A multivariate analysis of tourists' spending behaviour

Marta Disegna[1], Fabrizio Durante[2], and Enrico Foscolo[2]

Abstract According to the micro-economic theories regarding consumption behaviour, the determinants affecting the joint propensity of purchasing different goods and services are investigated. For this purpose, a copula-based model is suggested. This is in fact a suitable model to understand whether and how different expenditure categories are dependent with each other. A real application drawn from the tourism field is used to illustrate how the proposed approach works and to demonstrate its flexibility. The findings could guide local practitioners and managers in creating new promotional campaigns able to attract visitors willing to pay on a bundle of goods and services correlated with each other.

1 Introduction

The economic impact of tourism flows is often essential for those regions/local communities in which tourism is considered the major source of income [5]. In order to improve the economic effects of tourism visits, appropriate data and tools are needed to study the determinants of tourism expenditure and to analyse the tourists' spending behaviour in depth. In fact, as stated by [1], the use of micro-level makes it possible to observe individual choices regarding the consumption of a tourism commodity or service, and to analyse the heterogeneity and diversity that characterize individual tourism consumption behaviour. In other words, adopting a micro-level approach enables us to take both the consumer behaviour theory on the decision-making process to purchase, and the neoclassical economic theory of budget constraint, into consideration. In particular, we assume that the individual pur-

Faculty of Management, Bournemouth University, `disegnam@bournemmouth.ac.uk` · Faculty of Economics and Management, Free University of Bozen-Bolzano, `fabrizio.durante@unibz.it, enrico.foscolo@unibz.it`

chase process for a tourism good or service is a two-decision process [6], i.e. the decision to purchase a good followed by the decision on how much to spend on it. The economic theory of budget constraint is based on the assumption of weak separability between goods and services that leads tourists to allocate their budgets in accordance with a three-stage tourist spending process [8]: firstly, tourists decide how much of their budget to allocate for travel; secondly, they decide where to go on vacation; thirdly, they choose how to allocate their tourist budget among various goods and services offered by the selected destination. Obviously, the above-mentioned two economic theories are not disjointed but overlap; this means that an individual has to make a two-stage decision process in each stage comprised in the three-stage tourist spending process. This study contributes to this micro-economic tourism literature by analysing the first stage of the decision-making process (i.e. the propensity of tourist purchase) and the third stage of the budget allocation process (i.e. the allocation of tourist budget among various goods and services offered by a destination) simultaneously. In particular, this paper aims to analyse the factors involved in the decision to consume different categories of tourism goods and services simultaneously. To this end, we exploit the advantages of the copula-based models. Firstly, univariate Logit regressions are estimated per each category by considering a set of possible determinants. Then these regressions are grouped together by means of a copula, which is a multivariate distribution function that aims at describing the dependence among random outcomes in a flexible way [3]. The obtained model allows us to understand whether and how the different purchase decisions are correlated with each other. The methodology is illustrated by analysing a sample of international visitors to the South Tyrol region (Northern Italy).

2 The dataset

The dataset used in this study is drawn from the "International Tourism in Italy" annual survey, conducted by Bank of Italy in order to determine the tourism balance of payments. The survey offers detailed information on the amount of money (in Euro) spent in the five main categories of a typical travel budget: 1) Accommodation (Y_1), which also includes expenditure on food and beverages within accommodation premises; 2) Food & beverages (Y_2) consumed outside accommodation premises; 3) Shopping (Y_3), including souvenirs, gifts, clothes, etc. purchased only for personal use; 4) Internal transportation (Y_4) within the visited destination, including purchase of fuel; 5) Other services (Y_5), like museums, shows, entertainment, etc. In this study we focus on a subsample of 550 international visitors who spent time in the South Tyrol region (Northern Italy) in 2011 and whose main purpose for the trip was "tourism, holiday, leisure". South Tyrol is a highly competitive

destination compared to the rest of the Italian regions in the field of tourism, which is one of the leading sectors of the regional economy [2].

The sample consists of 87% tourists (i.e. people who stay at least one night in South Tyrol), and 13% day-visitors. Most of the sample stated they had incurred costs for tickets and/or transportation fuel (77%), souvenirs, gifts, items of clothing, or other things for personal use (69%), and food and beverages (87%) during the trip to South Tyrol. By contrast, only 34% of the sample stated they had incurred costs for other services, like museums, shows, entertainments, guided excursions, rented vehicles, or language courses. Table 1 describes the set of explanatory variables ($\mathbf{x}$) considered in this study. On average, visitors seem to be satisfied with the destination, giving a score value of 8.3 out of 10, they mainly come from Germany (62%) and Austria (15%), the two traditional markets for South Tyrol tourism, they are employed (53%), and more than half (60%) of the sample is aged over 44 years (see Table 1).

3 The methodology

Let Y_j be a dichotomous variable describing the decision to spend ($Y_j = 1$) or not ($Y_j = 0$) in the j-th tourism expenditure category, such as accommodation, transportation, and shopping. This study aims at modelling the dependence among these variables in order to understand whether the decision to spend or not spend in one category is correlated with the decision to spend in other categories, given a set of explanatory variables $\mathbf{x}$. Our main interest is to estimate the probability of spending in some (or all) categories given the set of covariates related to the tourist, namely

$$\mathbb{P}\left(Y_1 = y_1, \ldots, Y_d = y_d \mid \mathbf{x}\right)$$

for $y_j \in \{0, 1\}$.

To this end, based on the copula approach, we assume the multivariate probability distribution function of $\mathbf{Y}$ can be expressed as composition of a copula and suitable marginals. Specifically, following [7] the probability that $\mathbf{Y}$ is equal to $(y_1, \ldots, y_d) \in \{0, 1\}^d$, given a set of covariates $\mathbf{x}$, is obtained by the formula

$$\mathbb{P}\left(\ Y_1 = y_1, \ldots, Y_d = y_d \mid \mathbf{x}\right) =$$
$$= \sum_{v_1=0}^{y_1} \ldots \sum_{v_d=0}^{y_d} (-1)^{\sum_{k=1}^{d}(v_k+y_k)} C\left[F_1\left(v_1 \mid \mathbf{x}\right), \ldots, F_d\left(v_d \mid \mathbf{x}\right)\right] \qquad (1)$$

where the marginal dfs $F_1\left(\cdot \mid \mathbf{x}\right), \ldots, F_d\left(\cdot \mid \mathbf{x}\right)$ are suitable univariate model for $y_1, \ldots, y_d$ (e.g. Logit or Probit model), and $C \in \{C_\theta\}_\theta$ is a suitable copula belonging to a specific family indexed by the parameter $\theta \in \mathbb{R}^d$.

Table 1: Description of the explanatory variables.

Independent variable	Description	Mean (Median)
Characteristics of the trip		
Average level of satisfaction	Average level of satisfaction with some aspects of the destination (continuous values from 6 to 10)	8.334 (8.4)
Visit alone	1= the respondent makes the trip alone; 0 = otherwise	0.116 (0)
Number of paying travellers	Number of travellers who have shared the expenditure of the trip (discrete value from 1 to 7)	1.945 (2)
Number of times in Italy before		
Zero	1= the interviewee visits any city in Italy for the first time; 0 = otherwise	0.051 (0)
Up to 5 times	1= been in Italy from 1 to 5 times before the interview; 0 = otherwise	0.229 (0)
More than 5 times	1= been in Italy more than 5 times before the interview; 0 = otherwise (reference category)	0.720 (1)
Characteristics of the visitor		
Country of origin		
Austrian	1 = the respondent comes from Austria; 0 = otherwise	0.149 (0)
German	1 = the respondent comes from Germany; 0 = otherwise	0.618 (1)
Other country	1 = the respondent comes from a foreign country; 0 = otherwise (reference category)	0.233 (0)
Employment status		
Self-employed	1 = self-employed; 0 = otherwise	0.191 (0)
Office worker	1 = office worker; 0 = otherwise	0.225 (0)
Employee	1 = office employee; 0 = otherwise	0.311 (0)
Retired	1 = retired person; 0 = otherwise	0.220 (0)
Other	1 = other occupation; 0 = otherwise (reference category)	0.054 (0)
Age		
Less than 35 years old	1 = less than 35 years old; 0 = otherwise (reference category)	0.131 (0)
35-44 years old	1 = 35-44 years old; 0 = otherwise	0.267 (0)
45-64 years old	1 = 45-64 years old; 0 = otherwise	0.425 (0)
More than 64 years old	1 = 65 years old and over; 0 = otherwise	0.176 (0)

Notes: For the dichotomous variables, the mean value is equal to the proportion of 1's in the sample.

For instance, when $d = 2$, the conditional distribution of (Y_1, Y_2) given $\mathbf{x}$ can be calculated as in Table 2.

Table 2: Conditional distribution of (Y_1, Y_2) as a function of $\pi_j(\mathbf{x})$, $j = 1, 2$.

	$Y_2 = 0$	$Y_2 = 1$
$Y_1 = 0$	$C(\bar{\pi}_1, \bar{\pi}_2)$	$\bar{\pi}_1 - C(\bar{\pi}_1, \bar{\pi}_2)$
$Y_1 = 1$	$\bar{\pi}_2 - C(\bar{\pi}_1, \bar{\pi}_2)$	$1 - \bar{\pi}_1 - \bar{\pi}_2 + C(\bar{\pi}_1, \bar{\pi}_2)$

The estimation of model parameters for the copula and the logistic regressions can be made by maximizing the global log–likelihood given by

$$\sum_{i=1}^{n} \log \left\{ \sum_{v_1=0}^{y_1} \cdots \sum_{v_d=0}^{y_d} (-1)^{\sum_{k=1}^{d}(v_k+y_k)} C_\theta \left[\widehat{F_1}(v_1 \mid \mathbf{x}), \ldots, \widehat{F_d}(v_d \mid \mathbf{x}) \right] \right\} \quad (2)$$

$$\sum_{i=1}^{n} \log \left\{ \sum_{v_1=0}^{y_{1i}} \cdots \sum_{v_d=0}^{y_{di}} (-1)^{\sum_{k=1}^{d}(v_k+y_{ki})} C_\theta \left[\widehat{F_1}(v_1 \mid \mathbf{x}_i), \ldots, \widehat{F_d}(v_d \mid \mathbf{x}_i) \right] \right\} \quad (3)$$

where $\{y_{ji}, \mathbf{x}_i\}$ is the i-th observation, for $i = 1, \ldots, n$, and

$$\widehat{F_j}(0 \mid \mathbf{x}_i) = \frac{1}{1 + \exp\left(\mathbf{x}_i^\top \widehat{\beta_j}\right)}, \quad (4)$$

while

$$\widehat{F_j}(1 \mid \mathbf{x}_i) = 1 - \widehat{F_j}(0 \mid \mathbf{x}_i). \quad (5)$$

In practice, however, the maximization of Equation (3) is usually computationally cumbersome, since it requires the evaluation of the copula (or the copula density) on 2^d vertices of a rectangle in $[0, 1]^d$. To overcome such difficulties, we follow the approach by [4], which is based on composite likelihood estimation proposed by [10]. It can be summarized in the following two steps:

1. Univariate models for each marginal distribution are fitted separately. In particular, since each Y_j is described by the logistic regression model specified as follows:

$$\pi_j(\mathbf{x}) = \frac{\exp\left(\mathbf{x}^\top \beta_j\right)}{1 + \exp\left(\mathbf{x}^\top \beta_j\right)} \quad (6)$$

where β_j is the $(j + 1)-$dimensional vector including the intercept and the regression coefficients for the $j-$th variable. The estimation of the marginal models is performed by maximum likelihood and the estimates $\widehat{\beta_1}, \ldots, \widehat{\beta_d}$ are obtained. Notice that another binary model (e.g., Probit) may be used as well.

2. A suitable copula C on the estimated margins is fitted. Specifically, we suppose that C belongs to the family of multivariate Student t–copulas that are characterized by a parameter $\nu > 0$, called degree of freedom, and by the parameters of a correlation matrix $(\rho_{k\ell})$, $k, \ell = 1, \ldots, d$. Now, supposing that ν is held fixed, the estimation of the correlation parameters can be made by solving a system of $d(d-1)/2$ equations related to the score functions (for more details, see [4]).

The copula model considered in the previous steps is quite convenient. First, the selection of an appropriate dependence structure C can be performed independently from the inference on the marginal distributions. Moreover, the adopted copula family (Student t family) is quite flexible since it enables us to describe both negative and positive pairwise association among the random variables under consideration. This is particularly useful in our context, since two purchase decisions may be both positively and negatively correlated.

4 Model results and discussion

Following the two-step estimation method described in Section 3, five univariate logistic regression models for each variable in $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5)^\top$ were estimated using White's robust standard variance-covariance matrix [9]. The regression models were first estimated considering the whole set of explanatory variables presented in Table 1, then a stepwise procedure was adopted in order to sequentially drop the variables that were non-significant or significant only in one out of five equations for a $\alpha = 0.1$ significance level. Table 3 shows the reduced models obtained after this backward procedure.

Once the univariate marginal had been fitted, a score test of independence was performed to check whether the expenditures are conditionally uncorrelated given the explanatory variables. Table 4 shows the corresponding test statistics, denoted by z_{obs}. The resulting procedure should reject the assumption of conditional independence if z_{obs} is larger in absolute value than a critical value derived from the standard Gaussian distribution. As can be seen, at a confidence level of 95%, we can reject the assumption of independence between Accommodation (Y1) and Transportation (Y4), and Accommodation (Y1) and Other services (Y5); while at a confidence level of 90%, we can reject the assumption of independence between Food & Beverages (Y2) and Other services (Y5), and Transportation (Y4) and Other services (Y5). The other pairs, instead, seem to exhibit a weaker dependence.

Given the values of these statistics, for the sake of illustration we concentrated our attention on the trivariate model formed by the expenditures related to Accommodation (Y1), Transportation (Y4) and Other services (Y5), which exhibit a stronger evidence of association. Table 5 reports the composite likelihood estimates of the pairwise correlations, along with their

Table 3: Stepwise Logit regression coefficients.

Independent variables	Accommodation[a] (Y_1)	Food & Beverages[b] (Y_2)	Shopping[c] (Y_3)	Transportation[d] (Y_4)	Other services[e] (Y_5)
Average level of satisfaction	1.222*** (0.205)	0.152 (0.179)	0.285** (0.122)	1.134*** (0.174)	0.005 (0.121)
Retired	−1.214** (0.358)	−0.337 (0.293)	0.250 (0.244)	−0.523** (0.254)	−0.267 (0.249)
Austrian	−2.294*** (0.500)	−1.395** (0.480)	−0.300 (0.301)	−1.904*** (0.381)	−1.732*** (0.353)
German	0.262 (0.485)	−0.765* (0.432)	0.865*** (0.233)	−0.309 (0.339)	−1.214*** (0.220)
Number of paying travelers	0.658** (0.290)	0.253* (0.148)	0.155 (0.114)	1.380*** (0.363)	0.021 (0.099)
Constant	−7.935*** (1.680)	1.128 (1.492)	−2.341** (1.004)	−9.629*** (1.435)	0.291 (1.032)

Notes: ***Significant at $p \leqslant 0.01$, **Significant at $p \leqslant 0.05$, *Significant at $p \leqslant 0.1$. Whites robust standard variance-covariance matrix (White, 1980) has been used to estimate the robust standard errors provided in brackets.
[a] $N = 533$; Wald $\chi^2(5) = 93.38$; Prob $> \chi^2 = 0$; Log pseudolikelihood= -118.35; McKelvey & Zavoina's R^2=0.52
[b] $N = 533$; Wald $\chi^2(5) = 17.83$; Prob $> \chi^2 = 0$; Log pseudolikelihood=-185.41; McKelvey & Zavoina's R^2=0.10
[c] $N = 533$; Wald $\chi^2(5) = 39.99$; Prob $> \chi^2 = 0$; Log pseudolikelihood=-307.86; McKelvey & Zavoina's R^2=0.10
[d] $N = 533$; Wald $\chi^2(5) = 95.66$; Prob $> \chi^2 = 0$; Log pseudolikelihood=-187.74; McKelvey & Zavoina's R^2=0.57
[e] $N = 533$; Wald $\chi^2(5) = 42.96$; Prob $> \chi^2 = 0$; Log pseudolikelihood=-320.18; McKelvey & Zavoina's R^2=0.10

standard errors assuming either a Gaussian copula or a Student t–copula with degrees of freedom equal to 2, 5 or 10, respectively.

Table 4: Score test of independence (z_{obs}) among all possible pairs of response variables.

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1		1.4990	0.6304	3.2730	3.3255
Y_2			-0.5398	-1.1045	1.9434
Y_3				0.7583	-1.6897
Y_4					1.8465
Y_5					

Table 5: Estimates of the pairwise correlations and their standard errors in four meta-elliptical copula models.

Pair	t_2 $\hat{\rho}$	SE	t_5 $\hat{\rho}$	SE	t_{10} $\hat{\rho}$	SE	Gaussian $\hat{\rho}$	SE
Y_1-Y_4	-0.8672	0.0015	-0.7117	0.0010	-0.6229	0.0004	-0.5145	0.0000
Y_1-Y_5	-0.8921	0.0040	-0.7621	0.0048	-0.6746	0.0053	-0.5614	0.0059
Y_4-Y_5	-0.8526	0.0050	-0.7392	0.0048	-0.6822	0.0038	-0.6200	0.0013

To exploit such a model, Figure 1 reports the estimated probability of spending in all the three considered categories by varying all the explanatory variables, while the "Average level of satisfaction" is fixed at its average value. Analogously, Figure 2 shows the estimated probabilities by varying all the explanatory variables except for the "Number of paying travelers", which equals its average value. In both cases, the chosen copula model is the Gaussian copula. The models with a Student t–copula have also been considered, but the results are similar and, hence, are not reported here.

Regarding the country of origin, visitors from other foreign countries, excluding Germany and Austria, present higher estimated probabilities of spending on the three expenditure categories simultaneously, regardless of the values assumed by the explanatory variables. Austrian and retired tourists present the lowest estimated probabilities of spending simultaneously. This latter finding can probably be explained by geographical proximity that reduces the probability of their staying at least one night in a South-Tyrolean accommodation for holiday purposes.

Figure 1 reveals that the estimated propensity to spend on Accommodation, Transportation, and on Other services simultaneously increases if the number of paying travellers increases, but only up to four, because for a higher number of paying visitors the propensity becomes quite stable. The Austrian tourists, however, show quite different behaviour since the estimated probability assumes not negligible values only when the paying travellers are more than two, until stable levels are reached with bigger groups (i.e. six visitors). This finding is in line with the low estimated probability of spending on the three expenditure categories simultaneously within this group of visitors.

Focusing on Figure 2 we observe how the estimated probability of spending on Accommodation, Transportation, and Other services simultaneously is significantly affected by the level of satisfaction with the destination. In fact, in the literature it was often recognized that overall satisfaction stimulates higher profitability. The Austrian tourists show an increased estimated probability of spending only for very high satisfaction levels, but, again, this is probably due to the low estimated probability of spending as before.

To summarize, the highest estimated probabilities of spending on the three considered expenditure categories simultaneously was observed for employed foreign tourists (excluding those from Germany or Austria), who are overall very satisfied with the destination, and who have visited the South-Tyrol in a large groups in which 6 or 7 are paying travellers.

5 Conclusions

In this paper a copula–based approach is suggested for studying tourism consumption behaviour, i.e. the probability of spending at a given destination for different goods and services. A sample of international visitors
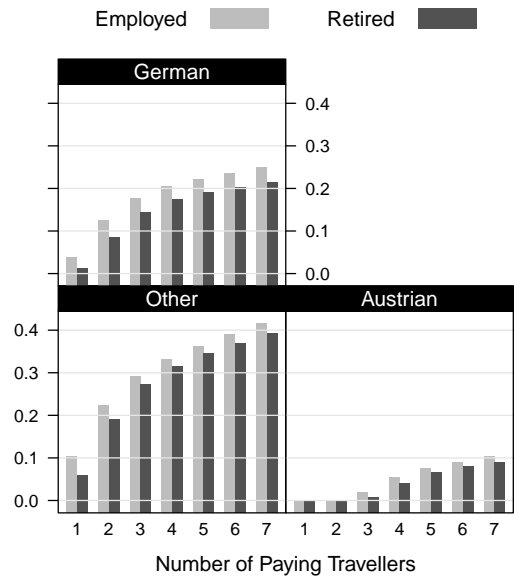
Fig. 1: Estimated probabilities obtained taken the "Average level of satisfaction" fixed at its average value and varying the other explanatory variables.
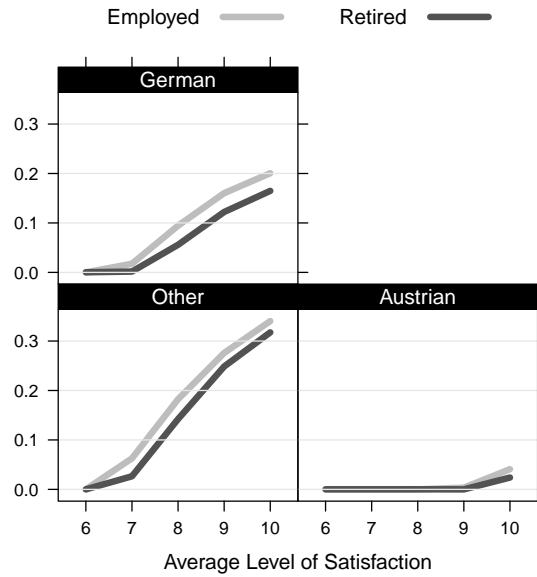

Fig. 2: Estimated probabilities obtained taken the "Number of paying travellers" fixed at its average value and varying the other explanatory variables.

to the South Tyrol region (Northern Italy) in 2011 was analysed to illustrate the main features of the method in tourism. The results suggest that a stronger dependence exists between the consumption of Accommodation, Transportation, and Other services; a weaker dependence exists between Food & Beverages, Other services, and Transportation; while the hypothesis of independence is not rejected ($\alpha = 0.1$) for the other combinations of tourism categories.

Focusing on the triplet of expenditures on Accommodation, Transportation, and Other services, the paper illustrates how a set of explanatory variables affect the joint probability of spending simultaneously on these three categories during the same trip. Somehow surprisingly, age and employment status, apart from being retired, do not significantly affect the joint consumption of these three commodities that is affected, on the other hand, by the number of paying travellers, the country of origin of the visitors, and the level of satisfaction towards the destination. To sum up, employed foreign visitors (excluding those from Germany and Austria), who are overall very satisfied with the destination, and who have visited the South-Tyrol in a large group in which 6-7 are paying travellers, present the highest estimated probabilities of simultaneously spending on Accommodation, Transportation, and Other services. Thus, our results highlight that the probability of spending on different tourism goods and services simultaneously is affected not only by economic variables, but also by other socio-demographic and psychographical variables; the level of satisfaction, in particular, plays an important role.

Overall, the findings are of potential interest in tourism management and/or local government policy, in order to know how visitors decide to allocate their travel budget among different combinations of tourism expenditure categories. Managing this information is fundamental for policy makers and marketing experts in order to improve the touristic supply and to implement specific marketing campaigns that offer a combination of different services (meals, lodging, shopping, etc.) according to tourists' preferences.

References

1. Alegre, J., Pou, L. (2004). Microeconomic determinants of the probability of tourism consumption. Tourism Economics, Vol 10, No 2, pp 125144.
2. ASTAT (2013). Turismo in Alto Adige. Anno turistico 2011/12. ASTAT, Provincia Autonoma di Bolzano (http://www.provincia.bz.it/astat/it/mobilita-turismo/).
3. Durante, F., Sempi, C. (2015). Principles of copula theory. Boca Raton, FL: CRC/Chapman and Hall.
4. Genest, C., Nikoloulopoulos, A., Rivest, L., and Fortin, M. (2013). Predicting dependent binary outcomes through logistic regressions and meta-elliptical copulas. Brazilian Journal of Probability and Statistics, Vol 27, No 3, pp 265284.
5. Hung, W.-T., Shang, J.-K., Wang, F.-C. (2012). Another look at the determinants of tourism expenditure. Annals of Tourism Research, 39(1): 495–498.

6. Pudney, S. (1989). Modelling Individual Choice: the Econometrics of corners, Kinks, and Holes. London: Basil Blackewll.

7. Song, P. X.-K. (2007). Correlated data analysis: modeling, analytics, and applications. Springer, New York, NY.

8. Syriopoulos, T., Sinclair, M. (1993). An econometric study of tourism demand: the AIDS model of US and European tourism in Mediterranean countries. Applied Economics, Vol 25, No 12, 15411552.

9. White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica, Vol 48, No 4, pp 817838.

10. Zhao, Y., abd Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. Canadian Journal of Statistics, Vol 33, No 3, pp 335356.