AN UPPER-BOUND ON SECOND-ORDER DEPENDENCY

Siwei Lyu

Computer Science Department
College of Computing and Information
University at Albany, State University of New York
slyu@albany.edu

ABSTRACT

In this work, we study the upper bound of second order statistical correlation. We provide a condition for a random variable reaching the upper-bound, and an algorithm that transform any variable to have the maximum second order statistical correlation.

***Index Terms*—** Second order dependency, Upper-bound, Information theory

1. INTRODUCTION

Statistical dependency is the pivotal subject in multivariate statistics and signal processing. The complete statistical dependency in a d-dimensional random vector $\mathbf{x}$ is measured by the multi-information (MI) [1], which is also the Kulback-Leibler divergence [2] between the joint distribution and the product of its marginals, as:

$$I(\mathbf{x}) = D_{\mathrm{KL}}\left(p(\mathbf{x}) \middle\| \prod_k p(x_k)\right) = \sum_{k=1}^{d} H(x_k) - H(\mathbf{x}),$$

where $H(\mathbf{x}) = \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ is the entropy of $\mathbf{x}$, and $H(x_k)$ denotes the differential entropy of the kth component of $\mathbf{x}$. When $\mathbf{x}$ is a multivariate Gaussian vector with covariance matrix C, its multi-information, $I(\mathbf{x})$, is given by

$$J(\mathbf{x}) = \sum_{k=1}^{d} \log c_{kk} - \log|\det C|, \qquad (1)$$

where c_{kk} is the $\mathrm{k^{th}}$ diagonal of C. As a Gaussian random variable only has second-order statistical dependencies, Eq.(1) also measures second-order statistical dependency in any random variable $\mathbf{x}$.

It is known that the measure of second order statistical dependency $J(\mathbf{x})$ is lower-bounded by zero, and the lower-bound is achieved with $\mathbf{x}$ whose components are de-correlated (equivalently, C is a diagonal matrix). For any random vector $\mathbf{x}$, we can transform it with the eigen-decomposition of the covariance matrix C, a procedure known as principal component analysis [3] to remove second-order dependency. PCA is also an important pre-processing step for methods aiming to remove higher-order dependency such as ICA [4, 5]. On the other hand, there is little previous work focusing on the *upper-bound* of the second-order statistical dependencies of a random variable $\mathbf{x}$. On the one hand, recent work on comparing dependency reduction methods often use second-order dependency as a base [6, 7]. It is therefore important to know the upper-bound of second-order dependency, which will quantify the maximum reduction in dependency achieved by PCA. On the other hand, in application fields (e.g., neural sciences [8]) there is a practical need to generate random signals with varying degree of second-order dependencies, where efficient algorithm that can turn a random variable with highest correlation is useful.

Here, we first provide a theoretical analysis on the upper-bound of the second-order dependencies, showing that it is achieved with a random variable whose covariance matrix has constant diagonal. We further provide an efficient algorithm that can transform any random vector to reach its upper-bound of second order dependencies using only orthonormal algorithms (i.e., rotations). We provide pseudo code for our algorithm.

2. LOWER AND UPPER BOUNDS OF $J(\mathbf{x})$

Here, we study the range of $J(\mathbf{x})$ with regards to an orthonormal transformation W of $\mathbf{x}$, i.e., $W^T W = W W^T = I_d$, where I_d is the d-dimensional identity matrix. Note that $W\mathbf{x}$ does not change the ℓ_2 norm of $\mathbf{x}$, and corresponds to a rotation of $\mathbf{x}$ in the d-dimensional space. The main reason for focusing on the orthonormal transformations is to avoid the effect of scaling on $J(\mathbf{x})$.

It is a well-known fact that $J(\mathbf{x})$ is lower bounded from zero, i.e., for any orthonormal W, we have $J(W\mathbf{x}) \geq 0$ [2], which can also be shown with the Hadamard inequality for positive definite matrix C,

$$|\det C| \leq \prod_{k=1}^{d} c_{kk}.$$

Furthermore, for a given random vector $\mathbf{x}$, this lower bound of $J(\mathbf{x})$ can be achieved by an orthonormal transformation of $\mathbf{x}$ using the eigenvectors of the covariance matrix of $\mathbf{x}$. Specifically, assuming $\text{cov}(\mathbf{x}) = C$, by definition, C is symmetric and positive definite, therefore according to the spectral theorem, we have $C = U\Lambda U^T$, where U is an orthonormal matrix corresponding to the eigenvectors of C, and Λ is a diagonal matrix whose diagonal corresponding to the positive eigenvalues of C.

In this work, however, we are interested in the upper-bound of $J(W\mathbf{x})$. We first give the upper-bound for $J(W\mathbf{x})$ using the following theorem.

Theorem 1 *For a full rank symmetric and positive definite matrix C, we have*

$$\sum_{k=1}^{d} \log c_{kk} \leq d \log \frac{\text{tr}\,(C)}{d}. \tag{2}$$

The equality holds if and only if

$$c_{kk} = \frac{\text{tr}\,(C)}{d}, \text{ for } k = 1, \cdots, d.$$

Proof (Theorem 1): First, we rewrite the left hand side of (2) as:

$$\sum_{k=1}^{d} \log c_{kk} = \log \prod_{k=1}^{d} c_{kk}.$$

Next, the arithmetic mean - geometric mean (AM-GM) inequality [9] states that we have

$$\left(\prod_{k=1}^{d} c_{kk}\right)^{1/d} \leq \frac{1}{d} \sum_{k=1}^{d} c_{kk}.$$

Raising to the dth power of both sides, we have

$$\sum_{k=1}^{d} \log c_{kk} = \log \prod_{k=1}^{d} c_{kk} \leq \log \left(\frac{\sum_{k=1}^{d} c_{kk}}{d}\right)^d = d \log \frac{\text{tr}\,(C)}{d}.$$

■

Now we apply Theorem 1 to the the second-order dependency $J(\mathbf{x})$ of a random vector $\mathbf{x}$ with regards to orthonormal transformations. Assuming W is an orthonormal transform and C is the covariance matrix of $\mathbf{x}$, it is not difficult to see that $\tilde{C} = W^T C W$ is the covariance matrix of $W\mathbf{x}$. Then we have

$$\begin{aligned}
J(W\mathbf{x}) &= \sum_{k=1}^{d} \log \tilde{c}_{kk} - \log|\det \tilde{C}| \\
&\leq d \log \frac{\text{tr}\left(W^T C W\right)}{d} - \log|\det W^T C W| \\
&\leq d \log \frac{\text{tr}\left(C W W^T\right)}{d} - 2\log|\det W^T| - \log|\det C| \\
&\leq d \log \frac{\text{tr}\,(C)}{d} - \log|\det C|
\end{aligned}$$

The last step in the proof uses the fact that $W^T W = I_d$ and $|\det W| = |\det W^T| = 1$. Furthermore, the upper bound of $J(W\mathbf{x})$ is reached when the covariance matrix of the transformed variable $\tilde{C} = W^T C W$ has constant diagonals.

It should be point out that this upper bound cannot be achieved is $\mathbf{x}$ is whitened, i.e., its covariance matrix is a multiple of the identity matrix as $C = \sigma I_d$. Under this case, for any orthonormal W, we have $W^T C W = \sigma W^T W = \sigma I_d$. In other words, no rotation will be able to change its second-order dependencies.

3. ACHIEVING THE UPPER-BOUND

While the previous result shows the upper bound of the second order dependency measure $J(\mathbf{x})$. In this section, we further show that this upper bound can be achieved for any non-whitened random vector $\mathbf{x}$ with an orthonormal transformation W. Our proof is constructive, in the sense that it also provides the algorithm that recovers such an orthonormal transform.

Our main result is summarized in the following theorem.

Theorem 2 *For a full rank symmetric and positive definite matrix C, there exists an orthonormal matrix W such that $W^T C W$ has constant diagonal.*

Proof (Theorem 2): We prove the theorem by an induction on the dimension of matrix C.

We first consider the case of $d = 2$. Because C is a 2×2 symmetric and positive definite matrix, based on the spectral theorem, we can decompose it as $C = U_1 \Lambda U_1^T$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

is a diagonal matrix with the eigenvalues of C on the diagonal, and U_1 denotes the orthonormal matrix with the eigenvectors of C. Next, note that we can use orthonormal matrix

$$U_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix},$$

and transform Λ to have constant diagonal as

$$U_2^T \Lambda U_2 = \begin{pmatrix} \frac{1}{2}(\lambda_1 + \lambda_2) & \frac{1}{2}(\lambda_1 - \lambda_2) \\ \frac{1}{2}(\lambda_1 - \lambda_2) & \frac{1}{2}(\lambda_1 + \lambda_2) \end{pmatrix}.$$

We can then multiply the two orthonormal matrices to form another orthonormal matrix $U = U_1 U_2$, which can transform

$$U^T C U = U_2^T U_1^T C U_1 U_2 = \begin{pmatrix} \frac{1}{2}(\lambda_1 + \lambda_2) & \frac{1}{2}(\lambda_1 - \lambda_2) \\ \frac{1}{2}(\lambda_1 - \lambda_2) & \frac{1}{2}(\lambda_1 + \lambda_2) \end{pmatrix}.$$

Next, consider a general $d\times d$ ($d \geq 2$) symmetric and positive definite matrix C that has eigen-decomposition as $C = U_1 \Lambda U_1^T$, where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

contains its eigenvalues and U_1 is the orthonormal matrix containing its eigenvectors.

Denote $\mathbf{1} = (1, \cdots, 1)^T$ as the d dimensional vector with all components being one. We consider the d-dimensional vector $\frac{1}{\sqrt{d}}\mathbf{1}$, and denote $U = (\mathbf{u}_1, \cdots, \mathbf{u}_{d-1})$ as the matrix that contains $d-1$ orthogonal unit vectors that span the quotient space of $\frac{1}{\sqrt{d}}\mathbf{1}$. In other words, $\frac{1}{\sqrt{d}}\mathbf{1}^T \mathbf{u}_i = 0$, and $\mathbf{u}_i^T \mathbf{u}_i = 1$ for any $i = 1, \cdots, d-1$. These can be obtained using, for instance, a Gram-Schmidt orthogonalization procedure [10].

Note that the matrix formed by combining all the vectors

$$\tilde{U} = \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1} & U \end{pmatrix}$$

is an orthonormal matrix, and

$$\frac{1}{\sqrt{d}}\mathbf{1}^T \Lambda \frac{1}{\sqrt{d}}\mathbf{1} = \frac{1}{d}\mathrm{tr}\,(C).$$

The last step uses the fact that the sum of eigenvalues of a matrix equals to its trace.

Putting together, we have

$$\begin{aligned} \tilde{U}^T \Lambda \tilde{U} &= \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1}^T \Lambda \frac{1}{\sqrt{d}}\mathbf{1} & \frac{1}{\sqrt{d}}\mathbf{1}^T \Lambda U \\ U^T \Lambda \frac{1}{\sqrt{d}}\mathbf{1} & U^T \Lambda U \end{pmatrix} \\ &= \begin{pmatrix} \frac{\mathrm{tr}(C)}{d} & \frac{1}{\sqrt{d}}\mathbf{1}^T \Lambda U \\ \frac{1}{\sqrt{d}} U^T \Lambda \mathbf{1} & U^T \Lambda U \end{pmatrix}, \end{aligned}$$

from which, we can see that

$$\begin{aligned} \mathrm{tr}\left(U^T \Lambda U\right) &= \mathrm{tr}\left(\tilde{U}^T \Lambda \tilde{U}\right) - \frac{1}{d}\mathrm{tr}\,(C) \\ &= \mathrm{tr}\,(C) - \frac{1}{d}\mathrm{tr}\,(C) = \frac{d-1}{d}\mathrm{tr}\,(C). \end{aligned}$$

Next, using the induction assumption that for any $(d-1)\times(d-1)$ symmetric and positive definite matrix, we can find an orthonormal matrix that can transform it to have constant diagonal. Assume that for matrix $U^T \Lambda U$, such an orthonormal matrix is V, such that $V^T U^T \Lambda U V$ has constant diagonal, whose diagonal elements all have value $\frac{1}{d-1}\mathrm{tr}\left(U^T \Lambda U\right) = \frac{1}{d-1}\frac{d-1}{d}\mathrm{tr}\,(C) = \frac{\mathrm{tr}(C)}{d}$. Therefore, we can form a matrix $U_2 = \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1} & UV \end{pmatrix}$, which is an orthonormal matrix as

$$U_2^T U_2 = \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1}^T \\ V^T U^T \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1} & UV \end{pmatrix} = I,$$

and

$$U_2 U_2^T = \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1} & UV \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{d}}\mathbf{1}^T \\ V^T U^T \end{pmatrix} = \tilde{U}\tilde{U}^T = I.$$

Furthermore, we have

$$U_2^T \Lambda U_2 = \begin{pmatrix} \frac{\mathrm{tr}(C)}{d} & \frac{1}{\sqrt{d}}\mathbf{1}^T \Lambda U V \\ \frac{1}{\sqrt{d}} V^T U^T \Lambda \mathbf{1} & V^T U^T \Lambda U V \end{pmatrix},$$

which has constant diagonal of value $\frac{1}{d}\mathrm{tr}\,(C)$. Finally, we form another orthonormal matrix $W = U_1 U_2$, and it is straightforward to see that $W^T C W = U_2^T \Lambda U_2$ has constant diagonal. ■

Based on the proof of Theorem 1, we obtain the following algorithm, which is given in MATLAB code in Fig.1.

```
function [U] = equalDiag(A)

D = size(A,1);
[V,e] = eig(A);
if D == 2
   U = [1 -1;1 1]/sqrt(2);
   U = V*U;
else
   v = ones(D,1)/sqrt(D);
   B = eye(D);
   B(:,1) = v;
   B = grams(B); % Gram-Schmidt
   U = B(:,2:end);
   U = V*[v U*equalDiag(U'*e*U)];
end

return
```

Fig. 1. *MATLAB code for finding the orthonormal transform making a p.d. matrix to have constant diagonal.*

4. CONCLUSION

In this work, we study the upper bound of second order statistical correlation. We provide a condition for a random variable reaching the upper-bound, and an algorithm that transform any variable to have the maximum second order statistical correlation.

Acknowledgement. This work is supported by a National Science Foundation of the United States CAREER Award IIS-0953373, and a National Science Foundation of the United States research grant IIS-1208463.

5. REFERENCES

[1] M. Studeny and J. Vejnarova, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in Graphical Models*, M. I. Jordan, Ed., pp. 261–297. Dordrecht: Kluwer., 1998.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, 2nd edition, 2006.

[3] I.T. Jolliffe, *Principal Component Analysis*, Springer, 2nd edition, 2002.

[4] A J Bell and T J Sejnowski, "The 'independent components' of natural scenes are edge filters," vol. 37, no. 23, pp. 3327–3338, 1997.

[5] J H van Hateren and A van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. R. Soc. Lond. B*, vol. 265, pp. 359–366, 1998.

[6] Matthias Bethge, "Factorial coding of natural images: how effective are linear models in removing higher-order dependencies?," *J. Opt. Soc. Am. A*, vol. 23, no. 6, pp. 1253–1268, 2006.

[7] S Lyu and E P Simoncelli, "Nonlinear extraction of 'independent components' of natural images using radial Gaussianization," *Neural Computation*, vol. 18, no. 6, pp. 1–35, 2009.

[8] O Schwartz, E J Chichilnisky, and E P Simoncelli, "Characterizing neural gain control using spike-triggered covariance," in *Adv. Neural Information Processing Systems (NIPS*01)*, T G Dietterich, S Becker, and Z Ghahramani, Eds., Cambridge, MA, May 2002, vol. 14, pp. 269–276, MIT Press.

[9] Michael D. Hirschhorn, "The am-gm inequality," *Mathematical Intelligencer*, vol. 29, no. 4, pp. 7, 2007.

[10] G. Strang, *Introduction to linear algebra*, Wellesley-Cambridge Press Wellesley, MA, 1993.