# ORTHOGONAL DECOMPOSITIONS OF MULTIVARIATE STATISTICAL DEPENDENCE MEASURES

*Ilan N. Goodman and Don H. Johnson*

Rice University, ECE Department, Houston, TX 77251-1892
igoodman@rice.edu, dhj@rice.edu

## ABSTRACT

We describe two multivariate statistical dependence measures which can be orthogonally decomposed to separate the effects of pairwise, triplewise, and higher order interactions between the random variables. These decompositions provide a convenient method of analyzing statistical dependencies between large groups of random variables, within which smaller "sub-groups" may exhibit dependencies separately from the rest of the variables. The first dependence measure is a generalization of Pearson's $\phi^2$, and we decompose it using an orthonormal series expansion of joint probability density functions. The second measure is based on the Kullback-Leibler distance, and we decompose it using information geometry. Applications of these techniques include analysis of neural population recordings and multimodal sensor fusion. We discuss in detail the simple example of three jointly defined binary random variables.

## 1. INTRODUCTION

Quantifying the statistical dependencies among jointly distributed random variables has never been simple. The most commonly used dependence measure, the correlation coefficient, only measures linear dependence between random variables, and applies only to pairs of variables. However, many applications involve large groups of variables that have very complicated relationships not captured by correlation. For example, neuroscientists are currently able to make recordings of tens to hundreds of neurons simultaneously, and it is known that these ensembles exhibit time-varying dependencies that may contribute to stimulus encoding [1]. Consequently, data-efficient dependence assessment techniques that can be applied to such high-dimensional random vectors are vital to understanding neural population codes. Another application involving complicated dependencies is multi-modal sensor fusion. In [2, 3] the authors describe methods to fuse audio and video data in order to locate the source of a sound. In fusing data from a large number of sources, traditional approaches usually resort to simplifying assumptions in order to facilitate an analytic solution; consequently they are applicable only to a narrow class of problems that fall within the modelling assumptions. Model-free statistical techniques are needed that are applicable to broader classes of problems.

Even simple examples show that, in general, groups of random variables can express more than just pairwise dependence. For example, consider $N$ binary random variables. Fully specifying their joint distribution requires $2^N - 1$ parameters; there are $N(N-1)/2$ pairwise correlations, which, together with the $N$ marginal probabilities, can only specify the joint distribution when $N = 2$. For larger groups, third and higher order dependencies also need to be determined. The challenge is to quantify these dependencies in a coherent, meaningful way.

We discuss here two dependence measures with stronger properties than correlation. We show how these measures can be decomposed such that each component quantifies the contribution of dependencies of each order separately, providing further detail about the intricacies of the interactions between random variables. In the neural population analysis example, decomposing the dependence measures could provide a level of detail about the interactions between neurons that was unavailable using conventional techniques. In the audio-visual example, the dependence measure decompositions could be used to identify complex relationships between the audio and video sources that would not be found using other techniques.

## 2. PHI-SQUARED DEPENDENCE MEASURE

*Pearson's* $\phi^2$ is a measure of the distance between a bivariate distribution and its independent counterpart, the product of the marginals[4]. To generalize this measure for multiple variables, let $\mathbf{X} = (X_1, \ldots, X_N)$ be a random vector with $X_n \in \mathcal{X}$, with joint probability distribution $p_{\mathbf{X}}(\mathbf{x})$ and marginal distributions $\{p_{X_n}(x_n)\}$, $n = 1, \ldots, N$. We define

$$\phi^2 = \int_{\mathbf{x} \in \mathcal{X}^N} \frac{p_{\mathbf{X}}^2(\mathbf{x})}{\prod_{n=1}^{N} p_{X_n}(x_n)} d\mathbf{x} - 1 \qquad (1)$$

$\phi^2$ equals zero if and only if the random variables are statistically independent, and increases without bound (in gen-

eral) with increasing dependence. For example, for a bivariate Gaussian distribution with correlation coefficient $\rho$, $\phi^2 = \rho^2/(1-\rho^2)$.

### 2.1. Expansion of $\phi^2$

In [5, 6] it was shown that for bivariate distributions, when $\phi^2 < \infty$, $\phi^2$ can be expanded in terms of the coefficients of an orthonormal expansion of the bivariate distribution. In general, when $\phi^2$ is bounded,

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_{n=1}^{N} p_{X_n}(x_n) \cdot \left[ 1 + \sum_{i_1,\cdots,i_N} a_{i_1 \cdots i_N} \prod_{n=1}^{N} \psi_{i_n}(x_n) \right] \tag{2}$$

The functions $\{\psi_{i_n}(x_n), i_n = 0, \ldots\}$ form an orthonormal basis with respect to a weighting function equal to the marginal probability function $p_{X_n}(x_n)$. These basis functions are chosen so that $\psi_0(x_n) = 1$. Then,

$$\phi^2 = \sum_{i_1,\cdots i_N} a^2_{i_1 \cdots i_N} \tag{3}$$

### 2.2. Dependence ordering in $\phi^2$

In the above expansion of a multivariate distribution function, we say that a the function $f(x_1, \ldots, x_N) = \prod_{n=1}^{N} \psi_{i_n}(x_n)$ is *of order* $k$, $2 \le k \le N$, if the variables $X_n$ can be re-indexed such that $f(x_1, \ldots, x_N) = g(x_1, \ldots, x_k)$. In other words, there are exactly $k$ non-constant factors in $f$[7]. If we let $\mathcal{I} = \{i_1, \ldots, i_N\}$ be the set of all indices in the expansion of $\phi^2$, $\mathcal{I}$ can be partitioned into logical components $\beta = \{\beta_k\}$, $k = 2, \ldots, N$, where each $\beta_k$ is the set of indices $(i_1, \ldots, i_N) \in \mathcal{I}$ for which the function $f$ in the expansion is of order $k$. Then,

$$\phi^2 = \sum_{k=2}^{N} \left( \sum_{(i_1,\ldots,i_N)\in\beta_k} a^2_{i_1 \cdots i_N} \right) \equiv \sum_{k=2}^{N} \phi^2_k \tag{4}$$

Thus $\phi^2$ is decomposed into $N-1$ components, where each component is the contribution of interactions of a different order within the ensemble of random variables.

### 2.3. Example: three binary random variables

Consider the simple example of three binary random variables $X_1$, $X_2$ and $X_3$, with $p_n = P[X_n = 1]$ and $\sigma_n = \sqrt{p_n(1-p_n)}$, $n = 1, 2, 3$. Using the orthogonal polynomials as a basis set for the marginal probability distributions, we have for each $X_n$

$$\psi_0(x_n) = 1 \qquad \psi_1(x_n) = (x_n - p_n)/\sigma_n$$

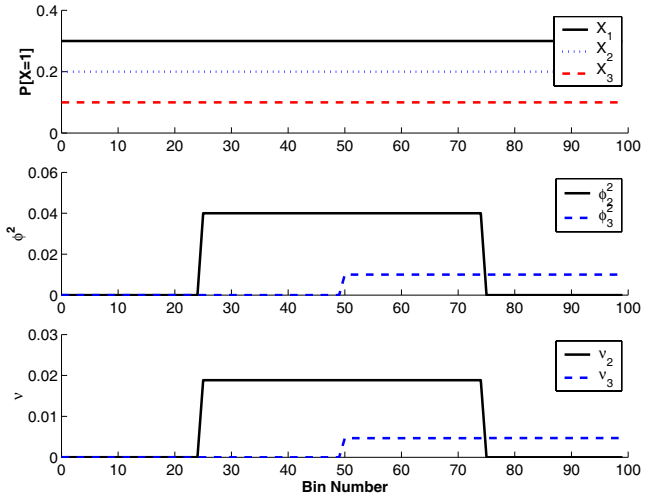


**Fig. 1**. Decomposition of dependence measures for three jointly distributed binary random variables. The joint distribution $p(x_1, x_2, x_3)$ was simulated in each bin. The top plot shows the marginal distributions $P[X_n = 1]$ for each variable; the marginal distributions are constant throughout the simulation. The middle plot shows the components of the $\phi^2$ dependence measure, computed as described in section 2.3. The bottom plot shows the components of the dependence measure $\nu$, computed as described in section 3.3. Between bin 25 and bin 75, there is a constant level of $2^{nd}$ order dependence. Between bin 50 and bin 100 there is a constant level of $3^{rd}$ order dependence.

Now the coefficients $a_{i_1 i_2 i_3}$ in the expansion of the joint distribution function are easy to compute:

$$\begin{aligned} a_{12} &= \rho_{12} \\ a_{13} &= \rho_{13} \\ a_{23} &= \rho_{23} \\ a_{123} &= \rho_{123} - \frac{p_1}{\sigma_1}\rho_{23} - \frac{p_2}{\sigma_2}\rho_{13} - \frac{p_3}{\sigma_3}\rho_{12} \end{aligned}$$

Here, $\rho_{ij} = (\mathbb{E}[x_i x_j] - p_i p_j)/\sigma_i \sigma_j$ and $\rho_{ijk} = (\mathbb{E}[x_i x_j x_k] - p_i p_j p_k)/\sigma_i \sigma_j \sigma_k$. Finally, we can compute the components of $\phi^2 = \phi^2_2 + \phi^2_3$:

$$\begin{aligned} \phi^2_2 &= a^2_{12} + a^2_{13} + a^2_{23} \\ \phi^2_3 &= a^2_{123} \end{aligned}$$

Figure 1 illustrates this decomposition for some simulated distributions.

## 3. KL DEPENDENCE MEASURE

Another useful measure to quantify the statistical dependencies between multiple random variables is the Kullback-Leibler (KL) distance between the joint probability function and its independent counterpart, the product of the

marginals. Again, we let $\mathbf{X} = (X_1, \ldots, X_N)$ be a random vector with $X_n \in \mathcal{X}$, with joint probability distribution $p_{\mathbf{X}}(\mathbf{x})$ and marginal distributions $\{p_{X_n}(x_n)\}$, $n = 1, \ldots, N$. We define

$$\nu = D(p||p_{ind}) = \int p_{\mathbf{X}}(\mathbf{x}) \log \frac{p_{\mathbf{X}}(\mathbf{x})}{\prod_{n=1}^{N} p_{X_n}(x_n)} d\mathbf{x} \quad (5)$$

$\nu$ equals zero if and only if the random variables are statistically independent, and increases without bound (in general) with increasing dependence. For example, for a bivariate Gaussian distribution with correlation coefficient $\rho$, $\nu = -\frac{1}{2}\log(1-\rho^2)$.

Using information geometry, we can decompose $\nu$ in a similar way to $\phi^2$. The decomposition described here is due to Amari[8].

### 3.1. Decomposition of KL dependence

Consider a family of parametric joint probability distributions $\mathcal{M} = \{p(\mathbf{x}; \boldsymbol{\xi})\}$, where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ is a vector of real-valued parameters that fully specify the distribution of the random vector $\mathbf{X}$. In information geometry, this family of distributions is viewed as an $m$-dimensional manifold with $\boldsymbol{\xi}$ as its coordinate system. For example, the joint distribution of $N$ binary random variables is specified by $2^N - 1$ parameters, and thus lies on an $m = 2^N - 1$ dimensional manifold.

In [8], manifolds with two different properties are discussed: *e-flat* manifolds and *m-flat* manifolds. An example of an $e$-flat manifold is the exponential family

$$\log p(\mathbf{x}; \boldsymbol{\theta}) = \sum_i \theta_i g_i(\mathbf{x}) - \psi(\boldsymbol{\theta}) \quad (6)$$

where the $\{g_i\}$ are model-specific functions and $\psi(\boldsymbol{\theta})$ is a normalizing function so that $p$ sums to 1. This family encompasses a broad class of distributions, including the discrete probability mass functions. An example of an $m$-flat manifold is the mixture family

$$p(\mathbf{x}; \boldsymbol{\eta}) = \sum_i \eta_i q_i(\mathbf{x}) \quad (7)$$

where the $\{q_i\}$ are probability density functions, $0 < \eta_i < 1$, and $\sum \eta_i = 1$. While it is not true in general that distributions in the exponential family are also in the mixture family, it is true that manifolds that are $e$-flat are also $m$-flat (and vice-versa). Consequently, it can be shown that for a distribution belonging to either family, the KL dependence measure can be decomposed as $\nu = \sum_{k=2}^{N} \nu_k$, where $\nu_k$ is the component due only to $k^{th}$-order interactions between the random variables.

### 3.2. Dependence ordering in $\nu$

We begin by partitioning the parameters such that each partition contains only parameters that describe the same dependence order. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be a random vector with $X_n \in \mathcal{X}$ and joint probability distribution $p_{\mathbf{X}}(\mathbf{x})$. If $p$ lies on an $e$-flat manifold $\mathcal{E} = \{p(\mathbf{x}; \boldsymbol{\theta})\}$ with coordinates $\boldsymbol{\theta}$, the manifold is also $m$-flat and can be written as $\mathcal{M} = \{p(\mathbf{x}; \boldsymbol{\eta})\}$ with coordinates $\boldsymbol{\eta}$. We rewrite the coordinates in terms of the ordered partitions $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N)$, where $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k$ are the set of all parameters that describe only the interactions of order $k$. Using this partition, we define the submanifolds

$$\begin{aligned} \mathcal{E}_k &= \{p(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta}_{k+1} = 0, \ldots, \boldsymbol{\theta}_N = 0\} \\ \mathcal{M}_k &= \{p(\mathbf{x}; \boldsymbol{\eta}) : \boldsymbol{\eta}_{k+1} = 0, \ldots, \boldsymbol{\eta}_N = 0\} \end{aligned} \quad (8)$$

Thus the submanifolds $\mathcal{E}_k$ and $\mathcal{M}_k$ contain only probability distributions that have no dependencies higher than order $k$. It can be shown that the two submanifolds are complementary and orthogonal at every point, and consequently using the Pythagoras theorem we obtain the decomposition

$$\nu = D(p||p^{(1)}) = \sum_{k=2}^{N} D\left(p^{(k)}||p^{(k-1)}\right) \equiv \sum_{k=2}^{N} \nu_k \quad (9)$$

where $p^{(k)}$ is the projection of $p$ *along the coordinates* $\boldsymbol{\eta}$ to the closest point on the submanifold $\mathcal{E}_k$. Note that this is not equivalent to simply setting $\boldsymbol{\theta}_{k+1} = \cdots = \boldsymbol{\theta}_N = 0$ in the exponential model to obtain a new distribution; rather, to satisfy the orthogonality condition we require a coordinate transformation to obtain a new set of $(\boldsymbol{\eta}_{k+1}, \ldots, \boldsymbol{\eta}_N)$ coordinates for the projection. Thus $\nu$ is decomposed into $N-1$ components, where each component is the contribution of interactions of a different order to the overall dependence between the random variables.

### 3.3. Example: three binary random variables

Consider the example from the previous section. We obtain the parameterization

$$\log p(\mathbf{x}, \boldsymbol{\theta}) = \sum_i \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \theta_{123} x_1 x_2 x_3 - \psi$$

which has the coordinates $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$, where $\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \theta_3)$, $\boldsymbol{\theta}_2 = (\theta_{12}, \theta_{13}, \theta_{23})$, and $\boldsymbol{\theta}_3 = (\theta_{123})$. Similarly, we define an $\boldsymbol{\eta}$ parameterization such that we obtain the coordinates $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3)$, where $\boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$, $\boldsymbol{\eta}_2 = (\eta_{12}, \eta_{13}, \eta_{23})$, and $\boldsymbol{\eta}_3 = (\eta_{123})$. Here, $\eta_i = \mathbb{E}[x_i]$, $\eta_{ij} = \mathbb{E}[x_i x_j]$, and $\eta_{123} = \mathbb{E}[x_1 x_2 x_3]$. We wish to obtain the projection $p^{(2)}$, which is found by setting $\boldsymbol{\theta}_3 = 0$ and solving for the new coordinate $\boldsymbol{\eta}_3$.

To find the transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, we first let $p_{x_1 x_2 x_3} = P[X_1 = x_1, X_2 = x_2, X_3 = x_3]$. Rewriting $\boldsymbol{\theta}$

in terms of these probabilities, we find that

$$\theta_{123} = \log \frac{p_{111}p_{100}p_{010}p_{001}}{p_{000}p_{110}p_{101}p_{011}} \tag{10}$$

Since $\boldsymbol{\eta}$ consists of expected values, we can also write $\boldsymbol{\eta}$ in terms of the probabilities

$$\boldsymbol{\eta} = \mathbf{A}\mathbf{p} \tag{11}$$

where $\mathbf{A}$ is a $(2^N - 1) \times (2^N - 1)$ invertible matrix and $\mathbf{p} = (p_{x_1x_2x_3})_{\mathbf{x}\in\mathcal{X},\, \mathbf{x}\neq(0,0,0)}$ is the vector of probabilities excluding $p_{000}$. Hence, to find $p^{(2)}$, we set equation (10) equal to zero and solve for the new $\tilde{\eta}_{123}$. Then we simply substitute $\tilde{\eta}_{123}$ back into equation (11) and solve for $\mathbf{p}$. Finally, we note that $p^{(1)}$ is simply the independent distribution, the product of the marginals. Thus we obtain the components of $\nu = \nu_2 + \nu_3$,

$$\begin{aligned} \nu_2 &= D(p^{(2)}||p^{(1)}) \\ \nu_3 &= D(p||p^{(2)}) \end{aligned}$$

Figure 1 illustrates this decomposition for some simulated distributions. The simulations show that the results of the two decompositions are similar.

## 4. CONCLUSION

We have described two multivariate statistical dependence measures that can be used to quantify dependencies between large groups of random variables. Both measures can be decomposed orthogonally into different orders of dependence, providing a greater level of detail about the dependencies. In each case, the decomposition applies to probability functions that meet certain conditions: (1) $\phi^2$ and $\nu$ must be bounded; more stringent constraints on the distributions are evident in each case, but both decompositions are valid in general for discrete distributions. (2) For discrete distributions, both decompositions are valid only if $P[\mathbf{X} = \mathbf{x}] > 0\ \forall \mathbf{x}$. In practice, this is ensured by using the Krichevsky-Trofimov [9] method to estimate the probability distributions.

Although both measures provide a strong expression of dependence, practical considerations dictate which measure should be used for a given application. For example, for large groups of random variables, decomposing $\nu$ requires solving large systems of non-linear equations, and hence may be too computationally intensive to be of practical use. $\phi^2$ is decomposed by computing sets of orthonormal functions on the marginal densities, which can be accomplished much more efficiently. It is important, however, to also consider the statistical properties of the two measures; this consideration is especially important in applications like neural population analysis that are severely data-limited. Empirical methods such as the bootstrap can be used to remove bias and estimate confidence intervals for both measures; however, we still require the complementary theoretical work to determine how much data are needed to achieve a specified confidence level for each measure.

## 5. REFERENCES


[1] M. Bezzi, M.E. Diamond, and A. Treves, "Redundancy and synergy arising from pairwise correlations in neuronal ensembles," *Journal of Computational Neuroscience*, vol. 12, no. 3, pp. 165–174, May-June 2002.

[2] J. W. Fisher III, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," *Advances in Neural Information Processing Systems*, Nov. 2000.

[3] J. Hershey and J. Movellan, "Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*, S. A. Solla, T. K. Leen, and K-R. Mller, Eds., pp. 813–819. MIT Press, 1999.

[4] H. Joe, "Relative entropy measures of multivariate dependence," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 157–164, March 1989.

[5] O.V. Sarmanov, "Maximum correlation coefficient (nonsymmetric case)," in *Selected Translations in Mathematical Statistics and Probability*, vol. 2, pp. 207–210. Amer. Math. Soc., 1962.

[6] H.O. Lancaster, "The structure of bivariate distributions," *Ann. Math. Statistics*, vol. 29, no. 3, pp. 719–736, September 1958.

[7] R. R. Bahadur, "A representation of the joint distribution of responses to *n* dichotomous items," in *Studies in Item Analysis and Prediction*, H. Solomon, Ed., pp. 158–168. Stanford University Press, 1961.

[8] S. Amari, "Information geometry on hierarchy of probability distributions," *IEEE Trans. Info. Theory*, vol. 47, no. 5, pp. 1701–1711, July 2001.

[9] R. Krichevsky and V. Trofimov, "The performance of universal encoding," *IEEE Trans. Information Theory*, vol. 27, no. 2, pp. 199–207, Mar. 1981.