where m refers to number of participation and N the total length of the aligned time series, given $m_A \leq m_B$, the probability of k alignments is

$$P(k|m_A, m_B, N) = \binom{m_A}{k} \left[\prod_{i=0}^{k-1} \frac{m_B - i}{N - i}\right] \left[\prod_{j=0}^{m_A-k-1} \frac{(N - m_B) - j}{N - k - j}\right]. \quad (19)$$

The first term in brackets accounts for total number of possible alignments in the 1's. The second term in brackets accounts for the non-alignment of the remaining 1's. This alignment and non-alignment can occur $m_A!/(m_A - k)!k!$ ways since any fight can align with any other for the same result. This is not the full expression because we must account for the possibility that the 1's fall into the gaps that we do not account at t_0 and t_N when we align the two sequences. Instead, we must weight over all the possible ways that 1's call fall into these gaps. There are four possible ways this could happen that we denote as

A_0B_0: A is 0 at t_0 and B is 0 at t_N.
A_1B_0: A is 1 at t_0 and B is 0 at t_N.
A_0B_1: A is 0 at t_0 and B is 1 at t_N.
A_1B_1: A is 1 at t_0 and B is 1 at t_N.

The respective probabilities for each of these events are

$$\begin{aligned}
P(A_0)P(B_0) = P(A_0B_0) &= \left(1 - \frac{m_A}{N}\right)\left(1 - \frac{m_B}{N}\right) \\
P(A_1)P(B_0) = P(A_1B_0) &= \frac{m_A}{N}\left(1 - \frac{m_B}{N}\right) \\
P(A_0)P(B_1) = P(A_0B_1) &= \left(1 - \frac{m_A}{N}\right)\frac{m_B}{N} \\
P(A_1)P(B_1) = P(A_1B_1) &= \frac{m_A}{N}\frac{m_B}{N}
\end{aligned}$$

The equalities in each equation holds because the alignments of A and B are independent. We can explain these probabilities by looking at the probabilities that a 1 will fall into one of the endpoints.

$$P(A_1) = \frac{m_A}{N}; \quad (20)$$

$$P(A_0) = 1 - \frac{m_A}{N}. \quad (21)$$

With these probabilities, we can weight each each of the probabilities from Eq

19:

$$P'(k|m_A, m_B, N) = P(A_0B_0)P(k|m_A, m_B, N)+ \\ P(A_1B_0)P(k|m_A - 1, m_B, N) + P(A_0B_1)P(k|m_A, m_B - 1, N)+ \\ P(A_1B_1)P(k|m_A - 1, m_B - 1, N). \quad (22)$$

We must be careful to remember the condition that $k \geq m_A$. This condition applies when $m'_A = m_A - 1$. We can stick this equation into the measure equations to get the variances of the measures. If we are to be careful, we must also account for the variances of the data series as a function of their length.

This brings up the question of whether our time shuffle nulls are appropriate. If we do not normalize by the variances but linearly with respect to the total number of participation relative to the maximum amount of participation, then we will have stronger signals for the sequences with more fights. Thus, if $N(A)$ participates once, its signal will be $1/30$ the signal of B who participates 30 times more often. As we discuss above, our current nulls do not make this distinction. Rare signals that coincide can be more or equally as important according to our measures. This possible null points to the underlying assumption that we are making here about the non-participation data points.

3.6 Role of non-participation

For individuals who participate rarely—which is most individuals in this system—we find predominance of non-participation data points (0's) relative to participation (1's). The measures we use treat these two different behaviors in different ways. The ΔP measure indirectly uses the 0's: it only measures time sequential relationships between 1's, but the 0's come into play only when we account for the possible alignments that can occur in null shuffles. At a certain ratio of 0's to the number of aligned 1's, ΔP will be non-zero and zero otherwise. $\Delta\rho$ and ΔD use the the 0's directly in their arithmetic implementations, but they treat them as less informative than the 1's because they are more common. It is the more surprising events that contain more information.

On the other hand, we know non-participation is informative because the protocol for sampling behavior was identical for all individuals, meaning the data set is not suffering from sampling biases. Instead, we are facing an observational obstacle; that is, the data set does not distinguish between different kinds of behaviors in non-participation that could be important. Thus, we should choose measures and their corresponding nulls to account for non-participation as a viable strategy, which all our measures do. The difficulty is that the uniformly labeled states may actually be quite different. What really may be happening is that previous fights may be playing a role in leading to the participation of both macaques so that there is an underlying temporal sequences whose temporal dependencies are crucial for resulting in such an alignment, which we may only be able to tell from a more fine-grained picture of non-participants.

We will return to this point briefly near the end of this report, but leave the answers to future research. Since there seem to be certain difficulties related to

the binary and sparse properties of this instance of the data set, we attempt alternate representations of the data in the next section.

4 Sparse coding

Although the previous sections deal with individual strategy analysis, we know that macaques form coalitions or alliances that have coherent strategies. Also, considering the cognitive efficiencies of encoding system conflicts in terms of these stable coalitions, we might expect that there might be some system-level encoding of these coalitions in time series (citation??). In their previous paper, DKF looked at pair, C_{11}, and triadic strategies, C_{21} C_{12},. They also explored coarse-graining by shuffling identities in the recommendations offered by inhibitory and excitatory connections to ask how specific triadic strategies are to specific individuals. We apply a different sort of coarse-graining by using a principled approach to dividing the group into coalitions of varying sizes. Thus, we find differently sized coalitions, not universally individual or paired coalitions, that best reflect the internal structure of this group.

In an unpublished paper, Daniels, Flack and Krakauer have compared different methods of representing the binary time series data including the frequency, spin glass and sparse bases models. They compare them by noting their efficiency of representation as measured by the entropy required for each model. The entropy is a proxy for the cognitive resources required by a macaque to process the system given the model of interest. Under the assumption that cognitive functions are efficient and limited, a model that can accurately represent group dynamics while conserving mental resources may allow some insight into properties of the system.

In dealing with cognition, we must frame the question in terms of how a certain system can perform a reliable, robust and efficient process of statistical inference to react and encode its interactions with the environment. In the context of macaques, there are multiple levels of analysis: first, we can investigate the cognitive processes on an individual level; second, the society may encode information about conflict history. Daniels, Krakauer, Flack (DaKF), in unpublished work, explore three different kinds of compression of the time series data at the group level including the frequency, spin glass and sparse coding models. Of these three models, we use their results from the sparse coding model of the reconstructed data set to look for patterns that may be more discernable when extracting the features of the data that are important according to this encoding.

4.1 Explanation

Olshausen et al. proposed the sparse coding algorithm as a model for how the primary visual cortex (V1) encodes the visual field by relying on previous evidence based on neuron firing rates. Instead of some orthogonal decomposition such as a Fourier series, sparse coding finds a basis that is "localized, oriented,

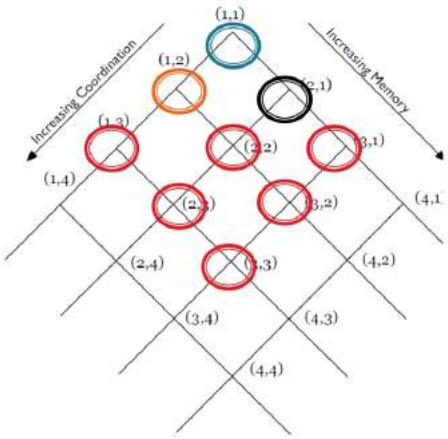


Figure 11: Graphical representation of the goals of sparse basis encoding. If we are to analyze the output of sparse coding in the light of what DKF did with the ΔP's, we can imagine that we are exploring the parameter space of strategies in this way. Instead of imposing a universal and uniform strategy for the group in different models as denoted by the blue, orange and black circles, we now attempt to straddle a natural subset of the strategy space by dividing the group up in a principled manner (red circles and other circles). Sparse coding returns groups of different sizes to us, reflecting the possibility that differently sized coalitions could demand varying cognitive resources.

and bandpass" [21] and overcomplete (in our case). This model is more realistic than previous decompositions because there is a smooth transition in the coefficients as small features in the image change whereas a Fourier series would require rapid and global changes in coefficients. Furthermore, an overcomplete code is more robust in the face of noise and degradation [21].

The motivation for investigating this model is that it seems to be quite close to real cognitive processes as well as efficient. By using the output of this sparse algorithm, we hope to reduce noise and highlight the important structure in the data. We also deal with the problem of explicitly binary data because we can perform the same analysis on the real-valued weighting coefficients of these dominant basis vectors. Finally, we obtain a principled and clear method of dividing the groups into natural subsets that explain dominant features of the data and allow us to naturally straddle the space of possible strategies (Figure 11).

One way of encoding a set of images is to find the set of basis functions, $\psi_i(\vec{x})$, and coefficients, b_i, that can represent the images, $I(\vec{x})$, such that the inner product of each image $I(\vec{x})$ with the spatial weighting function returns the coefficient.

$$b_i = \sum_{\vec{x}_j} \psi_i(\vec{x}_j) I(\vec{x}), \tag{23}$$

"where $\vec{x}_j$ denotes a discrete spatial position within the two-dimensional image"

[21]. This encoding involves solving for the invertible matrix $\psi_i(\vec{x}_j)$ such that some criterion is maximized. An alternative way for coding images is to use a generative model such that

$$I(\vec{x}) = \sum_i a_i \phi_i(\vec{x}). \tag{24}$$

These two methods can be interchangeable, but are not necessarily so. If the synthesis matrix is invertible, then we can describe it as an analysis model with independent, linear components. However, if the synthesis function set is overcomplete, there are multiple solutions for a_i that could represent the image.

Sparse basis encoding takes the second approach by minimizing two terms,

$$E(A, B, X) = \sum_{i,j} |X_{ij} - \tanh(BA)_{ij}|^2 + \lambda \sum_{k,j} S(A_{kj}|B|), \tag{25}$$

where the first term refers to the goodness of fights and the second to the sparseness of the basis set. X is the $l \times m$ matrix of m images each of dimension l, B is the $l \times N$ matrix of N basis vectors and A is the $N \times m$ matrix of reconstruction coefficients. Thus, the matrix product BA is the reconstructed time series that we put through a tanh function to prevent the overcomplete basis from overshooting the binary time series. In the second term, S is a Cauchy prior that penalizes high weight coefficients sublinearly and λ is a real-valued positive parameter that we set to adjust the sparseness of the returned basis. If λ is small, the second term will be unimportant relative to the first and we will not penalize for sparseness and will get something quite close to the original time series. If λ is large, then we will return a very sparse basis set that reconstructs the original time series badly. Essentially, the sparse coding algorithm consists of decomposing the matrix of fight series into a set of basis vectors and constrained to reconstruct the fight series with as few basis vectors as possible [5].

4.2 Results

By assuming the time series as originating from some temporally stationary generative process, DaKF used each fight as an image for the algorithm. They set the algorithm to return only 47 basis vectors with the motivation that the least sparse basis set should return each individual as one basis vector. After finding the middle point where the best fit is realized with respect to the given sparseness, they take the maximum 25 elements, by magnitude, in the basis. We call this coarse-grained basis our dominant basis vectors.

Depending on the initial conditions, this algorithm will produce multiple solutions via gradient descent. One of the bases and the one that we use for our analysis, consists of the 13 vectors,
We have not yet done an extensive comparison of the results from the different bases that we obtain from different initial conditions. From a cursory comparison with a few other randomly chosen bases, we do not find much in the number

Hp	Eo	Fo	Hh	Je	Eo	Fn	Dh	Eb	Eo	Sf	Cc	Fp
Ve	Qs			Vd	Vf	Th	Po	Ob	Fp	Zr	Ec	Sg
												Zu

Table 1: Percentage of significant measures in divided data set. "Dom. coeffs" refers to the coefficients of the 13 sparse basis vectors. "All coeffs." refers to the weighting coefficients for all 47 basis vectors before we coarse-grained to trim the number of individuals down to 25. "Reconstr." refers to the representation of the binary data found by the sparse coding algorithm. Remember that ΔP's are only defined for binary data.

Data Set	Sig. ΔP's	Sig. ΔDcor's	Sig. $\Delta\rho$	Total	Iter.'s
Time series	1.65%	3.46%	5.37%	61852	3000
Dom. coeffs.	N/A	5.94%	10.33%	4732	10000
All coeffs.	N/A	6.28%	11.73%	61852	10000
Reconstr.	N/A	6.07%	10.99%	13552	10000

of significant measures. We have not yet compared measures from basis to basis or the distributions. However, since most of the bases are quite similar and they represent the same data, we do not expect much change.

As we see in Figure 12, we do not see a decrease in the variance for which we hoped.

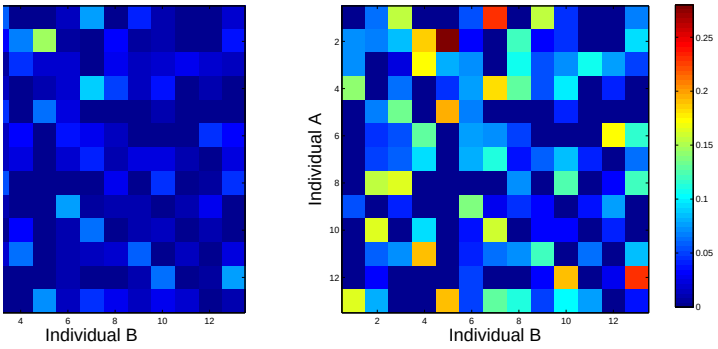
5 Temporal variance of ΔP

To conclusively show that the variance in the measures across days is anomalous, we must first check whether the length of the days has an effect on their variance. If we can show that variance is inversely related to the length of the observational periods, we may be able to conclude that above a certain level of aggregation of the data sets, we should expect stability and that observed changes indicate real changes. On the other hand, if we cannot show that any of our fluctuations can be distinguished from apparent changes that do not correspond to real changes, we may not be able to make any conclusive statement.

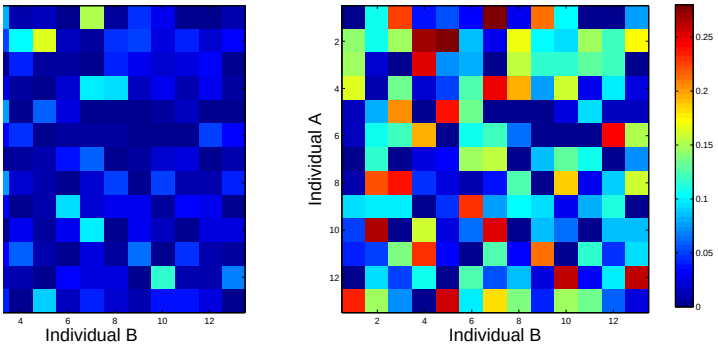
We will simulate data whose true underlying distribution we do know and see how the distributions of the values of the measures and the number of significant measures change with taking larger and larger aggregates of the data set.

6 Future avenues

Through these efforts, we have to believe that the current format of the data may not be suitable for analysis with measures that are explicitly designed for real-valued data. If we look at the distributions on which we are attempting to use these measures on, we realize that the four possible states represent the



(a)



(b)

Figure 12: **(a)** $\Delta\rho$ array between all 13 sparse bases with absolute values of means over all observational periods with more than 45 fights on the left. There are 14 such periods out of the 28. On the right, standard deviation over periods. Standard deviation is large relative to the mean. **(b)** Similar graphs except for ΔD.

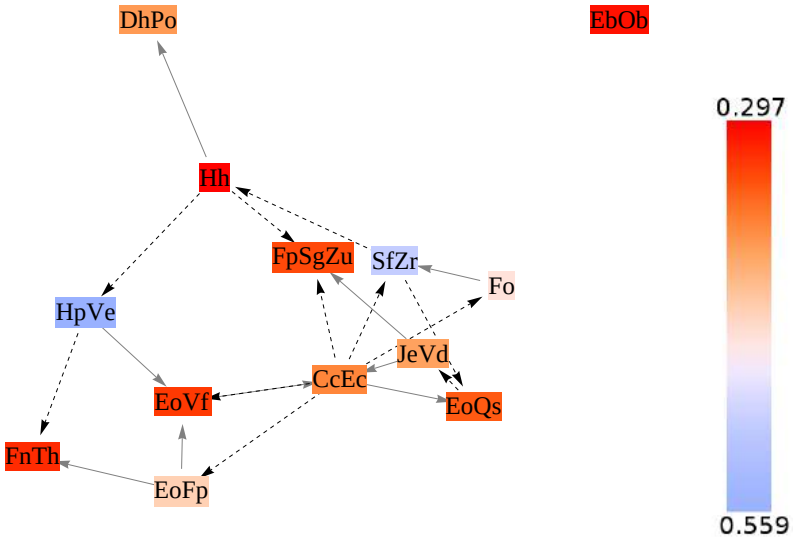


Figure 13: $\Delta\rho$ network for sparse bases over aggregated time series. There are 10 inhibitory connections and 9 excitatory connections. Refer to 8 for legend.

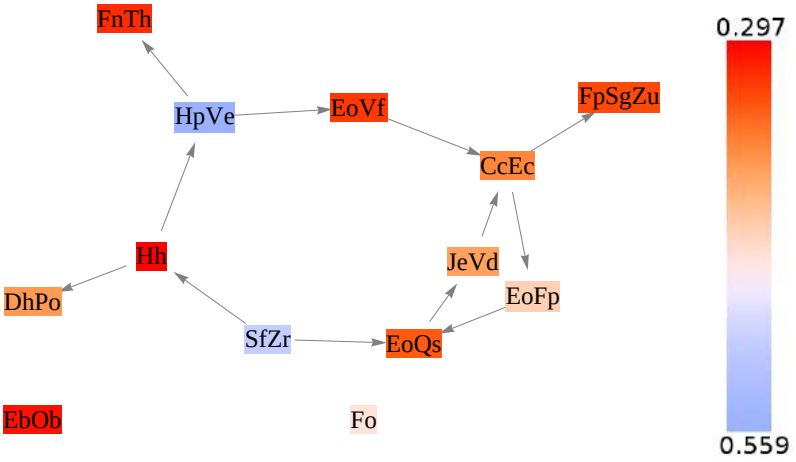


Figure 14: ΔD network for sparse bases over aggregated time series. Refer to 8 for legend.

corners of a square at $(0,0),(0,1),(1,0)$ and $(1,1)$. With $\Delta\rho$, we are trying to determine whether a line with a non-zero slope will plausibly pass through these points. Perhaps one future direction would be to invert the data set such that the 1's are replaced with 0's to see whether we can investigate ΔP values in non-participation. This inversion would not change the results of either $\Delta\rho$ or ΔD because they are symmetric with respect to the labeling of the states.

As we have tried in this paper, the fruitful approach with respect to other measures would be to reformat the data set in a way such that the patterns fall along some continuous dimension. This may require encoding of some kind or returning to the highly-resolved, original data set that could provide a more detailed description about the levels of participation in and roles in conflicts.

Another step we could take is to exploit the ability of distance covariance to take inputs of arbitrary dimension to compare sets of histories with current fights or future histories. This comparison would naturally introduce more dimensions without any adjustments made to the data set. This avenue leads naturally to the question of whether the sequential time step in bout time is the appropriate resolution for finding temporal patterns.

So far, we have concluded that several steps have been fruitful, and that we should continue with these approaches:

1. Reduce the number of possible states by aggregating the individuals such that our sample size becomes larger with respect to the number of possible states. This may involve using subsets of individuals known to compose an internal community. DKF showed evidence for community-based temporal patterns in conflict participation [7].

2. Encode the data in a format that reduces the dimensionality of the space of actors, increases the resolution of participation, and increases existing structure by reducing noise.

7 Appendix

7.1 Motivation for sequential step analysis

Although it may seem rather arbitrary to look at sequential time steps in bout time (maybe we could look at every other or consider the participation of a certain individual or look at real time etc.), we believe that we will find some significant correlation across sequential time steps. The strongest evidence we have is from the findings in the previous IGT paper [8]. It also makes sense that monkeys would remember the previous bout freshest in their mind for making decisions about the sequential fight (citation??). As for experimental evidence, we plotted the number of significant measures for different individual as a function of the time lag between fights in bout time. We find the most number of significant measures at one time step. This is not meant to be a rigorous argument; after all, this pattern might be a result of the amount of data that we have for single steps.

There could be several kinds of time dependencies in the data. One method is to look for sequential dependencies between time steps in the data. Reminiscent of autocorrelation, we can look for dependencies between in time lags; that is, adjacent fights, every other fight, etc. One measure we used to look for time lag is mutual information. From previous work, we know that this is not the complete picture because there seem to be multiple timescales at work in the system, some of which depend on real time [7].

7.2 Redefinition of ΔP

We modified the normalization of the null in Eq 26 since shuffling the data sets could mean $N(A) \neq N_{\text{null}}(A)$. We refer to this hereon as ΔP with the *new null*. For example, if we have the fight series

$$\begin{array}{c} \\ t_0 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ NA & NA \\ 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array},$$

we see that by aligning t_n for A with t_{n+1} for B, B only participates in 1 fight in the aligned time series. We see this because we must eliminate the first fight of B–it does not contribute to anything–and the first fight of B after the break for the same reasons. Correspondingly, we throw out the last fight of A before the break and at the very end. However, if we now shuffle the fights such that

we get

$$\begin{array}{c} \\ t_0 \\ t_1 \\ t_2 \\ t_3 \\ t_4 \end{array} \begin{array}{c} \begin{array}{cc} A_{\text{null}} & B_{\text{null}} \end{array} \\ \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ NA & NA \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array},$$

we find that B appears twice instead of once.

7.3 Previous work

In DKF's paper (2010), DKF took an inverse approach to determining the rules of conflict in a pigtailed macaque society, called Inductive Game Theory. Unlike traditional game theoretical approaches where the payoffs and interactions are postulated from which the behavior in the game are then explored, DKF constructed a Monte Carlo Markov chain model that used for its parameters the probability that macaques in a previous fight would predict the participation of another. The parameters for the $C_n m$ model was the correlation measure between adjacent fights

$$\Delta P(A \to B) = \frac{N(B|A) - N_{\text{null}}(B|A)}{N(A)} \tag{26}$$

"where $N_{\text{null}}(B|A)$ is the average from a large Monte Carlo set of null models generated by time-shuffling the series but not shuffling identities within fights" [8]. Although Eq 26 refers to the C_{11} model–where A and B represent individuals–,we can expand this model to account for larger subsets of the group as predictors, A, and as predicted, B. They, however, restricted the possible parameter space by known cognitive and social limitations of macaque behavior to C_{11}, C_{12} and C_{21}. For each model, they considered these ΔP relations to be either recommendations or inhibitions for joining a particular fight from all possible subsets of individuals or coalitions. Between all these different recommendations, they posited an AND or OR rule. In comparison with the true fight distribution, they discovered that neither the C_{11} AND or OR nor C_{12} AND or OR models returned the expected distribution but that the C_{21} OR model, the Triadic Hypothesis, was reasonably good at predicting the observed distribution.

In a following paper, DKF. looked for periodicities in the real time, not bout time [7]. Borrowing a method from astronomy to look for periodicities in discrete, irregularly sampled data–a more complicated form of Fourier analysis– Dedeo et al. found that certain subsets of individuals based on sex, social hierarchy and matrilineal relations showed characteristic frequencies in their participation in fights. Many of these subsets agreed with some functional relationship between the individuals such as kin, sex or dominance.

When we recalculated the measures between these different pairs for the $\Delta P(A, B)$ pairwise relationships, we found that DKF had calculated the ΔP

values without an inequality in determining the p-values. DKF found "513 detections of non-zero ΔP in the data" out of 2304, or 22.3%, at the 95% significance above and below cutoff [6]. When taking some time shuffled form of the data set and looking for artificial ΔP's, they found 480 detections, a comparable number signifying that there is some dependence between the $N(B|A)$. At the risk of being overly pedantic, we calculate the p-values by comparing the null shuffles with the observed $P(A, B)$ such that the p-value corresponds to the percentage of null shuffles that are greater than or as extreme as the observed statistic on either 5% end of the distribution. DKF's p-value's did not account for the null shuffles that were equal to the observed statistic. In many data sets, this issue would not exist, but the binary and sparse characteristics leads to many null shuffles that agree exactly with the $P(A, B)$. Given this quandary, we find a much smaller number of non-zero ΔP's especially in the lower end of the distribution.

For the lower ΔP's, we often have just a few alignments or no alignments such that accounting for the equality drastically reduces the number of ΔP's. For example, imagine the scenario where we have two individuals A and B that are mostly non-participants such that they do not align in any of the fights and have $P(A, B) = 0$. When we perform the null shuffle, we will again obtain most of the $P_{\text{null}}(A, B) = 0$. A few of these nulls will contain alignments and be greater than 0, but let us assume that they constitute less than 5% of the population of nulls. If we do not account for the inequality in calculating the ΔP's, we find that this statistic is significant because almost all the nulls are *just as extreme* as $P(A, B)$. Now, when we subtract the null, we will obtain a negative ΔP. This observation explains why only a few of the ΔP's are negative and why their values are smaller than the positive ΔP's. In the simulation, this observation will not change the results because all the inhibitory connections are filtered out by the OR logic, but now we have many fewer excitatory connections than before. We have not yet taken a look at how this might change the resulting conflict size distribution.

We also experimented with different nulls to get an idea of the shifts the number of ΔP's according to the two nulls. We compared the permutation test with the bootstrap test. By permutation, we refer to null shuffles where the fights are permuted so that the resulting distribution is exactly the same. By bootstrap, we refer to shuffles where we perform a bootstrap sampling of the original distribution such that the resulting distribution is not exactly the same. The difference between these two nulls is that permutation is a stronger null test; that is, there are fewer other possibilities that we are testing for the null model, specifically $N!$ possible permutations. The bootstrap shuffle is a weaker test since we have N^N possible sampled sets.

We also compared each of these shuffles with 1000 fake sets of data. A fake data set refers to a permuted shuffle of the original time series that we take as an original time series. On this permuted shuffle, we perform the same analysis with ΔP's as we do with the original time series. Then, we count the number of significant ΔP's to compare with the number of ΔP's from the original time series. To be clear, let us call the original order of fights α_0. If we are looking

at "Original null w/out replacement" in Table 2, we take α_0 and shuffle it 1000 times to get $\alpha_1, \alpha_2, ..., \alpha_{1000}$. From these permutations, we get 101 and 33 ΔP's at the high and low ends of the distribution. Then, we shuffle α_0 to get α_0'. We now perform the same analysis such that we compare it with 1000 null shuffles $\alpha_1', \alpha_2', ..., \alpha_1 000'$. We perform this entire process with $\alpha_0'', \alpha_0^{(3)} ..., \alpha_0^{(1000)}$. Then, we can get a p-value on the number of ΔP's we measured from each of these α_0^n's. We are effectively looking for the p-values on the p-values. We find a much smaller number of ΔP's and no negative ΔP's. This is an interesting observation because it tells us that any low number of ΔP's at the lower end is not anomalous in α_0, but a high number of ΔP's at the upper is. It is also surprising that the bootstrap shuffles have lower variance than the permutations. We would expect that the permutation tests result in a higher variance because bootstrapping increases the variance randomly sampling a slightly different distribution.

Table 2: Comparison of variance in ΔP's for different kinds of nulls. For the section labeled "w/out replacement", we refer to the permutation test. For the "w/ replacement sections, we refer to the bootstrap. We do the null shuffle 1000 times for each of these. The mean and standard deviations (denoted "Std.") refer to comparisons with 1000 fake data sets and the number of ΔP's in each of them.

Side of distrib.	Non-zero ΔP's out of 47^2	As %	Mean	Std.	p-value
Original null w/out replacement					
High at 95%	101	4.57%	73.0	10.7	0.009
Low at 95%	33	1.49%	50.1	8.27	0.983
Original null w/ replacement					
High at 95%	80	3.62%	55.2	9.0	0.003
Low at 95%	21	0.95%	37.0	6.8	0.992
New null w/out replacement					
High at 95%	136	6.16%	105	13.4	0.012
Low at 95%	47	2.13%	86.6	12.1	0.998
New null w/ replacement					
High at 95%	121	5.48%	94.8	12.1	0.021
Low at 95%	45	2.04%	73.9	10.5	0.997

In Table 2, we find two interesting results. First, the variance on the bootstrap is comparable to the variance on the permutation test, but larger in both cases. We would expect that the variance on the null with bootstrapping to be larger than the permutation null because the distribution changes with the former but remains exactly the same with the latter. With the original null, the standard deviation of the null permutations are 10.7 and 8.27 for ΔP's at the high and low ends of the distribution. On the other hand, the bootstrap yields 9.0 and 6.8. A possible explanation is that aligning the two sequences after sampling from the original distribution introduces or removes variance in a way that we do not foresee.

Second, the number of negative ΔP is anomalously low. Out of the null shuffles or bootstraps, we find that having as few inhibitory connections as we observe in the original data is unusual: most nulls have a greater number of negative ΔP's such that our p-values on the number of significant negative ΔP's detected are all > 0.9. These data suggests that the ΔP analysis of this data set on sequential time steps does not detect inhibitory relationships, but that does not mean inhibitory connections do not exist.

There are two possible explanations since we observe and expect inhibitory relationships in this system. It may be that inhibitory relationships act over longer time scales that our sequential time analysis does not capture. However, that is contrary to what we observe in Figure 13, where there seem to be a plethora of inhibitory connections even though that originates from sequential bout analysis. This observation leads to the second hypothesis that inhibitory connections do not appear at an individual scale: inhibitory relationships apply to higher order relationships. One avenue for exploring the second hypothesis would be to look at different pairs of individuals to expand on the triadic interaction hypothesis from DKF's paper on Inductive Game Theory. Another avenue would be to shuffle sparse bases data and perform the same analysis as in Table 2.

7.4 Székely, Rizzo and Bakirov: Theorem 1

Now, we present a cursory explanation to motivate Theorem 1 in Székely and Rizzo's original paper [33]. These integrals apply, obviously, only when X is continuous. However, we deal with finite realizations of a discrete X; thus, we use the empirical characteristic function

$$f_X^n = \frac{1}{n}\sum_{k=1}^{n} e^{i\langle t, X_k\rangle}. \tag{27}$$

For the joint distribution of X and Y, we define the two dimensional function

$$f_{X,Y}^n(t,s) = \frac{1}{n}\sum_{k=1}^{n} e^{i\langle t, X_k\rangle + i\langle s, Y_k\rangle}. \tag{28}$$

We have a set of inputs X and corresponding outputs Y, denoting each pair as $Z = (X, Y)$. Each X_i and Y_i represent random vectors of dimension p and q, respectively.[9] We can calculate DCOV in the following way. We begin with the characteristic function:

Now, let us take the simple case where X and Y are of dimensions $p = q = 1$ for demonstrative purposes. Remember that the integral that we wish to evaluate is

$$\int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t) f_Y(s)|^2 d\omega, \tag{29}$$

[9]Random in the sense that their deviation from their underlying distributions is random. Here, we use X to refer to the distribution and x or X_i as a realization of the variable.

where we follow the same notation as in Székeley 2007.

Then, we must evaluate

$$|f^n_{X,Y}(t,s) - f^n_X(t)f^n_Y(s)|^2 = (f^n_{X,Y}(t,s) - f^n_X(t)f^n_Y(s))(\overline{f^n_{X,Y}(t,s)} - \overline{f^n_X(t)f^n_Y(s)})$$
$$= |f_{X,Y}(t,s)|^2 + |f_X(t)f_Y(s)|^2 - 2\overline{f_{X,Y}(t,s)}f_X(t)f_Y(s). \tag{30}$$

Although 30 may look pretty complicated in sum form, it turns out that many of the terms become irrelevant when we integrate. First, let us simplify this into a usable form. I will only simplify the joint distribution as an example. Let

$$u_k := \langle t, X_k \rangle ,$$
$$v_k := \langle s, Y_k \rangle .$$

Then, we can write using Euler's identity $e^{\pm ix} = \cos x \pm i \sin x$,

$$f_{X,Y}\overline{f_{X,Y}} = \frac{1}{n^2}\left(\sum_{k=1}^{n}(\cos u_k + i\sin u_k)(\cos v_k + i \sin v_k)\right)\cdot ...$$
$$\left(\sum_{l=1}^{n}(\cos u_l - i\sin u_l)(\cos v_l - i\sin v_l)\right)$$
$$= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}(\cos u_k \cos v_k - ...)(\cos u_l \cos v_l - ...)$$
$$= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n}\cos u_k \cos u_l \cos v_k \cos v_l - ...$$

where we omit the rest of the terms because they will not matter. We now use the trigonometric identity

$$\cos u \cos v = \cos(u \pm v) \pm \sin u \sin v$$

to find

$$f_{X,Y}\overline{f_{X,Y}} = \frac{1}{n^2}\sum_{k,l=1}^{n}\cos(u_k - u_l)\cos(v_k - v_l) - ...$$
$$= \frac{1}{n^2}\sum_{k,l=1}^{n}\cos[(X_k - X_l)t]\cos[(Y_k - Y_l)s] + V_1,$$

where V_1 represents the terms that will vanish in the integral. We use the distributive property of scalar products to separate out the terms, and the scalar product is implicit in the parenthetical notation. This equation is the same as on page 2775 in [33].

With these same tricks, we can find the other equations on the same page that I reiterate here:

$$f_X^n(t)f_X^n(s)\overline{f_X^n(t)f_X^n(s)} = ...$$
$$\frac{1}{n^2}\sum_{k,l=1}^{n}\cos[(X_k - X_l)t]\frac{1}{n^2}\sum_{k,l=1}^{n}\cos[(Y_k - Y_l)s] + V_2 \quad (31)$$

$$f_{X,Y}^n(t,s)\overline{f_X^n(t)f_Y^n(s)} = ...$$
$$\frac{1}{n^3}\sum_{k,l,m=1}^{n}\cos[(X_k - X_l)t]\cos[(Y_k - Y_m)s] + V_3. \quad (32)$$

We integrate each of these separately with the weight function from Lemma 1 $\omega(t,s) = \{c_p c_q |t|_p^{1+p}|s|_q^{1+q}\}^{-1}$ and not before pulling the sum out in front of the integral. Here, we reiterate Lemma 1 [31]

LEMMA 1:
If $0 < \alpha < 2$, *then for all* x *in* $\mathbb{R}^d$

$$\int_{\mathbb{R}^d}\frac{1-\cos\langle t,x\rangle}{|t|_d^{d+\alpha}}\,dt = C(d,\alpha)|x|^\alpha, \quad (33)$$

where

$$C(d,\alpha) = \frac{2\pi^{d/2}\Gamma(1-\alpha/2)}{\alpha 2^\alpha\Gamma((d+\alpha)/2)} \quad (34)$$

and $\Gamma(\cdot)$ *is the complete gamma function. The integrals at 0 and* ∞ *are meant in the principal value sense:* $\lim_{\epsilon\to 0}\int_{\mathbb{R}^d\{\epsilon B+\epsilon^{-1}B^c\}}$, *where* B *is the unit ball (centered at 0) in* $\mathbb{R}^d$ *and* B^c *is the complement of* B.

For the proof, see the above citation. Lemma 1 then suggests a weight function of the form for normalization purposes

$$\omega(t,s;\alpha) = (C(p,\alpha)C(q,\alpha)|t|_p^{p+\alpha}|s|_q^{q+\alpha})^{-1}, \quad 0 < \alpha < 2 \quad (35)$$

and we end up with an equation of the form

$$||f_{X,Y}^n(t,s) - f_X^n(t)f_Y^n(s)||^2 = S_1 + S_2 - 2S_3. \quad (36)$$

where S_1, S_2 and S_3 are given by Eqs 2.15, 2.16 and 2.17 in the paper. Thus, we see how by careful accounting, we can arrive at a simpler form of $\mathcal{V}^2$ that involves no tricky integrals and complex variables. In fact, the computational implementation of this algorithm is trivial in the sense that it only involves the calculation of Euclidean distances and the summation of matrices. As for the last part of the proof verifying the algebraic identity

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = S_1 + S_2 - 2S_3, \quad (37)$$

this step requires messy accounting with many terms, and we will not reproduce that here.

References

[1] S Achard. "Asymptotic properties of a dimension-robust quadratic dependence measure." In: *Comptes Rendus Mathematique* 346.3-4 (Feb. 2008), pp. 213–216. URL: http://linkinghub.elsevier.com/retrieve/pii/S1631073X07004438.

[2] a J Bell and T J Sejnowski. "An information-maximization approach to blind separation and blind deconvolution." In: *Neural computation* 7.6 (Nov. 1995), pp. 1129–59. URL: http://www.ncbi.nlm.nih.gov/pubmed/7584893.

[3] Austin Burt and Robert Trivers. *Genes in Conflict: The Biology of Selfish Genetic Elements*. Harvard University Press, 2008.

[4] Leslie Cope. "Discussion of: brownian distance covariance." In: *The Annals of Applied Statistics* 3.4 (2009), pp. 1279–1281.

[5] B. Daniels, D. Krakauer, and J. Flack. 2011.

[6] Simon Dedeo, David C Krakauer, and Jessica Flack. "Supporting Information for Inductive Game Theory and the Dynamics of Animal Conflict Aggregate Properties of the Time Series." In: *Conflict* (2010), pp. 1–13.

[7] Simon DeDeo, David C. Krakauer, and Jessica C. Flack. "Evidence of strategic periodicities in collective conflict dynamics." In: *Arxiv preprint* (2011), pp. 1–22. arXiv:arXiv:1101.1556v1. URL: http://arxiv.org/abs/1101.1556.

[8] Simon DeDeo, David C Krakauer, and Jessica C Flack. "Inductive Game Theory and the Dynamics of Animal Conflict." In: *PLoS computational biology* 6.5 (2010).

[9] J. C. Flack, David C. Krakauer, and Frans B. M. de Waal. "Robustness mechanisms in primate societies: a perturbation study." In: *Proc. R. Soc. B* (272 2005), pp. 1091–1099.

[10] Jessica C. Flack et al. "Policing stabilizes construction of social niches in primates." In: *Nature* 439 (), pp. 426–429.

[11] G F Harpur and R W Prager. "Development of low entropy coding in a recurrent network." In: *Network (Bristol, England)* 7.2 (May 1996), pp. 277–84. URL: http://www.ncbi.nlm.nih.gov/pubmed/16754387.

[12] E. T. Jaynes and G. L. Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[13] Michael R. Kosorok. "Discussion of: brownian distance covariance." In: *The Annals of Applied Statistics* 3.4 (2009), pp. 1270–1278.

[14] Michael R Kosorok. "On Brownian Distance Covariance and High Dimensional Data." In: *The annals of applied statistics* 3.4 (Jan. 2009), pp. 1266–1269. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2889501\&tool=pmcentrez\&rendertype=abstract.

[15] Abba M. Krieger and Paul E. Green. "Generalized measures of association for ranked data with an application to prediction accuracy." In: *Journal of Classification* 10.1 (Jan. 1993), pp. 93–114. URL: http://www.springerlink.com/index/10.1007/BF02638455.

[16] W.H. Kruskal. "Ordinal measures of association." In: *Journal of the American Statistical Association* 53.284 (1958), pp. 814–861. URL: http://www.jstor.org/stable/2281954.

[17] David J. C. Mackay. *Information Theory*. Cambridge University Press, 2003.

[18] A Micheas and K Zografos. "Measuring stochastic dependence using -divergence." In: *Journal of Multivariate Analysis* 97.3 (Mar. 2006), pp. 765–784. URL: http://linkinghub.elsevier.com/retrieve/pii/S0047259X05000515.

[19] Michael a. Newton. "Introducing the discussion paper by Székely and Rizzo." In: *The Annals of Applied Statistics* 3.4 (Dec. 2009), pp. 1233–1235. URL: http://projecteuclid.org/euclid.aoas/1267453932.

[20] B a Olshausen and D J Field. "Natural image statistics and efficient coding." In: *Workshop on Information Theory and the Brain*. Vol. 7. 2. University of Stirling, Scotland, May 1995, pp. 333–9. URL: http://www.ncbi.nlm.nih.gov/pubmed/16754394.

[21] Bruno A Olshausen and David J Fieldt. "Strategy Employed by V1 ?" In: *Science* 37.23 (1997), pp. 3311–3325.

[22] Liam Paninski. "Estimation of Entropy and Mutual Information." In: *Neural Computation* 15.6 (June 2003), pp. 1191–1253. URL: http://www.mitpressjournals.org/doi/abs/10.1162/089976603321780272.

[23] Karl Pearson. "On the probable error of a coefficient of mean square contingency." In: *Biometrika* 10.4 (1915), pp. 570–573. URL: http://www.jstor.org/stable/2331842.

[24] A. Rényi. "On Measures of dependence." In: *Acta Mathematica* 10 ().

[25] Joseph Lee Rodgers and W. Alan Nicewander. "Thirteen Ways to Look at the Correlation Coefficient." In: *The American Statistician* 42.1 (Feb. 1988), p. 59. URL: http://www.jstor.org/stable/2685263?origin=crossref.

[26] Sohan Seth and José C. Príncipe. "Generalized measure of association." 2010.

[27] Sohan Seth and Jose C. Principe. "Variable Selection: A Statistical Dependence Perspective." In: *2010 Ninth International Conference on Machine Learning and Applications* (Dec. 2010), pp. 931–936. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5708971.

[28] David Sherrington. In: *Arxiv preprint* (2009), pp. 1–20.

[29] C. Spearman. "The proof and measurement of association between two things." In: *The American journal of psychology* 15.1 (1904), pp. 72–101. URL: http://www.jstor.org/stable/1412159.

[30] JF Steffensen. "On certain measures of dependence between statistical variables." In: *Biometrika* 26.1/2 (1934), pp. 251–255. URL: http://www.jstor.org/stable/2332058.

[31] G. Székely and M. Rizzo. "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method." In: *Journal of classification* 22 (2 2005), pp. 151–183.

[32] Gabor J Szekely and Maria L. Rizzo. "Brownian distance covariance." In: *The Annals of Applied Statistics* 3.4 (2009), pp. 1236–1265.

[33] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. "Measuring and testing dependence by correlation of distances." In: *The Annals of Statistics* 35.6 (Dec. 2007), pp. 2769–2794. URL: http://projecteuclid.org/euclid.aos/1201012979.