

On Testing for the Absence of Monotone Dependence in a Random Vector

Sulla Verifica dell'Ipotesi di Indifferenza per un Vettore Aleatorio

Lilla Di Scala¹

Dipartimento di Matematica, Università di Pavia, e-mail: lilla@dimat.unipv.it

Riassunto: La misura del grado di associazione fra variabili quantitative e la verifica inferenziale di tale associazione sono da sempre argomenti di indubbio interesse statistico. In questo lavoro si propone un nuovo test non-parametrico per l'ipotesi di assenza di associazione, nel senso specifico dell'assenza di dipendenza monotona (o *indifferenza*) per un vettore aleatorio bidimensionale. Tale test adotta come statistica l'indice di omofilia proposto da Cifarelli e Regazzini (1990) e viene condotto sfruttando nuovi risultati asintotici ottenuti per la versione empirica di tale indice. A titolo di esempio, si verifica la co-regolazione fra due profili genici in un contesto di microarray.

Keywords: indifference, Frechet class, Hadamard differentiability, d -delete jackknife.

1. An Index of Homophily

The term *monotone dependence* indicates the tendency of two random variables to increase/decrease concomitantly or contrariwise. The former is usually referred to as *concordance* (or positive dependence); the latter as *discordance* (or negative dependence). The concept of an *index of homophily* to provide a one-number summary of monotone dependence was first introduced by Gini, see for example Gini (1916). The idea was to measure dependence relative to the greatest attainable intensity consistent with the observed marginals. Therefore, given X and Y with marginal cdf's F and G respectively, the corresponding bivariate distribution H is considered in the Frechet class $\Gamma(F, G)$ and the index of homophily will thus be, generally, a functional $\mathcal{I} : \Gamma(F, G) \rightarrow \mathbb{R}$. Let H^+ and H^- be the upper and lower bounds, respectively, of $\Gamma(F, G)$ with regards to the Positive Quadrant Dependence (PQD) ordering, i.e. $H^-(x, y) \leq H(x, y) \leq H^+(x, y)$, $\forall (x, y) \in \mathbb{R}^2$, $\forall H \in \Gamma(F, G)$. The cdf's H^+ and H^- represent, respectively, maximal and minimal concordance. In fact, within the bivariate Frechet class, it can be easily checked that the above inequalities also hold for the corresponding survival functions: $S^-(x, y) \leq S(x, y) \leq S^+(x, y)$, where $S(x, y) = 1 - F(x) - G(y) + H(x, y)$. Moreover, Kimeldorf and Sampson (1978) proved that, in the case of continuous marginals, if $H = H^+(H^-)$ then there exists g monotone increasing (decreasing) such that $Y \stackrel{a.s.}{=} g(X)$. Cifarelli and Regazzini (1990) showed that, even removing the hypothesis of continuity, H^+ and H^- still correspond to the concept of maximal (or complete) monotone dependence, according to a definition which generalizes that of Kimeldorf and Sampson (1978).

Various measures of monotone dependence have been proposed in the literature. The *monotone correlation* index ρ^* of Kimeldorf and Sampson (1978), for which the authors prove good theoretical properties, is then generally difficult to compute. The index σ of Schweizer and Wolff (1981), which is the normalized L_1 -distance of H from indepen-

¹Indirizzo per corrispondenza: Via Ferrata 1, 27100 Pavia

dence $H^\perp = FG$, actually measures dependence without discriminating between concordance and discordance. Among these measures there also are known *cograduation* indexes, such as Kendall's τ , Spearman's ρ and Gini's, which basically consider the variables as ordinal rather than quantitative. Gini's index, in particular, adopts the concept of dependence in terms of computing the "distance" of H from maximal (H^+) and minimal (H^-) dependence. It is with this construction in mind that Cifarelli and Regazzini (1990) proposed a general measure of monotone dependence $\Theta_a(H) : \Gamma(F, G) \rightarrow [-1, 1]$, thus providing a formal framework to the Ginian idea of homophily. Its general version is thus defined as the (normalized) expected value

$$\Theta_a(H) = \frac{1}{K_a} \left\{ \mathbb{E}_H[\gamma_a] - \mathbb{E}_H[\delta_a] \right\}$$

where $a : \mathbb{R} \rightarrow [0, +\infty]$ is continuous, increasing and convex, $a(x) = 0$ for $x \leq 0$ and

$$\begin{aligned} \gamma_a(P = (P_x, P_y)) &= \inf_{Q \in S_{H^-, P}^v} a(d(P, Q)) + \inf_{Q \in S_{H^-, P}^h} a(d(P, Q)) \\ \delta_a(P = (P_x, P_y)) &= \inf_{Q \in S_{H^+, P}^v} a(d(P, Q)) + \inf_{Q \in S_{H^+, P}^h} a(d(P, Q)) \end{aligned}$$

being d the Euclidean distance on $\mathbb{R}^2$, $S_{H, P}^v := (\{P_x\} \times \mathbb{R}) \cap S_H$ the vertical section in $P_x \in S_F$ of the support S_H of H and similarly $S_{H, P}^h := (\mathbb{R} \times \{P_y\}) \cap S_H$ the horizontal section in $P_y \in S_G$, while K_a is the normalizing constant (function of F and G). The index acquires a simpler form both in the case of continuous marginals, which is suitable for theoretical considerations on the population index, and also in the empirical case, where it is easily computable even in the presence of ties; see Di Scala (2004) for details.

Cifarelli and Regazzini (1990) have shown that the index is well defined for every $H \in \Gamma(F, G)$ such that $\mathbb{E}_{H^-}[\delta_a] < \infty$ and $\mathbb{E}_{H^+}[\gamma_a] < \infty$ and that it is consistent with the PQD ordering on $\Gamma(F, G)$, thus completing such partial ordering. If $\Theta_a(H) = 0$, X and Y are said to be *mutually indifferent*. Under the hypothesis of continuity, indifference holds for independent variables but also whenever the average departure from H^+ compensates that from H^- , for example when $H = (H^+ + H^-)/2$. Lastly, the index is such that $\Theta_a(H^+) = 1$ and $\Theta_a(H^-) = -1$, hence discriminating between positive and negative dependence.

2. Testing for Indifference

The main issue of this short paper is to propose a new inferential procedure to test for indifference. This kind of tests is neglected in the statistical literature, although ample space has been dedicated to tests for independence which make use of concordance/discordance measures, such as Gini's cograduation index. Since these tests have no power to distinguish between independence and indifference, they should be properly considered as tests for indifference, see Conti (1994). The test proposed here explicitly acknowledges indifference as the null hypothesis, to which the significance level α is associated.

Cifarelli, Conti and Regazzini (1996) carry out a test of the hypothesis of indifference by making use of asymptotic results obtained for a class of cograduation indexes which includes Gini's. The asymptotic distribution of the test statistic is obtained via the theory of empirical process convergence, as the continuity of the marginal cdf's and the specific functional form of the index allow the passage to uniform marginals and thus working

with copulas. Conti (1994) deals with the limit distribution of a non-normalized general measure of monotone dependence through the theory of U -statistics, also furnishing a consistent estimator of its asymptotic variance.

In order to carry out the proposed test, the asymptotic distribution of the empirical homophily index $\Theta_{a,n}$ is determined under the null hypothesis of indifference. The strategy adopted is that of showing that the non-normalized index $\Psi_a(H) = \Theta_a(H) \cdot K_a$ is a Hadamard differentiable functional, hence applying the functional delta method to obtain asymptotic normality, see Di Scala (2004). In fact, the following theorem holds, where $\Delta(S_F \times S_G)$ is the set of bi-dimensional cdf's with support on $S_F \times S_G$ and $C(S_F \times S_G)$ is the set of all continuous functions on $S_F \times S_G$.

Theorem. *Let F and G be continuous, strictly increasing and with bounded support. Furthermore, let F and G be continuously differentiable on their support with strictly positive derivatives. Assume that $a(x) = \int_0^x b(t)dt$, $\forall x \in \mathbb{R}$, where b is continuous on $\mathbb{R}$. Then the functional $\Psi_a : \Delta(S_F \times S_G) \rightarrow \mathbb{R}$ is Hadamard differentiable in $H \in \Gamma(F, G)$ tangentially to $C(S_F \times S_G)$. Let $\Psi'_{a,H}$ be the above derivative; then*

$$\sqrt{n}(\Psi_{a,n} - \Psi(H)) \xrightarrow{\mathcal{L}} \Psi'_{a,H}(\beta_H) \sim \mathbb{N}(0, \sigma_H^2),$$

as $n \rightarrow \infty$, where β_H is an appropriate Brownian sheet, i.e. a centered Gaussian process with a.s. continuous sample paths and covariance function

$$\text{Cov}[\beta_H(x_1, y_1), \beta_H(x_2, y_2)] = H(x_1 \wedge x_2, y_1 \wedge y_2) - H(x_1, y_1)H(x_2, y_2).$$

(Note that the theorem's hypothesis on a holds, for example, if $a(x) = x^\alpha \mathbb{I}_{\{x \geq 0\}}$ with $\alpha > 1$; while it does not hold for the simple homophily index, although simulation studies did not show any difference between $\alpha > 1$ and $\alpha = 1$.) As a consequence of this result, since the empirical version $K_{a,n}$ of the normalizing constant is a consistent estimator of K_a , see Di Scala (2004), under the null hypothesis $H_0 : \Theta_a(H) = 0$, it holds that

$$\sqrt{n}\Theta_{a,n} \xrightarrow{\mathcal{L}} \mathbb{N}(0, \sigma_H^2/K_a^2), \quad \text{for } n \rightarrow \infty.$$

However, the asymptotic variance is a function of H and its form is quite complicated, so that determining a consistent estimator for it is not immediate. Therefore, in order to construct a rejection region of asymptotic level α for the test, the approach adopted will be that of Politis and Romano (1994). These authors make use of a subsampling technique, namely the d -delete jackknife, to construct large sample confidence regions for a parameter of interest, say $\Theta_a(H)$. Their result requires the minimal assumption that the functional $\sqrt{n}\Theta_{a,n}$ weakly converges to a limit law as $n \rightarrow \infty$, which is satisfied under H_0 as per the above theorem. This assumption guarantees that the sampling distribution of $\sqrt{n}\Theta_{a,n}$ may be well approximated by the distribution of the same statistic computed over the $\binom{n}{m}$ subsets of size m of the n i.i.d. observations (X_i, Y_i) , where m is such that $m/n \rightarrow 0$ and $m \rightarrow \infty$ as $n \rightarrow \infty$. As computing $\binom{n}{m}$ times the test statistic may be burdensome, Politis and Romano (1994) note that the same conclusions hold if one chooses randomly s subsets out of the $\binom{n}{m}$ (with or without replacement) such that $s \rightarrow \infty$.

3. Co-regulation of Expressions

After completion of the human genome map, the goal of the current genomic phase, termed *functional genomics*, is to annotate this map by understanding each gene's function and their interplay. Microarray technology has been developed for this aim and is

able to monitor the protein-coding activity (*expression*) of thousands of genes at a time. In particular, microarrays are being used as a prognostic assessment tool to predict survival in cancer patients. Beer *et al.* (2002) identified a list of 100 genes which were found to be related to survival in patients affected by lung adenocarcinoma. It is of interest to determine if these genes are co-regulated, i.e. if they are up/down-regulated concomitantly, thus implying that they may be functionally related. In order to verify this, data from Bhattacharjee *et al.* (2001) are used, where expression profiles of cancer patients were monitored on a partially overlapping set of genes as that of Beer *et al.* (2002). Genes *CASP4* and *FUT3* are picked as an example. Choose $a(x) = x^2$ and let $\alpha = 0.05$. The quadratic homophily index is $\Theta_{x^2, n=203} = -0.10$. Letting $m = n^{2/3}$, see Politis and Romano (1994) for the choice of m , and implementing a Monte Carlo algorithm with n^2 iterations, the test confirms that there is not enough evidence to reject the hypothesis $H_0 : \Theta_a(H) = 0$ against the alternative $H_0 : \Theta_a(H) \neq 0$; the approximated p -value is 0.09.

Acknowledgements. The present work is part of the author's Ph.D. thesis. Sincere thanks go to the supervisor Professor E. Regazzini, University of Pavia, for the constant advice.

References

- Beer D.G., Kardia S.L.R., Huang C., Giordano T.J., Levin A.M., Misek D.E., Lin L., Chen G., Gharib T.G., Thomas D.G., Lizyness M.L., Kuick R., Hayasaka S., Taylor J.M.G., Iannettoni M.D., Orringer M.B. and Hanash S. (2002) Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature*, 8, 816–824.
- Bhattacharjee A., Richards W.G., Staunton J., Li C., Monti S., Vasa P., Ladd C., Beheshti J., Bueno R., Gillette M., Loda M., Weber G., Mark E.J., Lander E.S., Wong W., Johnson B.E., Golub T.R., Sugarbaker D.J. and Meyerson M. (2001) Classification of human lung carcinoma by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proceedings of the National Academy of Science*, 98, 13790–13795.
- Cifarelli D.M., Conti P.L. and Regazzini E. (1996) On the asymptotic distribution of a general measure of monotone dependence, *Annals of Statistics*, 24, 1386–1399.
- Cifarelli D.M. and Regazzini E. (1990) On monotone dependence, Technical Report 17, IAMI.
- Conti P.L. (1994) Asymptotic inference for a general measure of monotone dependence, *Journal of the Italian Statistical Society*, 2, 213–241.
- Di Scala L. (2004) *Analysis of a class of monotone dependence measures with an application to microarray data*, Ph.D. thesis, Dottorato in Statistica Matematica, Università degli Studi di Pavia.
- Gini C. (1916) Indici di concordanza, in: *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*, 1419–1461.
- Kimeldorf G. and Sampson A.R. (1978) Monotone dependence, *Annals of Statistics*, 6, 895–903.
- Politis D.N. and Romano J.P. (1994) Large sample confidence regions based on subsamples under minimal assumptions, *Annals of Statistics*, 22, 2031–2050.
- Schweizer B. and Wolff E.F. (1981) On nonparametric measures of dependence for random variables, *Annals of Statistics*, 9, 879–885.