A note on the total covariation of multivariate random variables

Silia Vitoratou* and Ioannis Ntzoufras †

*Institute of Psychiatry - Biostatistics Department, King's College London, UK, e-mail: vasiliki.vitoratou@kcl.ac.uk.

†Department of Statistics, Athens University of Economics and Business, e-mail: ntzoufras@aueb.gr; corresponding author of this article.

Abstract

Dependencies between random variables are routinely assessed through their variance-covariance matrix which provides information in a pairwise manner. Here, we explore potential ways to summarize dependencies using a single measure, the total covariation index. This measure quantifies the total covariation and can be thought as the multivariate analogue of the covariance. The total covariation index can be implemented to assess dependencies among variables or identify departures from independence assumptions which are embodied in many popular statistical models. Due to its univariate nature and its simplicity, it can be also used to compute the covariation among products of random variables or sums of products of random variables. Moreover this index allows for the decomposition of the total covariation related to two sets of variables into between and within sets covariation.

Keywords: *Covariance decomposition, Covariance of products, Covariance of sums of products, Divergence from independence.*

1 Introduction

When two random variables are considered, their dependencies are assessed through their covariance (or correlation). In the case that more than two random variables are considered, then the variance-covariance matrix is implemented. The variance-covariance matrix consists of the corresponding pairwise covariances and therefore the dependencies between the N random variables is assessed in a pairwise manner. In this work, we focus on summarizing the dependencies between all variables under consideration using the *total covariation index* which can combine all relevant information into a single univariate measurement.

The assessment of the total covariation of N variables can be attractive in various ways. First, it quantifies the magnitude of the dependencies via a single value. Secondly, non-zero values of the index denote divergences from independence for a particular set of random variables, even though the reverse statement is not true (that is, zero values do not ensure independence). Third, it is feasible to explore the factors that influence the dependencies over all random variables which cannot be done in a straightforward manner using the variance-covariance matrix. An example is given by Vitoratou et al. (2014) where a total covariation index is used to assess possible departures from the conditional independence assumption in Bayesian latent variable models and identify which factors affect them. Finally, the total covariation index can be used to simplify the computation of covariances among products of random variables and sums of products of random variables.

The structure of the paper is as follows. Following Vitoratou et al. (2014), in Section 2 we define the total covariation. We study this index and we provide some further properties. In Section 3, we illustrate how it facilitates the computation of the covariance among products of random variables and/or sums of products of variables. In Section 4, we describe the decomposition of the total covariation related to two sets of variables, in terms of the indexes of each set. Finally, the paper closes with a short discussion and a comment on potential

applications of our findings in latent variable models.

2 Indexes of total covariation

The total covariance index (TCI) between N random variables $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$ was introduced by Vitoratou et al. (2014) and can be thought as the multivariate extension the covariance between two random variables. The TCI is formally defined as

$$TCI(\boldsymbol{Y}) = E\Big(\prod_{i=1}^{N} Y_i\Big) - \prod_{i=1}^{N} E(Y_i)\,. \tag{1}$$

From the above definition, it is clear that in the case of two random variables ($N = 2$) then TCI is equal to the covariance between Y_1 and Y_2. As opposed to the covariance matrix, the information provided by (1) is based on all N variables simultaneously, rather than being limited to two variables at each time. However, while (1) provides an index of the total covariation, it involves only expectations. Nevertheless, Eq. (1) requires the computation of an N-dimensional integral (the expectation of the product of the variables), which, in many occasions, can be analytically intractable. For all these reasons, it is reasonable to seek a more convenient computationally expression of TCI.

Dependencies between N random variables are often conceptualized via the joint cumulants (see Hu, 1991; Hasebe and Saigo, 2011 and references within), given by

$$\kappa(\boldsymbol{Y}) = \sum_{\pi \in \mathcal{P}_N} (|\pi| - 1)!(-1)^{(|\pi|-1)} \prod_{b \in \pi} E\Big(\prod_{i \in b} Y_i\Big), \tag{2}$$

where π runs through $\mathcal{P}_N$ which is the set of all possible partitions of $\{1, ..., N\}$, $|\pi|$ is he number of elements in the partition π, and b runs through the list of all blocks of the partition π. Based on (1) and (2), we define here the total covariation in terms of joint cumulants as

follows

$$TCI(\boldsymbol{Y}) = \sum_{\pi \in \mathcal{P}_N} \prod_{b \in \pi} \kappa(\boldsymbol{Y}_b) - \prod_{i=1}^{N} \kappa(Y_i),\ \boldsymbol{Y}_b = \{Y_i : i \in b\} \subseteq \boldsymbol{Y}. \tag{3}$$

The first term in (2) is the expectation of the product of the N variables and $\kappa(Y_i)$ is the first order cumulant which is equal to expectation $E(Y_i)$ of the random variable Y_i. Therefore, in addition to the expectations, (3) implements the covariances of each pair of variables as well as higher order cumulants. However, the number of all partitions π of a set $\{1, ..., N\}$ is given by the Bell number B_N (Rota, 1964) which increases exponentially with N (for example, when $N = 10$ the number of possible partitions is $B_{10} = 115975$), making the calculation of the total covariation using (3) less attractive for most applications.

An alternative expression to (1) and (3) can be obtained by the following equation which uses only first and second order joint cumulants (see Vitoratou et al., 2014)

$$TCI(\boldsymbol{Y}) = Cov_{(N)}(\boldsymbol{Y}) + \sum_{k=1}^{N-2} \left[\left(\prod_{i=N-k+1}^{N} E(Y_i) \right) Cov_{(N-k)}(\boldsymbol{Y}) \right]. \tag{4}$$

The term $Cov_{(k)}(\boldsymbol{Y}),\ k = 2, ..., N$ will be hereafter refereed to as the *item-product covariance*, defined as

$$Cov_{(k)}(\boldsymbol{Y}) = Cov\left(\prod_{i=1}^{k-1} Y_i, Y_k \right). \tag{5}$$

Note that for random variables centered to zero ($E(Y_i) = 0$), then $TCI(\boldsymbol{Y}) = Cov_{(k)}(\boldsymbol{Y})$. Using (4), TCI can be computed in a straightforward manner for arbitrary large N since it does neither involve the evaluation of the N-dimensional integral involved in the expectation of the product of the random variables as in (1), nor it requires to consider the large number of partitions as in (3). In addition, it demonstrates the effect of the bivariate covariances in the total covariation, along with their expectations.

Moreover, Vitoratou et al. (2014) implement the Cauchy-Schwartz inequality to derive an

upper bound for the absolute value of $TCI(\boldsymbol{Y})$; see Collorary 3.2 in Vitoratou et al. (2014). This upper bound is an increasing function of the expectations, the variances and the number of random variables under consideration which intuitively implies similar relationships with the TCI itself. Empirical evidence from simulated data confirmed these effects on the sample estimates of TCI.

In the following of this section, we proceed further and provide provide some additional results for the $TCI(\boldsymbol{Y}_N)$; the subscript N is used here to denote the number of random variables in $\boldsymbol{Y}$. Expression (4) implies that the total covariation among N random variables is assessed through a weighted sum of N-1 covariance terms. For each additional variable added, its expectation serves as a weight that adjusts the variable's contribution to the total covariation. The expression that links two successive terms in (4) is given by:

$$TCI(\boldsymbol{Y}_{k+1}) = Cov_{(k+1)}(\boldsymbol{Y}_{k+1}) + E(Y_{k+1})\ TCI(\boldsymbol{Y}_k),\ k = 2, ..., N-1, \tag{6}$$

since

$$\begin{aligned} TCI(\boldsymbol{Y}_{k+1}) - Cov_{(k+1)}(\boldsymbol{Y}_{k+1}) &= E\left(\prod_{i=1}^{k+1} Y_i\right) - \prod_{i=1}^{k+1} E(Y_i) - \left[E\left(\prod_{i=1}^{k+1} Y_i\right) - E(Y_k)E\left(\prod_{i=1}^{k} Y_i\right)\right] \\ &= E(Y_{k+1})\left[E\left(\prod_{i=1}^{k} Y_i\right) - \prod_{i=1}^{k} E(Y_i)\right] \\ &= E(Y_{k+1})\, TCI(\boldsymbol{Y}_k). \end{aligned}$$

From (6), we derive directly an upper bound for the absolute value of the index, given by $|TCI(\boldsymbol{Y}_{N+1})| \leq |Cov_{(N+1)}(\boldsymbol{Y}_{N+1})| + |E(Y_{N+1})\ TCI(\boldsymbol{Y}_N)|$. Therefore, if the additional variable is centered around its mean, the total covariation of the augmented set is bounded by the item-product covariance (with respect to the additional item), that is

$$-Cov\left(\prod_{i=1}^{N} Y_i, Y_{N+1}\right) \leq TCI(\boldsymbol{Y}_{N+1}) \leq Cov\left(\prod_{i=1}^{N} Y_i, Y_{N+1}\right) \ . \tag{7}$$

In the next sections, we use the $TCI(\boldsymbol{Y}_N)$ to derive the covariance in more complex settings involving products of different groups of random variables.

3 On the covariance between products and/or between sums of products of random variables

With respect to the covariance between products of variables, Brown and Alexander (1991) provide an expression based on the findings of Goodman (1960), Goodman (1962) and Bohrnstedt and Goldberger (1969). Their aim was to decompose the covariance and detect which variable(s) contribute most on the total variation. In order to achieve that, Brown and Alexander (1991) implement the sums over all possible k-tuplets ($k = 1, ..., N$) and r-tuplets ($k = 1, ..., M$) of variables, where N and M is the number of variables involved in each of the two products. This approach can be useful in relation to interaction terms where the dimensionality (N or M) is no larger than three. Nevertheless, when the objective is to summarize the information (rather than to decompose it in variance components) in cases with large number of random variables under consideration, $TCI(\boldsymbol{Y})$ can provide a reasonable alternative to Brown and Alexander (1991) formula.

TCI is employed here to facilitate the computation of the covariance of products and sums of products, concluding to expressions that can be efficiently evaluated no matter how large is the number of random variables involved in them. To begin with, let us consider two sets of variables, $\boldsymbol{X} = \{X_1, \ldots, X_M\}$ and $\boldsymbol{Y} = \{Y_1, \ldots, Y_N\}$. The covariance between the products of variables in each set can be computed according to the following lemma.

Lemma 3.1 *The covariance between two products of random variables,* $\prod_{i=1}^{N} Y_i$ *and* $\prod_{j=1}^{M} X_j$*, is given by*

$$Cov\left(\prod_{i=1}^{N} Y_i, \prod_{j=1}^{M} X_j\right) = TCI(\boldsymbol{Y} \cup \boldsymbol{X}) - E\left(\prod_{j=1}^{M} X_j\right) TCI(\boldsymbol{Y}) - \left[\prod_{i=1}^{N} E(Y_i)\right] TCI(\boldsymbol{X}), \quad (8)$$

where $TCI(\boldsymbol{Y} \cup \boldsymbol{X})$ *stands for the total covariation index of the full set. The proof is given in the Appendix.*

Expression (7) is simplified if the variables in one of the sets are mutually independent. In the special case where the variables within both sets are independent, then the covariance in (8) equals the total covariation of the $N + M$ variables, $TCI(\boldsymbol{Y} \cup \boldsymbol{X})$. If additionally all variables have zero means, then the covariance of the products equals the item-product covariance of the full set. Lemma 3.1 can be further implemented to derive the covariance of sums of products of random variables since

$$\begin{aligned} &Cov\left(\left[\sum_{k=1}^{S}\prod_{i=1}^{N_k} Y_{ki}\right], \left[\sum_{k'=1}^{S'}\prod_{j=1}^{M_{k'}} X_{k'j}\right]\right) = \sum_{k=1}^{S}\sum_{k'=1}^{S'} Cov\left(\left[\prod_{i=1}^{N_k} Y_{ki}\right], \left[\prod_{j=1}^{M_{k'}} X_{k'j}\right]\right) \\ &= \sum_{k=1}^{S}\sum_{k'=1}^{S'}\left\{TCI(\boldsymbol{Y}_{k+} \cup \boldsymbol{X}_{k'+}) - E\left(\prod_{j=1}^{M_{k'}} X_j\right) TCI(\boldsymbol{Y}_{k+}) - \left[\prod_{i=1}^{N_k} E(Y_i)\right] TCI(\boldsymbol{X}_{k'+})\right\}, \end{aligned}$$

where $\boldsymbol{Y}_{k+} = (Y_{k1}, Y_{k2}, \ldots, Y_{kN_k})^T$ and $\boldsymbol{X}_{k'+} = (X_{k'1}, X_{k'2}, \ldots, X_{k'M_k})^T$.

4 Decomposing the total covariation when two groups of random variables are under consideration

The results of Section 3 provide valuable intuition about the sources of covariation related to $\boldsymbol{X}$ and $\boldsymbol{Y}$. Solving (8) with respect to the covariation of the full set $TCI(\boldsymbol{Y} \cup \boldsymbol{X})$ leads

to

$$TCI(\boldsymbol{Y} \cup \boldsymbol{X}) = Cov\left(\prod_{i=1}^{N} Y_i, \prod_{j=1}^{M} X_j\right) + E\left(\prod_{j=1}^{M} X_j\right) TCI(\boldsymbol{Y}) + \left[\prod_{i=1}^{N} E(Y_i)\right] TCI(\boldsymbol{X}). \quad (9)$$

Therefore, the total covariation does not simply coincide with the sum of the TCIs of the two groups of variables $\boldsymbol{Y}$ and $\boldsymbol{X}$ but takes also under consideration any between groups dependencies.

If the N and M variables within each group/set are independent then the latter two terms in (9) become zero and the total covariation coincides with the covariance in the first term. That is, the covariance of the products of the variables at each set consists of the between sets covariation. Conversely, if there are no dependencies between the variables that belong to different sets, the covariance in the first term will be zero since "if the two groups are expectation-independent and covariance-independent the covariance of the products is zero" (Bohrnstedt and Goldberger, 1969). On the other hand, the within group total covariation is given by the corresponding index, $TCI(\boldsymbol{X})$ or $TCI(\boldsymbol{Y})$. The latter two terms in (9) involve these indexes, each weighted by the expectations referring to the other set. Based on these observations we decompose the total covariation into within and between groups covariation, according to the following definition.

Definition 4.1 *Let $\boldsymbol{Y} = \{Y_i\}_{i=1}^{N}$ and $\boldsymbol{X} = \{X_j\}_{j=1}^{M}$ denote two groups or sets of random variables. The total covariation of the full set $(\boldsymbol{Y} \cup \boldsymbol{X})$ can be decomposed into between and within sets covariation*

$$TCI(\boldsymbol{Y} \cup \boldsymbol{X}) = BCI(\boldsymbol{Y}, \boldsymbol{X}) + WCI(\boldsymbol{Y}, \boldsymbol{X}), \quad (10)$$

where the **between sets covariation** *is given by*

$$BCI(\boldsymbol{Y},\boldsymbol{X}) = E\left(\prod_{i=1}^{N} Y_i \prod_{j=1}^{M} X_j\right) - E\left(\prod_{i=1}^{N} Y_i\right) E\left(\prod_{j=1}^{M} X_j\right)$$
$$= Cov\left(\prod_{i=1}^{N} Y_i, \prod_{j=1}^{M} X_j\right) \tag{11}$$

and the **within sets covariation** *is given by*

$$WCI(\boldsymbol{Y},\boldsymbol{X}) = E\left(\prod_{i=1}^{N} Y_i\right) E\left(\prod_{j=1}^{M} X_j\right) - \prod_{i=1}^{N} E\left(Y_i\right) \prod_{j=1}^{M} E\left(X_j\right)$$
$$= E\left(\prod_{j=1}^{M} X_j\right) TCI(\boldsymbol{Y}) + \left[\prod_{i=1}^{N} E(Y_i)\right] TCI(\boldsymbol{X}). \tag{12}$$

In equations 10–12 we refer to the covariance attributed to different sets of random variables, in contrast to ANOVA where we measure the between and within variance of realizations/observations of the same random variable. Identifying groups or sets of variables based on their covariance is often the objective in multivariate analysis, as for instance in cluster and factor analysis models.

5 Discussion: future research

The total covariation among N random variables can be assessed in a straightforward manner via the index introduced by Vitoratou et al. (2014). In this article, we have demonstrated that TCI can be employed to compute the covariance of products and/or sums of products that involve arbitrary large number of variables. Additionally, the total covariation associated with groups (sets) of variables was studied here in relation to the TCI. Since the TCI measures the covariation of sets of variables, is directly linked to multivariate analysis techniques. It can be used in applications related to factor analysis (or latent variable models in general)

whose objective is to infer on the existence of unobserved (latent) variables based on the covariation of observed (manifest) items. Below we discuss specific problems encountered in the field where the use of the TCI seems promising.

To begin with, a fundamental assumption in latent variable models is the conditional independence assumption, that is, conditional on the latent variables the observed items are assumed conditionally independent. Non-zero values of the TCI represent *divergence from independence* for the variables involved, even though the reverse statement is false. Hence, it can be employed to test (reject) the independence assumption and more interestingly in the study of the factors that influence its violation. Such an example is given in Vitoratou et al. (2014) where the TCI is employed in order to study the impact of the dimensionality and the components' variability on Monte Carlo integration (MCI). Based on our results, similar studies may be conducted for the covariance of products and/or sums of products in different model settings.

A second area of potential application of the TCI, is the *item selection* problem often emerging in applications, as for instance in psychometrics. The item-total correlation (that is, the correlation of an item with the sum-score of the rest of the items) is extensively used by practitioners as a rule of thump to decide whether an item is inconsistent ($r<0.2$) or redundant ($r>0.7$) in the assessment of a latent trait. The TCI and the item-product covariance can play a central role in this procedure since the first is bounded by the latter when centered variables are added in the current set of variables under study; see equation (7). The exact mechanism that affects the total covariation for the augmented set is described by (6). Therefore, the TCI can be potentially used in order to construct a test for the contribution of an item to the total covariation and to decide upon its inclusion or exclusion.

Finally, the decomposition of the covariance in relation to sets of variables into between and within sets covariation (see Definition 4.1) can provide information towards the latent structure. As an appetizer, consider the case of two sets of items, each representing the items

that load on a specific latent variable. If the between sets covariation is substantially lower than the within sets covariation, then there is an indication that the items are not allocated properly. In fact, among all partitions of the items into two sets, the one which maximizes the within sets covariation (and therefore minimizes the between sets covariation) may be suggestive of the true latent structure.

While the above observations provide intuition and denote possible applications of the TCI, future research is required in order to construct appropriate hypothesis testing framework.

APPENDIX

Proof of Lemma 3.1

$$
\begin{aligned}
&Cov\left(\prod_{i=1}^{N} Y_i, \prod_{j=1}^{M} X_j\right) = E\left(\prod_{k=1}^{N+M} Y_k\right) - E\left(\prod_{i=1}^{N} Y_i\right) E\left(\prod_{j=1}^{M} X_j\right) \\
&= E\left(\prod_{k=1}^{N+M} Y_k\right) - \left[TCI(\boldsymbol{Y}) + \prod_{i=1}^{N} E(Y_i)\right]\left[TCI(\boldsymbol{X}) + \prod_{j=1}^{M} E(Y_i)\right] \\
&= TCI(\boldsymbol{Y} \cup \boldsymbol{X}) - TCI(\boldsymbol{Y})TCI(\boldsymbol{X}) - \left[\prod_{i=1}^{N} E(Y_i)\right] TCI(\boldsymbol{X}) - \left[\prod_{j=1}^{M} E(X_j)\right] TCI(\boldsymbol{Y})
\end{aligned}
$$

$$
= TCI(\boldsymbol{Y} \cup \boldsymbol{X}) - E\left(\prod_{j=1}^{M} X_j\right) TCI(\boldsymbol{Y}) - \left[\prod_{i=1}^{N} E(Y_i)\right] TCI(\boldsymbol{X}),
$$

since $\quad TCI(\boldsymbol{Y})\, TCI(\boldsymbol{X}) = E\left(\prod_{j=1}^{M} X_j\right) TCI(\boldsymbol{Y}) - \left[\prod_{j=1}^{M} E(X_j)\right] TCI(\boldsymbol{Y}).$ $\square$

References

Bohrnstedt, G. W. and Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64(328):1439–1442.

Brown, D. and Alexander, N. (1991). The analysis of the variance and covariance of products. *Biometrics*, 47(2):429–444.

Goodman, L. A. (1960). On the exact variance of products. *Journal of the American Statistical Association*, 55(292):708–713.

Goodman, L. A. (1962). The variance of the product of K random variables. *Journal of the American Statistical Association*, 57(297):54–60.

Hasebe, T. and Saigo, H. (2011). Joint cumulants for natural independence. *Electronic Communications in Probability*, 16(44):491–506.

Hu, S.-L. J. (1991). Probabilistic independence and joint cumulants. *Journal of Engineering Mechanics*, 117(3):640–652.

Rota, G.-C. (1964). The number of partitions of a set. *The American Mathematical Monthly*, 71(5):498–504.

Vitoratou, S., Ntzoufras, I., and Moustaki, I. (2014). Explaining the behavior of joint and marginal Monte Carlo estimators in latent variable models with independence assumptions. *Statistics and Computing. In press.*