Clustering-based measurement of dependence

Raffaella Piccarreta, Marco Bonetti, Sergio Venturini
Istituto di Metodi Quantitativi, Università Bocconi
Viale Isonzo 25, 20137 Milano, Italy
{raffaella.piccarreta, marco.bonetti, sergio.venturini}@unibocconi.it

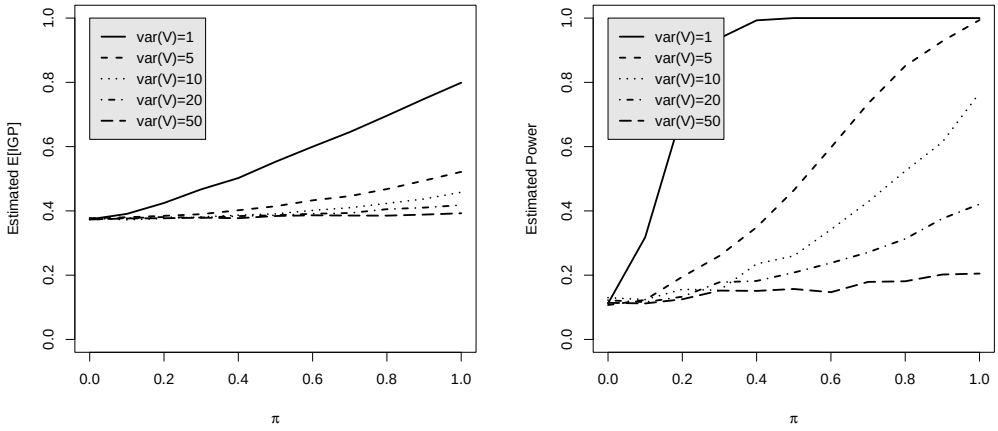
Abstract: A measure of the dependence of a multivariate response variable upon a categorical variable is introduced. Its characteristics are explored via simulations by referring to a specific mixture association model. Inferential aspects are investigated using a permutation test approach. We present preliminary results. We propose an extension to the case of several categorical explanatory variables.

Keywords: Association, IGP, Permutation tests

1 Introduction

Kapp and Tibshirani (2007) introduce the IGP (In Group Proportion) measure within the context of validating clusters. Let P_T be a partition of N observations on a multivariate variable Y into K clusters. Here, T denotes the (training) data set. Suppose that new observations on Y are available in a second dataset D. It is of interest to assign them to one of the previously determined clusters; the IGP has been introduced to evaluate the adequacy of the chosen assignment procedure. (The classification procedure may be defined in different ways).
Let $C_T(i)$ indicate the cluster to which the i-th observation in D is assigned, with $i = 1,\ldots,N_D$ and N_D indicating the size of D. Denote by P_D the resulting partition of D.
The (overall) IGP is defined as the proportion of cases in D that are classified to the same group as their nearest neighbor. More precisely, let $NN(i) \in D$ indicate the nearest neighbor of the i–th observation in D, and let $C_T(NN(i))$ denote the cluster to which $NN(i)$ is assigned. The IGP is therefore $IGP(P_D) = \frac{1}{N_D}\sum_{i=1}^{N_D} 1[C_T(NN(i)) = C_T(i)]$.
We propose to use the IGP index to measure the extent of the association between one set of response variables, Y, and one explanatory variable X_1, both observed on a *single* dataset of size N, yielding for each observation the values $(Y_i, X_{1,i})$. We start by considering the case of one categorical nominal variable X_1. Note that the (K_1) levels of X_1 define a partition of the N observations. Let $W_{1,i} = 1(X_{1,i} = X_{1,NN(i)})$ be the indicator of the event "observation i and its nearest neighbor $NN(i)$ share the same value of X_1." Then the IGP index in this context is defined as $IGP_Y(X_1) = \frac{1}{N}\sum_{i=1}^{N} W_{1,i}$. If the responses Y are related to X_1, then X_1 should provide a good partition also with respect to Y, thus yielding a high value of $IGP_Y(X_1)$.
Note that this can be viewed as a sort of nonparametric ANOVA problem on possibly multivariate responses. Below we explore the main features of this IGP-based approach through simulations based on a specific multivariate association model, with particular attention to the inferential aspects of the approach. We then propose an extension of this IGP measure to the case of several categorical explanatory variables $X_1,\ldots,X_K$.

Figure 1: *Estimated $E[IGP_Y(X_1)]$ (left) and power (right).*



2 IGP: A small simulation study and inferential aspects

Consider the association model between X_1 and Y such that Y is a mixture of two distributions f_V and f_Z with mixing parameter $\pi \in [0,1]$. Also, V and Z are distributed as two mixtures, each of three components (f_1^V, f_2^V, f_3^V) and (f_1^Z, f_2^Z, f_3^Z) respectively, with mixing vectors $(\alpha_1^V, \alpha_2^V, \alpha_3^V)$ and $(\alpha_1^Z, \alpha_2^Z, \alpha_3^Z)$, and all bivariate normals with different means and with variance-covariance matrices equal to $\sigma_V^2 I_2$ and to $\sigma_Z^2 I_2$. V and Z are taken to be independent. The categorical explanatory variable X_1 is defined as the mixture component from which V is generated. Notice that this induces three groups whose within dispersion is related to the standard deviation σ_V. Thus, Y is related to X_1 if the association parameter π assumes values close to 1. If π assumes low values, Y does not depend upon X_1 (through V) but, rather, upon Z. In particular, the value $\pi = 0$ in this model corresponds to the null hypothesis of *no association*.
We conducted a small simulation study to explore the relationship between π and the IPG. For fixed values of σ_V^2, σ_Z^2 we repeatedly generated samples of size N from the model above, and estimated the expected value of the IGP measure over the simulated samples. We used 1000 simulated datasets of size 100 for each value of π. As an illustration, the left panel in Figure 1 shows the monotonicity that was observed across the simulations (results refer to the case $\sigma_Z^2 = 5$; similar patterns were observed for different values). This behavior suggests that *IGP* may be considered a reasonable measure of dependence.
However, a confounding effect exists in general between association (as measured by π) and the strength of the structure in Y. For example, if $\pi = 1$ but Y has weak structure (equivalently, if Y coincides with V but the variance σ_V^2 is very large) then the groups induced by V will not retain information on the dispersion of Y. This situation will practically coincide with the case of no association, even though $\pi = 1$. This behavior appears to be a general characteristic of this problem, and should be kept in mind when interpreting the index. In other words, the ability of the IGP to measure the level of association depends upon the fact that there is some structure in Y to begin with. If Y has no structure, so that the X_1-groups can essentially be viewed as a random selection from the observations' labels, then *any* measure of association will be useless. Consistently with this observation, when the Y structure is highly dispersed the IGP index does not reach its theoretical maximum value of one and, moreover, it shows a low sensitivity to the strength

of association (i.e., it increases very slowly as π increases). Under the null hypothesis of no association ($\pi = 0$ in our model) it can be shown that $E[IGP_Y(X_1)] = \sum_{k=1}^{K_1} p_k^2$, where p_k is the probability that X_1 takes its $k-th$ level. This value can be computed exactly for the simulated model above from the theoretical parameters α_k^Y. For example, for the parameter values that were used one finds that $E[IGP_Y(X_1)] = .375$ under H_0. (This null value can be noted in the left panel of Figure 1). On actual data, the quantity $E[IGP_Y(X_1)]$ under H_0 can be estimated from the observed counts in the K groups induced by X_1.
To test H_0 one can use a permutation distribution approach, i.e. extract random permutations from the set of the N X-group labels associated to the Y-observations. For each permutation of the labels the IGP is computed, and the p-value for $IGP_Y(X_1)$ is obtained as the proportion of IGP values that are more extreme (larger) than the observed $IGP_Y(X_1)$. A small p-value (less than a fixed level α) indicates rejection of H_0 in favor of the alternative hypothesis of association. To evaluate the power of this procedure one can simulate many datasets, and for each determine whether the permutation test would reject H_0 at a chosen alpha level. Thus one can easily estimate the power of the test to reject H_0 for different values of σ_V^2 and σ_Z^2, for various alternative values of π. Note that the rejection probability that one obtains with this procedure is averaged over all the possible group label counts that could be observed when distributing N observations over K groups. In other words, in our model the average is taken over a multinomial distribution having parameters $(N, (\alpha_V^1, \alpha_V^2, \alpha_V^3))$. In Figure 1 (right panel) the estimated powers of permutation tests are reported for the case when $\alpha = 0.1$ for various combinations of values of π and σ_V^2 (results refer to the case $\sigma_Z^2 = 5$; similar patterns were observed for different values). It is worth noting that the power appears to be increasing with π but its maximum value depends upon the dispersion within Y. This phenomenon is consistent with the discussion above on the confounding effect of π and the variance of Y.

3 Multiple IGP indices

Now, consider the case of K variables X_j, $j = 1, \ldots, K$ measured on the N individuals, so that we have the N covariate vectors $(X_{1,i}, \ldots X_{K,i})$, $i = 1, \ldots N$. Call $W_{j,i} = 1(X_{j,i} = X_{j,NN(i)})$ the indicator of the event "i and its nearest neighbor $NN(i)$ share the same value of variable X_j," and $\overline{W}_{j,i} = 1(X_{j,i} \neq X_{j,NN(i)}) = 1 - W_{j,i}$. Consider a subset of the K covariates: without loss of generality, to simplify notation let these covariates be the first h of the K covariates, i.e. the covariate vector $(X_1, \ldots, X_h)$, $1 \leq h \leq K$. We define the IGP measure for $(X_1, \ldots, X_h)$ as the quantity

$$IGP_Y(X_1, \ldots X_h) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{h} \left\{ \sum_{\rho \in C_{j,h}} \left[\left(\prod_{u=1}^{h} W_{u,i}^{\rho_u} \overline{W}_{u,i}^{1-\rho_u} \right) \pi_\rho \right] \right\},$$

where $\rho = (\rho_1, \ldots, \rho_h)^T$ is a vector of zeros and ones and $C_{j,h}$ is the set of all vectors that have *at most* j ones and $(h - j)$ zeros.
The problem of properly extending a measure of association to the multiple case is in general non-trivial, and in particular in our case in which the "dependent" variable $W_{i,j}$ is a function of *all* observations. This is a consequence of the use of the nearest neighbor approach. Our proposal above is very general, and allows for a possible definition of a partial index of the subset of variables $X_{h_1+1}, \ldots, X_h$ given that $X_1, \ldots, X_{h_1}$ ($h_1 < h$) already entered the model as $IGP(X_{h_1+1}, \ldots, X_h | X_1, \ldots, X_{h_1}) = IGP(X_1, \ldots, X_h) - IGP(X_1, \ldots, X_{h_1})$. Here, too, inference can be based on permutation distributions techniques.

Note that the weighting function π_ρ assigns different weights to the various intersections of the events indicated by $W_{j,i}$ and $\overline{W}_{j,i}$. By selecting different functions π_ρ one can construct IGP-type indices that differently measure the degree of similarity in the covariate space between each observation and its nearest neighbor. In particular, the following three special cases can be identified: (i) The *Intersection* model: $\pi_\rho = 1$ for $\rho = (1,1,1)$, zero otherwise. This IGP index measures the association between Y and the variable obtained by combining all categories of the h explanatory variables; (ii) The *Union* model: $\pi_\rho = 1$ for all $\rho \in C_{j,h}$; (iii) The *Additive* model: $\pi_\rho = \frac{1}{h}\left(\sum_{u=1}^{h} \rho_u\right)$, or the proportion of the h variables such that $X_{j,i} = X_{j,NN(i)}$.
The intersection model is conservative: it considers i and $NN(i)$ to be similar in the covariate space only if they share the value of *all* covariates $X_1, \ldots, X_h$. At the other end of the spectrum, the union model considers i and $NN(i)$ to be similar when they take the same value of *at least one* of the variables. The similarity measure used in the additive model is proportional to the simple matching coefficient. It can be easily shown that in this case $IGP_Y(X_1, \ldots X_h)$ is the average of the marginal IGP indices $IGP_Y(X_1), \ldots, IGP_Y(X_h)$. A consequence of this is that a natural (descriptive) model selection procedure would introduce variables with decreasing marginal IGP (as this guarantees the slowest possible decrease in the joint IGP) and therefore this choice of the function π_ρ does not use the joint information of the variables in the model building process – which is not satisfactory. Regardless of the choice of the weighting function, in the case of only one variable the general definition given above reduces to a quantity that is proportional to the original definition of IGP as given in Kapp and Tibshirani (2007).
Also, note that all the three cases above have specialized the weighting function π_ρ to be a function of the number of variables taking the same value for i and its nearest neighbor: i.e., the weighting functions are all of the form $\pi_\rho = \pi(\rho) = \pi\left(\sum_{u=1}^{h} \rho_u\right)$. A possibility that compromises between the ones seen above is the choice $\pi_\rho = \left[\frac{1}{h}\left(\sum_{u=1}^{h} \rho_u\right)\right]^2$, which downweights the observations that have few covariate values in common with their nearest neighbor. A limited simulation study suggests that this choice performs well in selecting the relevant covariates and in reflecting the degree of dependence between Y and X.

4 Conclusions

The use of the IGP as a measure of association seems promising. As any other measure of X/Y–association, the IGP reflects both dependency and the amount of "explainable" structure in Y, and hence rejection of the null hypothesis strongly suggests the existence of association. This approach only requires the Y-distances (or dissimilarities) between all possible pairs of cases, and the procedure can be applied whatever the dissimilarity measure: for example, one can consider time series (one for each case), sequence data (e.g. categorical time series or genetic sequences), and other situations where Y is complex but a dissimilarity measure between two cases can be defined.

References

Kapp A. and Tibshirani R. (2007) Are clusters found in one dataset present in another dataset?, *Biostatistics*, 8, 9–31.