Pairwise Symmetry Decomposition Method for Generalized Covariance Analysis

Tsuyoshi Idé
IBM Research, Tokyo Research Laboratory
1623-14 Shimo-tsuruma, Yamato, Kanagawa 242-8502, Japan
goodidea@jp.ibm.com

Abstract

We propose a new theoretical framework for generalizing the traditional notion of covariance. First, we discuss the role of pairwise cross-cumulants by introducing a cluster expansion technique for the cumulant generating function. Next, we introduce a novel concept of symmetry decomposition of probability density functions according to the $C_{4\mathrm{v}}$ group. By utilizing the irreducible representations, generalized covariances are explicitly defined, and their utility is demonstrated using an analytically solvable model.

1. Introduction

Correlation analysis for multivariate systems is one of the major topics in data mining. In spite of the importance, however, most of the practical correlation analysis methods for real-valued data are essentially based on the traditional Gaussian distribution. The (partial) covariance matrix is the measure of correlation between variables in Gaussian distributions. However, it is well-known that the covariance can correctly describe phenomena only in the vicinity of a linear correlation. One typical example is the fact that the covariance is zero if (x, y) is distributed on a circle. Although x is strongly correlated with y in this case, the covariance clearly fails in capturing the correlation.

To capture the nonlinearities, kernel-based methods have been actively studied for the last decade. However, kernel methods are essentially "black boxes," where what kind of correlations one discovers depends greatly on a possibly accidental choice of a suitable kernel.

In this paper, we propose a new theoretical framework for generalizing traditional covariance analysis. First, in Section 2, under an approximation called sparse correlation approximation, we show that pairwise cross-cumulants can suffice to describe nonlinear correlations without unwanted disturbances of the heterogeneity. Next, in Section 3, we introduce a novel concept of symmetry decomposition of probability density functions (pdf) according to the $\mathrm{C}_{4\mathrm{v}}$ group. To generalize the notion of covariance, we propose an idea of regarding pairwise functional relationships as two-dimensional (2D) geometric patterns in the 2D configuration spaces, where the irreducible representations of the $\mathrm{C}_{4\mathrm{v}}$ group are utilized to characterize the patterns. To the best of the author's knowledge, this is the first attempt to reduce the task of pattern recognition to discovery of irreducible representations. After giving explicit definitions of the generalized covariances in Section 4, we demonstrate the capability of the generalized covariances based on an analytically solvable model in Section 5.

2. Pairwise Cross-Cumulants

Consider a system whose internal state is described with an n-dimensional random vector $\boldsymbol{x} = (x_1, ..., x_n)^T$. We expect that the pdf $p(\boldsymbol{x})$ contains all of the information about the internal structure of the system. The statistical properties of p are completely determined by the cumulant generating function $\Psi(\boldsymbol{s})$:

$$\Psi(\boldsymbol{s}) \equiv \ln \int d\boldsymbol{x}\; p(\boldsymbol{x}) \exp(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x}) = \ln \left\langle \exp(\boldsymbol{s}^{\mathrm{T}}\boldsymbol{x}) \right\rangle, \quad (1)$$

where $\langle \cdot \rangle$ denotes the expectation with respect to $p(\boldsymbol{x})$. The multivariate cumulants are defined as the coefficients of the Taylor expansion with respect to $\boldsymbol{s}$. For example, we have $\langle x_i x_j \rangle_{\mathrm{c}} = \langle x_i x_j \rangle$ and $\langle x_i^2 x_j^2 \rangle_{\mathrm{c}} = \langle x_i^2 x_j^2 \rangle - \langle x_i^2 \rangle \langle x_j^2 \rangle - 2\langle x_i x_j \rangle^2$ for zero-mean data (which we assume hereafter). Here we introduced a notation of *cumulant average* $\langle \cdot \rangle_{\mathrm{c}}$ to represent multivariate cumulants [4]. For the readers' convenience, we summarize the relationships between the cross-cumulants and the moments in Table 1. For higher order cumulants, the following properties are well-known [6]:

Theorem 1 *A cross-cumulant is zero if there is at least one pair of statistically independent variables inside* $\langle \ \rangle_{\mathrm{c}}$.

Theorem 2 *All of the higher order cumulants of* $k \geq 3$ *vanish for the Gaussian.*

Thus, one needs to take account of the higher cumulants in order to go beyond the Gaussian distribution. Conversely, one may think that approximating $\Psi(\boldsymbol{s})$ using a finite number of higher order terms might be reasonable, since traditionally the Gaussian has been used as a practical solution for real-valued multivariate correlation analysis.

Let us rewrite $\Psi(\boldsymbol{s})$ as $\Psi(\boldsymbol{s}) = K_1 + K_2 + ... + K_i + ...$, where K_i denotes the summation of terms including i different variables. For instance, K_2 is given by

$$K_2 = \sum_{i \neq j} \left[\frac{1}{2!} s_i s_j \langle x_i x_j \rangle_\mathrm{c} + \frac{1}{3!} {s_i}^2 s_j \langle {x_i}^2 x_j \rangle_\mathrm{c} + \dots \right].$$

We call each term of K_i an i-body cluster after statistical physics [4].

Now let us approximate $\Psi(\boldsymbol{s})$ using a finite number of clusters. We here make an assumption of *sparse correlation*. Under this assumption, the larger the number of different variables inside $\langle\ \rangle_\mathrm{c}$, the greater the likelihood that the cumulant vanishes. Thus, the contribution of higher order clusters would be negligible in the cluster expansion. Therefore, the lowest nontrivial approximation should be $\Psi(\boldsymbol{s}) \simeq K_1 + K_2$, which we call the sparse correlation approximation (SCA).

Mathematically, this approximation would be valid when the average number of correlated variables on each variable is on the order of one. However, the leading term which describe the correlation must be K_2, even when the condition does not exactly hold. Since the one-body cluster K_1 only gives us the information about the marginal distribution of the individual variables, our basic quantities for correlation analysis are the pairwise cross-cumulants of the type $\langle {x_i}^\mu {x_j}^\nu \rangle_\mathrm{c}$, where μ and ν are nonzero integers. To be robust against the diversity of nonlinear correlations, we further approximate K_2 to include only the terms of $\mu+\nu \leq 4$, i. e. we confine ourselves within the *fourth order SCA*.

According to Theorem 1, the pairwise cross-cumulants are zero if the two variables are statistically independent. Note that this fact has nothing to do with *how* independent they are. In fact, a constant distribution and an uncorrelated Gaussian distribution equally give the value zero. This is a very good property in that the heterogeneity is mitigated in terms of statistical independence.

3. Symmetry Decomposition of pdf

Let $\boldsymbol{r} = (x, y)^T$ be a pair of variables arbitrarily chosen from $\boldsymbol{x}$. To facilitate the discussion, we adopt Dirac's bra-ket notation to represent the pairwise pdf $p(\boldsymbol{r})$ [3]. We define $|p\rangle$, a state vector in a Hilbert space $\mathcal{H}$, by $\langle \boldsymbol{r}|p\rangle = p(\boldsymbol{r})$, where $|\boldsymbol{r}\rangle \in \mathcal{H}$ is the position eigenket [5]. The inner product between $|f\rangle, |h\rangle \in \mathcal{H}$ can be calculated as $\langle f|h\rangle = \int d\boldsymbol{r} f(\boldsymbol{r}) h(\boldsymbol{r})$. Since $\langle \boldsymbol{r}|\boldsymbol{r}'\rangle$ equals to Dirac's delta function $\delta(\boldsymbol{r} - \boldsymbol{r}')$, we have $\langle \boldsymbol{r}|p\rangle = \int d\boldsymbol{r}' \langle \boldsymbol{r}|\boldsymbol{r}'\rangle \langle \boldsymbol{r}'|p\rangle = p(\boldsymbol{r})$,

order	cumulant	moment
2	$\langle xy \rangle_\mathrm{c}$	$\langle xy \rangle$
3	$\langle xy^2 \rangle_\mathrm{c}$	$\langle xy^2 \rangle$
3	$\langle x^2 y \rangle_\mathrm{c}$	$\langle x^2 y \rangle$
4	$\langle x^2 y^2 \rangle_\mathrm{c}$	$\langle x^2 y^2 \rangle - \langle x^2 \rangle \langle y^2 \rangle - 2\langle xy \rangle^2$
4	$\langle xy^3 \rangle_\mathrm{c}$	$\langle xy^3 \rangle - 3\langle xy \rangle \langle y^2 \rangle$
4	$\langle x^3 y \rangle_\mathrm{c}$	$\langle x^3 y \rangle - 3\langle x^2 \rangle \langle xy \rangle$

Table 1. Relationships between the pairwise cumulants and the moments for $\langle x \rangle = \langle y \rangle = 0$.

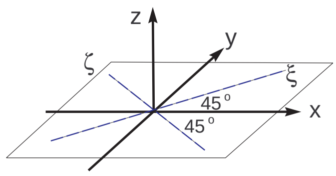

Figure 1. Symmetry axes of the $\mathrm{C}_{4\mathrm{v}}$ group.

as expected. A cross-moment $\langle x^\mu y^\nu \rangle$ is now represented as $\langle p|x^\mu y^\nu \rangle$, where $|x^\mu y^\nu\rangle$ is defined by $\langle \boldsymbol{r}|x^\mu y^\nu\rangle = x^\mu y^\nu$.

It is quite interesting to study the symmetry properties of $|x^\mu y^\nu\rangle$. Consider a set of symmetry operations defined within the xy-space. We request the operations to satisfy the axioms of a group: For a group G associated with a product operation $\circ$, (1) *Closure*. For $\forall a \in G$ and $\forall b \in G$, $(a \circ b) \in G$. (2) *Associativity*. $(a \circ b) \circ c = a \circ (b \circ c)$ for all a, b, and $c \in G$. (3) *Identity*. There exists $e \in G$, such that $e \circ g = g = g \circ e$ for all $g \in G$. (4) *Inverse*. For each $g \in G$, there exists the g', the inverse of g, such that $g' \circ g = g \circ g' = e$.

Unexpectedly, there are only 32 groups that can satisfy the axioms with rotations and mirror reflections [2]. Among the 32 point groups, the most general one within the 2D xy-space is a group named $\mathrm{C}_{4\mathrm{v}}$. Figure 1 shows the symmetry axes of this group, which contains eight symmetry operations:

$$\mathrm{C}_{4\mathrm{v}} = \{e, C_4, C_2, {C_4}^3, \sigma_x, \sigma_y, \sigma_\xi, \sigma_\eta\},$$

where e is the identity element. Operations C_4, C_2, and ${C_4}^3$ are $\pi/2$-, π-, and $3\pi/2$-rotations around the z-axis, respectively. Mirror reflections with respect to the xz-, yz-, ξz- and ηz-planes are represented as $\sigma_x, \sigma_y, \sigma_\xi, \sigma_\eta$, respectively. Note that the sufficiency of such symmetry operations is not necessarily guaranteed when they are arbitrarily chosen. It is the axioms of a group that guarantees the sufficiency of symmetry operations.

For $|f\rangle \in \mathcal{H}$ and $g \in \mathrm{C}_{4\mathrm{v}}$, we define $g|f\rangle$ by $\langle \boldsymbol{r}|g|f\rangle = \langle g^{-1}\boldsymbol{r}|f\rangle = f(g^{-1}\boldsymbol{r})$. A space spanned by a set of linearly independent bases $\{|\phi_1\rangle, .., |\phi_l\rangle\} \subset \mathcal{H}$ is said to be an invariant subspace with respect to a group G if

$$g|\phi_j\rangle = \sum_{i=1}^{l} |\phi_i\rangle D_{ij}(g) \quad (2)$$

is satisfied for $\forall g \in G$. Specifically, a state in this subspace remains in the same subspace even after transformation by any operation of G. The matrix $D_{ij}(g)$ is called a representation matrix for g.

Using the fact that $\langle \boldsymbol{r}|C_4|x^\mu y^\nu\rangle = y^\mu(-x)^\nu$ and $\langle \boldsymbol{r}|\sigma_x|x^\mu y^\nu\rangle = x^\mu(-y)^\nu$, etc., one can easily see that the state vector $|xy\rangle$ spans a one-dimensional (1D) invariant subspace. In fact, the representation matrices are 1 for $e, C_2, \sigma_\xi, \sigma_\eta$, and -1 for $C_4, {C_4}^3, \sigma_x, \sigma_y$. Since any direct-product space spanned by bases such as $|\phi_i\rangle \otimes |\phi_j\rangle$ can be an invariant subspace, the dimension of an invariant subspace does not have an upper bound. On the other hand, some lower bounds exist: a fundamental result of the theory of finite groups is that any invariant subspace can be expressed as a direct sum of a finite number of types of irreducible representation spaces [2]. This fact leads to the orthogonal relation between the irreducible representations

$$\langle \varphi^{(\gamma)}|\varphi^{(\gamma')}\rangle \propto \delta_{\gamma,\gamma'}, \quad (3)$$

where γ and γ' are symbols to identify irreducible representations, and $\delta_{\gamma,\gamma'}$ is Kronecker's delta function. Therefore, a pairwise marginal distribution function $p(\boldsymbol{r})$ can be *decomposed with respect to the symmetries* [1] as

$$p(\boldsymbol{r}) = \sum_{\gamma} \pi^{(\gamma)}(\boldsymbol{r}), \quad (4)$$

where we used the $\boldsymbol{r}$-representation for clarity. These facts prove the following theorem:

Theorem 3 *A state vector $|\varphi^{(\gamma)}\rangle$ in an irreducible representation subspace γ satisfies $\langle \varphi^{(\gamma)}|p\rangle = \langle \varphi^{(\gamma)}|\pi^{(\gamma)}\rangle$, i.e., it works as a symmetry filter for $|p\rangle$.*

Generally, irreducible representations are classified by their characters, i.e., the trace of representation matrices. For the $\mathrm{C_{4v}}$ group, there are known to be five irreducible representations named $\mathrm{A_1}$, $\mathrm{A_2}$, $\mathrm{B_1}$, $\mathrm{B_2}$, and E. The E representation is 2D while the others are 1D. For $\gamma = \mathrm{E}$, the function $\varphi^{(\gamma)}$ in Eq. (4) can be understood as a linear combination of the two orthogonal bases of E. Comparing the aforementioned result with the character table in Table 2, we have an important theorem:

Theorem 4 *$\{|xy\rangle\}$ spans the B_2 representation.*

This theorem clearly shows the way to generalize the notion of covariance. Out of the five irreducible representations, only a single symmetry has been used so far. Now, one can utilize the other symmetries to describe correlations.

[1] It is instructive to consider another group called $\mathrm{C}_i = \{e, I\}$, where I denotes space inversion. In this case, this decomposition corresponds to that between even and odd functions.

$\mathrm{C_{4v}}$	e	$C_4, {C_4}^3$	C_2	σ_x, σ_y	σ_ξ, σ_η
$\mathrm{A_1}$	1	1	1	1	1
$\mathrm{A_2}$	1	1	1	-1	-1
$\mathrm{B_1}$	1	-1	1	1	-1
$\mathrm{B_2}$	1	-1	1	-1	1
E	2	0	-2	0	0

Table 2. The character table of the $\mathrm{C_{4v}}$ group.

4 Generalized Covariances

As an example beyond $|xy\rangle$, consider a state vector $|x^2y^2\rangle \in \mathcal{H}$. It is easy to verify $\langle \boldsymbol{r}|g|x^2y^2\rangle = x^2y^2$ for $\forall g \in \mathrm{C_{4v}}$, so that all of the representation matrices are the 1×1 identity matrix. Table 2 shows that this state spans the $\mathrm{A_1}$ representation. Similarly, one can verify that $|xy^3\rangle + |x^3y\rangle$ is $\mathrm{A_2}$, and $\{|xy^2\rangle + |x^2y\rangle, |xy^2\rangle - |x^2y\rangle\}$ span the E representation.

Let us relate these states to the pairwise cross-cumulants. For $\mathrm{C_{4v}}$, the following theorem holds (proof omitted):

Theorem 5 *$\langle p|x^\mu y^\nu\rangle$ has the same symmetry as $\langle x^\mu y^\nu\rangle_\mathrm{c}$.*

Using this, we define the *generalized covariances in the $\mathrm{C_{4v}}$ sense* as [2]

$$\begin{aligned}
\mathsf{C}(\mathrm{B_2}) &= \langle xy\rangle_\mathrm{c} && (5)\\
\mathsf{C}(\mathrm{E_1}) &= \left[\langle xy^2\rangle_\mathrm{c} + \langle x^2y\rangle_\mathrm{c}\right]/2 && (6)\\
\mathsf{C}(\mathrm{E_2}) &= \left[\langle xy^2\rangle_\mathrm{c} - \langle x^2y\rangle_\mathrm{c}\right]/2 && (7)\\
\mathsf{C}(\mathrm{A_1}) &= \langle x^2y^2\rangle_\mathrm{c} && (8)\\
\mathsf{C}(\mathrm{A_2}) &= \left[\langle xy^3\rangle_\mathrm{c} - \langle x^3y\rangle_\mathrm{c}\right]/2 && (9)
\end{aligned}$$

where the two degrees of freedom in the E representation are distinguished using the subscripts 1 and 2. Clearly, these are symmetric cross-cumulants up to the fourth order according to the $\mathrm{C_{4v}}$ group. To make those quantities dimensionless, it is useful to divide by $\left[\langle x^2\rangle\langle y^2\rangle\right]^{(\mu+\nu)/4}$. Evidently, this normalization factor transforms according to $\mathrm{A_1}$, so that the symmetries of the generalized covariances are not affected. We call the normalized covariances the *generalized correlation coefficients*.

5. Experiment

To see the capability of detecting nonlinear correlations, consider a theoretical model of a correlated time series as

$$\begin{aligned}
x(t) &= \sqrt{2}\cos(\omega_1 t + \alpha)\\
y(t) &= \sqrt{2}\sin(\omega_2 t + \beta)
\end{aligned}$$

[2] One cannot construct a $\mathrm{B_1}$ representation using the quantities $\langle x^\mu y^\nu\rangle_\mathrm{c}$ when $\mu, \nu > 0$ and $\mu + \nu \leq 4$.

with a constant pdf over the time domain. Clearly, the averages $\langle x \rangle$ and $\langle y \rangle$ are zeros, and the variances $\langle x^2 \rangle_{\mathrm{c}}$ and $\langle y^2 \rangle_{\mathrm{c}}$ are ones. Utilizing the fact that $\langle \sin(at+b) \rangle = 0$ unless the constant a is zero, we can derive analytical expressions for the generalized covariance:

$$
\begin{aligned}
\mathsf{C}(\mathrm{B}_2) &= \delta_{\omega_1,\omega_2} \sin \Omega_1^{\beta,\alpha} \\
\mathsf{C}(\mathrm{E}_1) &= -\frac{\delta_{\omega_1,2\omega_2}}{\sqrt{2}} \cos \Omega_2^{\alpha,\beta} + \frac{\delta_{2\omega_1,\omega_2}}{\sqrt{2}} \sin \Omega_2^{\beta,\alpha} \\
\mathsf{C}(\mathrm{E}_2) &= -\frac{\delta_{\omega_1,2\omega_2}}{\sqrt{2}} \cos \Omega_2^{\alpha,\beta} - \frac{\delta_{2\omega_1,\omega_2}}{\sqrt{2}} \sin \Omega_2^{\beta,\alpha} \\
\mathsf{C}(\mathrm{A}_1) &= -\frac{\delta_{\omega_1,\omega_2}}{2} \left[1 + 2\sin^2(\Omega_1^{\alpha,\beta}) \right] \\
\mathsf{C}(\mathrm{A}_2) &= \frac{\delta_{\omega_1,3\omega_2}}{4} \sin \Omega_3^{\alpha,\beta} - \frac{\delta_{3\omega_1,\omega_2}}{4} \sin \Omega_3^{\beta,\alpha},
\end{aligned}
$$

where we used the symbol $\Omega_c^{a,b} = a - bc$.

Figure 2 shows trajectories and generalized covariances for several combinations of the parameters, as shown beside each of the trajectories. The trajectories are well known as Lissajous' trajectories. It is remarkable that the generalized covariances effectively detects the nonlinearities in the trajectories from (b) through (e), where the traditional covariance $\mathsf{C}(\mathrm{B}_2)$ takes the value of zero.

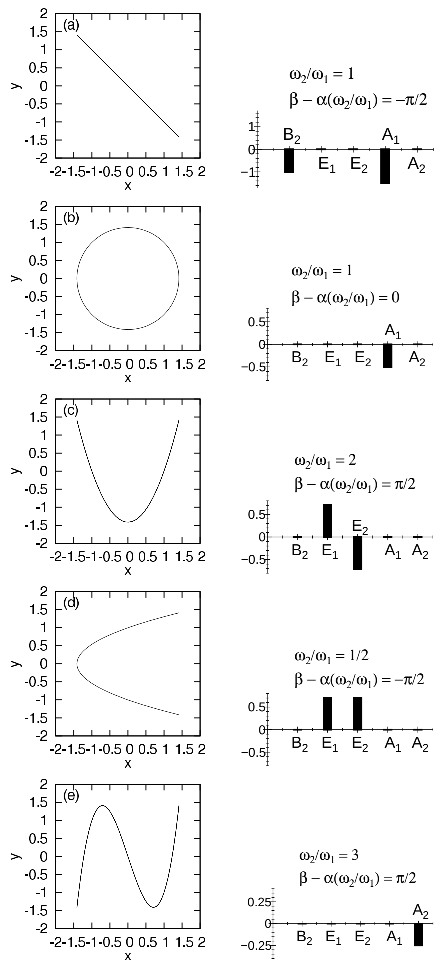

Figure 2. Lissajous' trajectories and their generalized covariance.

6. Conclusion

We have proposed a new theoretical framework for generalized covariance analysis, considering the limitations of the existing methods. To summarize, first, we showed that the pairwise cross-cumulants can be viewed as the nontrivial simplest correction terms to the Gaussian distribution.

Next, we have proposed a new method for extending the traditional covariance based on the $\mathrm{C}_{4\mathrm{v}}$ group. The key idea is to think of pairwise functional relationships as 2D geometric patterns. To the best of the author's knowledge, this is the first work to use the irreducible representations of a specific group as a tool for pattern recognition. We found that the traditional covariance exploits only one of the five irreducible representations, and we defined generalized covariances according to the other irreducible representations.

Finally, we have demonstrated the utility of the generalized covariances using an analytically solvable model. An application to an anomaly detection task [1] using a real-world time-series data set will be published elsewhere.

References

[1] T. Idé and K. Inoue. Knowledge discovery from heterogeneous dynamic systems using change-point correlations. In *Proc. of 2005 SIAM International Conference on Data Mining (SDM 05)*, pages 571–575, 2005.

[2] T. Inui, Y. Tanabe, and Y. Onodera. *Group Theory and Its Applications in Physics, 2nd ed.* Springer-Verlag Telos, 1996.

[3] R. Kondor and T. Jebara. A kernel between sets of vectors. In *Proc. the 20th International Conference on Machine Learning*, 2003.

[4] R. Kubo. Generalized cumulant expansion method. *Journal of the Physical Society of Japan*, 17(7):1100–1120, 1962.

[5] J. J. Sakurai. *Modern Quantum Mechanics*. Addison Wesley, 2nd Ed., 1994.

[6] A. Stuart and J. K. Ord. *Kendall's Advanced Theory of Statistics, Volume 1*. Arnold Publishers, 6th ed., London, 1998.