This article was downloaded by: [Knihovna Univerzity Palackeho]
On: 06 May 2013, At: 06:06
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Modelling interval data with Normal and Skew-Normal distributions

Paula Brito [a] & A. Pedro Duarte Silva [b]
[a] Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal
[b] Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal
Published online: 11 May 2011.

To cite this article: Paula Brito & A. Pedro Duarte Silva (2012): Modelling interval data with Normal and Skew-Normal distributions, Journal of Applied Statistics, 39:1, 3-20

To link to this article: http://dx.doi.org/10.1080/02664763.2011.575125

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Journal of Applied Statistics
Vol. 39, No. 1, January 2012, 3–20

Taylor & Francis
Taylor & Francis Group

Modelling interval data with Normal and Skew-Normal distributions

Paula Brito[a]* and A. Pedro Duarte Silva[b]

[a] *Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Porto, Portugal;* [b] *Faculdade de Economia e Gestão & CEGE, Universidade Católica Portuguesa at Porto, Porto, Portugal*

(*Received 11 October 2009; final version received 21 March 2011*)

A parametric modelling for interval data is proposed, assuming a multivariate Normal or Skew-Normal distribution for the midpoints and log-ranges of the interval variables. The intrinsic nature of the interval variables leads to special structures of the variance–covariance matrix, which is represented by five different possible configurations. Maximum likelihood estimation for both models under all considered configurations is studied. The proposed modelling is then considered in the context of analysis of variance and multivariate analysis of variance testing. To access the behaviour of the proposed methodology, a simulation study is performed. The results show that, for medium or large sample sizes, tests have good power and their true significance level approaches nominal levels when the constraints assumed for the model are respected; however, for small samples, sizes close to nominal levels cannot be guaranteed. Applications to Chinese meteorological data in three different regions and to credit card usage variables for different card designations, illustrate the proposed methodology.

Keywords: symbolic data; parametric modelling of interval data; statistical tests for interval data; Skew-Normal distribution; ANOVA; MANOVA

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

1. Introduction

In classical multivariate data analysis, data are represented in an $n \times p$ data-matrix where n "individuals" (usually in rows) take exactly one value for each variable (usually in columns). This structure is however too simple to represent more complex data, where the information for an individual on each variable is not reduced to a single value. Symbolic data analysis [5,7,11] provides a framework where new variable types allow to take directly into account variability and/or uncertainty associated to each single "individual", avoiding restrictive summarizations to impose a fit to the classical representation structure. Symbolic data extend the classical tabular model by allowing multiple, possibly weighted, values for each variable. New variable types – interval, categorical multi-valued and modal variables – are introduced.

*Corresponding author. Email: mpbrito@fep.up.pt

ISSN 0266-4763 print/ISSN 1360-0532 online
© 2012 Taylor & Francis
http://dx.doi.org/10.1080/02664763.2011.575125
http://www.tandfonline.com

In this paper, we focus on the analysis of interval data, i.e., where elements are characterized by variables whose values are intervals on $\mathbb{R}$. Interval data may occur in many different situations. We may have *native* interval data, when describing ranges of variable values – for example, daily stock prices or temperature ranges. Interval variables also allow dealing with imprecise data, coming from repeated measures or confidence interval estimation. A natural source of interval data is the aggregation of huge data bases, when real values describing the individual observations lead to intervals describing the aggregated data. Original symbolic data – concerning, for instance, descriptions of biological species or technical specifications – constitute yet another possible source of interval data.

In the context of interval data, mention should be made to Interval Calculus [9,15], a discipline that has derived rules for dealing with interval values.

Most existing methods developed so far for the analysis of symbolic data consider non-parametric exploratory approaches [7,11]. Our goal is to develop parametric inference methodologies based on probabilistic models for interval variables. In the proposed approach, each interval is represented by its midpoint and log-range, for which Normal or Skew-Normal distributions are assumed. Therefore, as we do not operate directly on intervals, rules of Interval Calculus do not apply in our framework. The intrinsic nature of the interval variables leads to special structures of the variance–covariance matrix, which are represented by five different possible configurations. We show that in Normal models maximum likelihood inference can be carried out with simple analytical formulae in four out of the five configurations. In all configurations of the Skew-Normal model, as well as in the remaining configuration of the Normal model, maximum likelihood inference is performed with numerical optimization procedures, which are studied and implemented. This methodology is then considered in the context of (M)ANOVA testing, and statistical properties are analysed through a simulation experiment.

The proposed approach provides the adequate tool for hypothesis testing when variables are intrinsically interval-valued.

Consider a situation where you wish to analyse whether some performance variables (e.g. sales, nb. customers, etc.) vary among different towns for shops of a given chain, for each of which you have daily data. Although shops may be considered independent, daily observations for a given shop are not independent, so that the whole set of daily observations may not be considered a random sample. Also, if we summarize data for a given shop by an average, the information as concerns its variation from day to day is lost. Aggregating these data in the form of intervals enables us to keep relevant information. A multivariate analysis of variance (MANOVA) for the resulting interval data then allows investigating whether performance variables vary across towns.

To the best of our knowledge, few authors have addressed inference problems in this context. Some steps in this direction have been taken in [6]. Related problems have been considered in [12] where some inferential procedures for "interval-valued random sets" are developed, based on the rules of Interval Calculus and appropriate definitions of integrals and expected values for these sets. These procedures have been applied to analysis of variance (ANOVA) and ANCOVA in [10]. In our paper, a simpler but more general approach is pursued, by modelling an interval directly by its midpoint and range. Krätschmer [14] has addressed the problem of parameter estimation with random fuzzy sets, following a different approach.

The structure of this paper is the following: Section 2 introduces interval variables, presents the interval representation to be used in the sequel and fixes notation. A modelling based on the Normal distribution and its application to M(ANOVA) is presented in Section 3. In Section 4, the Skew-Normal distribution is reviewed, and applied to the modelling and analysis of interval data. Section 5 presents a simulation study, while Section 6 describes two applications to real data. Section 7 concludes the paper, presenting perspectives for further research. Analytical gradients required for the numerical optimization of the log-likelihood in the Skew-Normal model, for all configurations, are given in the appendix.

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 1. Matrix I of interval data.

	Y_1	...	Y_j	...	Y_p
s_1	$[l_{11}, u_{11}]$	...	$[l_{1j}, u_{1j}]$	...	$[l_{1p}, u_{1p}]$
...	...		...		...
s_i	$[l_{i1}, u_{i1}]$	...	$[l_{ij}, u_{ij}]$	...	$[l_{ip}, u_{ip}]$
...	...		...		...
s_n	$[l_{n1}, u_{n1}]$	...	$[l_{nj}, u_{nj}]$	...	$[l_{np}, u_{np}]$

2. Parametric representation of interval data

Given a set of n "individuals" $S = \{s_1, \ldots, s_n\}$, an interval variable is defined by an application $Y : S \to T$ such that $s_i \to Y(s_i) = [l_i, u_i]$, where T is the set of intervals of an underlying set $O \subseteq \mathbb{R}$. Let I be an $n \times p$ matrix representing the values of p interval variables on S. Each $s_i \in S$ is represented by a p-uple of intervals, $I_i = (I_{i1}, \ldots, I_{ip})$, $i = 1, \ldots, n$, with $I_{ij} = [l_{ij}, u_{ij}]$, $j = 1, \ldots, p$ (see Table 1). The value of an interval variable Y_j for each $s_i \in S$ is hence defined by the bounds l_{ij} and u_{ij} of $I_{ij} = Y_j(s_i)$. For modelling purposes, a preferable equivalent parametrization consists in representing $Y_j(s_i)$ by the midpoint $c_{ij} = (l_{ij} + u_{ij})/2$ and range $r_{ij} = u_{ij} - l_{ij}$ of I_{ij}.

In the next sections, we propose a parametric model for interval data, based on this representation.

3. Normal model

Let us consider each interval I_{ij} represented by its midpoint c_{ij} and range r_{ij}. In the first model, we assume that the joint distribution of the midpoints C and the logs of the ranges R is multivariate Normal, i.e., $R^* = \ln(R)$, $(C, R^*) \sim N_{2p}(\mu, \Sigma)$, with $\mu = [\mu_C^t, \mu_{R^*}^t]^t$ and $\Sigma = \left(\begin{smallmatrix} \Sigma_{CC} & \Sigma_{CR^*} \\ \Sigma_{R^*C} & \Sigma_{R^*R^*} \end{smallmatrix}\right)$ where μ_C and μ_{R^*} are p-dimensional column vectors of the mean values of, respectively, the midpoints and log-ranges, and Σ_{CC}, Σ_{CR^*}, Σ_{R^*C} and $\Sigma_{R^*R^*}$ are $p \times p$ matrices containing their variances and covariances.

This model has the advantage that it allows for a straightforward application of classical inference methods. If we consider the intervals' midpoints as location indicators of the variables' values, assuming that they follow a joint Normal distribution corresponds to the usual Gaussian assumption for classical data. By considering the log transformation of the ranges, we overcome the difficulties created by their limited domain. An obvious implication of this model is that the marginal distributions of the midpoints are Normals and those of the ranges are Log-Normals. In Section 4, we consider more general models that try to alleviate some of the known limitations of the multivariate Normal distribution.

It should be emphasized that the midpoint c_{ij} and the range r_{ij} of the value of an interval variable are two quantities related to one only variable, and should therefore not be considered separately. One contribution of this work is to offer parameterizations of the global covariance matrix that take into account the link that may exist between midpoints and log-ranges of the same or different variables. Intermediate parameterizations between the non-restricted and the non-correlation setup usually considered are particularly relevant for the specific case of interval data.

In the most general formulation, we allow for non-zero correlations among all midpoints and log-ranges (configuration 1); other cases of interest are:

- Midpoints (respectively, ranges) of different variables may be correlated, the midpoint of each variable may be correlated with its range, but no correlation between midpoints and ranges of different variables is allowed (configuration 2);

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

- The interval variables Y_j are independent, but for each variable, the midpoint may be correlated with its range (configuration 3);
- Midpoints (respectively, ranges) of different variables may be correlated, but no correlation between midpoints and ranges is allowed (configuration 4);
- All midpoints and ranges are uncorrelated, both among themselves and between each other (configuration 5).

Table 2 summarizes the different possibilities. We note that in this framework imposing non-correlations with log-ranges is equivalent to imposing non-correlations with ranges. This follows from the normality assumption and the well-known equivalence between non-correlation and independence in the multivariate Normal distribution. Note that configuration 2 is a particular case of 1 and both 3 and 4 are particular cases of 2, configuration 5 being a particular case of all the others.

It should be remarked that in cases 3, 4 and 5, Σ can be written as a diagonal by blocks matrix, after a possible rearrangement of rows and columns. This is directly the case for configurations 4 and 5; in configuration 3, the rows and columns of Σ must be rearranged such that the row (resp. column) corresponding to the midpoint of each variable is immediately followed by the row (resp. column) corresponding to its range. It then follows that in configuration 4 the matrix Σ is formed by two $p \times p$ blocks, whereas in configuration 3 there are p 2×2 blocks and in configuration 5 the $2p$ blocks are single real elements.

Testing configurations 3, 4 and 5 against 1 amounts to testing for the independence of sets of variables. Tests for this problem are well known and may be found, for instance, in [16,17]. In any case, for this problem, as well as for testing configuration 2 against 1 and configurations 3, 4 and 5 against a more general configuration other than 1, the likelihood ratio principle may be applied.

3.1 *Maximum likelihood estimation*

Let $X_i = [C_i^{\mathrm{t}}, R_i^{*\mathrm{t}}]^{\mathrm{t}}$ be the $2p$-dimensional column vector comprising all the midpoints and log-ranges for s_i. Let $\bar{X}$ be the sample mean of the X_i's. The maximum likelihood estimators of μ and Σ under configuration 1 are obviously the classical ones, $\hat{\mu} = \bar{X}$ and $\hat{\Sigma} = (1/n)\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{t}} := (1/n)E$. We will now show that the maximum likelihood estimators of μ and Σ for configurations 3, 4 and 5 are obtained from the non-restricted estimators simply replacing by zeros the null parameters in the model for Σ.

For all configurations, the likelihood function is

$$L(\mu, \Sigma) = (2\pi)^{-np}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \mu)^{\mathrm{t}}\Sigma^{-1}(X_i - \mu)\right) \tag{1}$$

Table 2. Different configurations.

Configuration	Characterization	Σ
1	Non-restricted	Non-restricted
2	C_j not-correlated with R_ℓ^*, $\ell \neq j$	$\Sigma_{CR^*} = \Sigma_{R^*C}$ diagonal
3	Y_j's non correlated	Σ_{CC}, $\Sigma_{CR^*} = \Sigma_{R^*C}$, $\Sigma_{R^*R^*}$ all diagonal
4	C's non-correlated with R^*'s	$\Sigma_{CR^*} = \Sigma_{R^*C} = 0$
5	All C's and R^*'s are non-correlated	Σ diagonal

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

and the log-likelihood can be written as

$$\ln L(\mu, \Sigma) = -np \ln(2\pi) - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \operatorname{tr} E\Sigma^{-1} - \frac{n}{2}(\bar{X} - \mu)^{\mathrm{t}} \Sigma^{-1} (\bar{X} - \mu). \tag{2}$$

Since Σ^{-1} is symmetric positive definite, the quadratic form term will be a minimum only when μ is equal to $\bar{X}$, so that the maximum-likelihood estimate of the mean vector is always $\bar{X}$, as usual. Then the maximization of the likelihood function with respect to Σ reduces to maximizing

$$\ln L(\mu, \Sigma) = \text{constant} - \frac{n}{2} \ln |\Sigma| - \frac{1}{2} \operatorname{tr} E\Sigma^{-1}. \tag{3}$$

In configurations 3, 4 and 5, Σ is subject to the constraints shown in Table 2. As we have already seen, in these cases Σ can be written as a diagonal by blocks matrix, after a possible rearrangement of rows and columns. Lemma 1 shows that the maximum of $\ln L(\mu, \Sigma)$ in Equation (3) can be obtained by separately maximizing with respect to each block of Σ.

LEMMA 1 *When Σ is diagonal by blocks*

$$\Sigma = \begin{pmatrix} \Sigma_1 & & & \\ & \Sigma_2 & & 0 \\ 0 & & \cdots & \\ & & & \Sigma_q \end{pmatrix},$$

the maximum of the likelihood function (1) *is reached when $\Sigma_h = \hat{\Sigma}_h = E_h/n$ for $h = 1, \ldots, q$ where E_h is the block of E corresponding to Σ_h.*

Proof As shown above, maximizing Equation (1) after replacing μ by $\bar{X}$ reduces to maximizing Equation (3) with respect to Σ. Since in this case $|\Sigma| = |\Sigma_1||\Sigma_2| \ldots |\Sigma_q|$ and

$$\Sigma^{-1} = \begin{pmatrix} \Sigma_1^{-1} & & & \\ & \Sigma_2^{-1} & & 0 \\ 0 & & \cdots & \\ & & & \Sigma_q^{-1} \end{pmatrix}.$$

Equation (3) may be written as

$$\ln L(\mu, \Sigma) = \text{constant} - \sum_{h=1}^{q} \left(\frac{n}{2} \ln |\Sigma_h| + \frac{1}{2} \operatorname{tr} E_h \Sigma_h^{-1} \right). \tag{4}$$

It is well known (see, for instance [17]) that Equation (4) is maximized when $\Sigma_h = \hat{\Sigma}_h = E_h/n$.

This result is not valid for configuration 2, since in that case Σ cannot be written as a block diagonal matrix. As far as we can tell there is no closed form expression for the maximum likelihood estimator in this case; however Equation (3) can always be maximized by standard numerical procedures. In particular, in our implementation, we used the "L-BFGS-B" quasi-Newton algorithm with bounded variables, proposed in [8]. This uses function values and gradients to build up a picture of the surface to be optimized. The optimization procedure takes as arguments the elements of μ and the non-null elements of L, where $\Sigma = LL^{\mathrm{t}}$ denotes the lower-triangular Cholesky decomposition of Σ. ■

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

3.2 *Application to ANOVA and MANOVA*

Different inferential methodologies may be approached by the modelling presented above. For instance, model-based clustering and statistical tests in regression analysis require a parametric modelling of the data generating process. Here, we will discuss ANOVA and MANOVA in this context.

First let us point out that since each interval variable Y_j is modelled by the pair (C_j, R_j^*), it follows that an analysis of variance of Y_j is now accomplished by a two-dimensional MANOVA of (C_j, R_j^*).

To fix ideas, let us assume a one-way design, where the factor has k levels, representing k groups, and n_ℓ be the number of observations in group ℓ. Let $X_{ij} = [C_{ij}, R_{ij}^*]^{\mathrm{t}}$ be the two-dimensional column vector comprising the midpoint and log-range of variable Y_j for s_i. Moreover, let $\bar{X}_{\bullet j\ell}$ and $\mu_{\bullet j\ell}$ be sample and population means of the X_{ij}'s in group ℓ, and $\bar{X}_{\bullet j\bullet}$ the corresponding global sample mean. The null hypothesis in this case consists in stating that all $\mu_{\bullet j\ell}$ are equal across groups.

We will adopt a likelihood ratio approach since different likelihood ratio statistics may be derived for each different configuration in Table 2. For alternative statistics, such as the Lawley–Hotelling trace, the Pillai's trace and the maximum root statistics (see [17]), it is not clear how a similar adaptation could be done. In particular, if given entries in the variance–covariance matrix are set to zero the classical distributions can no longer be guaranteed. These distributions have been deduced assuming that the corresponding matrices follow Wishart distributions, which implies that all the matrix entries are random variables. In the situation at hand some of these entries will be fixed scalars, therefore violating the classical assumptions.

Following the same reasoning as in the previous section, it follows that for configurations 3, 4 and 5, the likelihood ratio statistic is $\lambda = (|E_{\text{alt}}|/|E_{\text{null}}|)^{n/2}$ where E_{null} and E_{alt} are 2×2 matrices corresponding to the null and alternative hypothesis respectively. E_{null} is obtained from $E_j = \sum_{i=1}^{n}(X_{ij} - \bar{X}_{\bullet j\bullet})(X_{ij} - \bar{X}_{\bullet j\bullet})^{\mathrm{t}}$ by replacing the null entries corresponding to each configuration; likewise, E_{alt} is obtained from $\sum_{\ell=1}^{k}\sum_{i=1}^{n_\ell}(X_{ij} - \bar{X}_{\bullet j\ell})(X_{ij} - \bar{X}_{\bullet j\ell})^{\mathrm{t}}$ in the same manner.

Configuration 1 is the classical one, and for configuration 2 the likelihood ratio statistics may be obtained by numerical methods. In all cases, under the null hypothesis, $2\ln\lambda$ follows asymptotically a chi-square distribution with $n - k$ degrees of freedom.

A simultaneous analysis of all the Y's may be accomplished by a $2p$-dimensional MANOVA, following the same procedure.

4. Skew-Normal model

A limitation of the Normal model proposed in the above section is that it imposes a symmetrical distribution on the midpoints and a specific relation between mean, variance and skewness for the ranges. A more general model that overcomes these limitations may be obtained by considering the family of Skew-Normal distributions (see, for instance, [1–4]). This distribution generalizes the Gaussian distribution by introducing an additional shape parameter, while trying to preserve some of its mathematical properties.

The univariate standard Skew-Normal distribution has density

$$f(z;\alpha) = 2\phi(z)\Phi(\alpha z), \quad z \in \mathbb{R}, \tag{5}$$

where α is the shape parameter and ϕ and Φ the density and the distribution function of a $N(0, 1)$ variable, respectively.

A general Skew-Normal variable may be obtained by the transformation $X = \xi + \omega Z$ where Z is a standard Skew-Normal variable and ξ and ω are location and scale parameters. In this

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

case, we write $X \sim \mathrm{SN}(\xi, \omega^2, \alpha)$. The density of a p-dimensional Skew-Normal distribution is given by

$$f(y; \alpha, \xi, \Omega) = 2\phi_p(x - \xi; \Omega)\Phi_p(\alpha^{\mathrm{t}}\omega^{-1}(x - \xi)), \quad x \in \mathbb{R}^p, \tag{6}$$

where now ξ and α are p-dimensional vectors, Ω is a symmetric $p \times p$ positive-definite matrix, ω is a diagonal matrix formed by the square-roots of the diagonal elements of Ω and ϕ_p, Φ_p are, respectively, the density and the distribution function of a p-dimensional standard Gaussian vector.

The variance–covariance matrix of a p dimensional Skew-Normal distribution is given by Azzalini [2]

$$\mathrm{Var}(X) = \Sigma = \Omega - \omega\mu_Z\mu_Z^{\mathrm{t}}\omega, \tag{7}$$

where μ_Z is a vector of expected values for standard Skew-Normal variables,

$$\mu_Z = \sqrt{\frac{2}{\Pi}}\delta \text{ with } \delta = \frac{\omega^{-1}\Omega\omega^{-1}\alpha}{\sqrt{(1 + \alpha^{\mathrm{t}}\omega^{-1}\Omega\omega^{-1}\alpha)}}.$$

Azzalini and Capitanio [3] have obtained the following expression for the log-likelihood of a p-dimensional Skew-Normal distribution:

$$l = \text{constant} - \frac{1}{2}n \ln|\Omega| - \frac{n}{2}\mathrm{tr}(\Omega^{-1}V) + \sum_i \zeta_0(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi)), \tag{8}$$

where $V = n^{-1}\sum_i (x_i - \xi)(x_i - \xi)^{\mathrm{t}}$ and $\zeta_0(x) = \ln(2\Phi(x))$. These authors maximize Equation (8) in two steps. First, they separate the ξ and α arguments from Ω by specifying a new parameter $\eta = w^{-1}\alpha$. Then, the maximization with respect to Ω is given by the well-known result $\hat{\Omega} = V$. The maximization with respect to η and ξ is then performed numerically.

As an alternative to the Normal model, introduced in Section 3, we now consider that (C, R^*) follow jointly a $2p$-multivariate Skew-Normal distribution. This model encompasses mixed models with marginal Normal random variables, for which the corresponding α parameter is null.

In the Skew-Normal model, we will again consider, in addition to the unrestricted configuration, particular cases where midpoints and log-ranges may or may not be correlated among themselves or with each other. Note, however, that now non-correlation is not equivalent to independence and non-correlation with log-ranges is not equivalent to non-correlation with ranges. However, in the context of our models, enforcing null correlations with log-ranges is more natural than with ranges and seems preferable in terms of mathematical tractability.

In the ANOVA and MANOVA of interval variables based on the likelihood ratio principle, we need to maximize Equation (8) for the null (mean vectors equal across groups) and the alternative hypothesis. In this case, the optimal likelihood solution for restricted configurations in Table 2 may not be obtained by simply replacing corresponding entries in the appropriate matrices. This is due to the fact that the restrictions now imply nonlinear relations between the parameters in Ω and α. Given that $\Sigma = \Omega - \omega\mu_Z\mu_Z^{\mathrm{t}}\omega$, a null covariance $\Sigma(j, j')$ implies that $\Omega(j, j') = \Omega(j, j)^{1/2}\mu_{Z_j}\Omega(j', j')^{1/2}\mu_{Z_{j'}}$ or, equivalently

$$\Sigma(j, j') = 0 \Rightarrow \Omega(j, j') = \frac{2}{\pi}\frac{1}{1 + \alpha^{\mathrm{t}}\omega^{-1}\Omega\omega^{-1}\alpha}\Omega_j^{\mathrm{t}}\omega^{-1}\alpha\alpha^{\mathrm{t}}\omega^{-1}\Omega_{j'} \tag{9}$$

where Ω_j denotes the jth column of matrix Ω.

Therefore, for configurations 2–5 in Table 2, we impose condition (9) for the null elements of matrix Σ in each case. This condition defines a system of nonlinear equations on the $\Omega(j, j')$, which may be solved by standard numerical procedures.

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

The maximization of the log-likelihood (8) under configurations 2–5 cannot be performed in two separate steps as in the non-restricted case (configuration 1) since, due to conditions (9), α can no longer be separated from Ω. Therefore, we propose to employ a quasi-Newton algorithm, using as arguments all the elements from ξ and α and the free non-null elements of Ω. We replicate the procedure several times from different starting points (to avoid local optima). In order to implement this algorithm, we need the gradients of the log-likelihood, for which the required derivatives are given in the appendix.

5. Simulation

To study the behaviour of some of the above proposed tests, a restricted simulation study was performed. We generated interval data by simulating midpoints and log-ranges from multivariate Normal and Skew-Normal distributions; for the Skew-Normal distribution the α parameter has been set so that the univariate skewness coefficient equals 0.75. We have considered in all cases $k = 3$ balanced groups, and a full factorial design was employed, with the following factors:

- Number of interval variables: $p = 1$ and $p = 5$.
- Sample size: $n = 60 = 20 + 20 + 20$ (small sample), $n = 300 = 100 + 100 + 100$, and $n = 600 = 200 + 200 + 200$ (large sample) elements.
- Group separation:
  - (i) no separation;
  - (ii) midpoints well-separated, log-ranges well-separated;
  - (iii) midpoints well-separated, log-ranges badly-separated;
  - (iv) midpoints badly separated, log-ranges well-separated;
  - (v) midpoints badly separated, log-ranges badly separated.
- Data configuration (see Table 2): Normal with configurations 1, 3, 4 and 5, Skew-Normal with configuration 1.

The reason for not having considered configuration 2 in the Normal model and configurations 2–5 in the Skew-Normal model in the full factorial design was their heavy computational requirements. However, in the illustations presented below, both models with all configuations are considered.

"No-separation" corresponds to having the mean values (location parameter in the case of Skew-Normal distributions) of all variables (both for midpoints and log-ranges) in all groups equal to 0; in the case of "good-separation" these means are set to 0, 0.25 and 0.5, respectively; finally, "bad-separation" is defined by having these means equal to 0, 0.1 and 0.2. Variances are always set to 1.

Correlations have been set as follows. It has been assumed that all interval variables have been ordered with neighbouring variables having the strongest correlations. Then, for all j, the variable $X_{j(q)}$ (with $q = 1$ for midpoints and $q = 2$ for log-ranges) were generated according to a factorial model

$$X_{j(q)} = \mu_{j(q)} + 0.9^{j/2}(\beta_{1(q)} z_{mpt} + \beta_{2(q)} z_{lnr} + \beta_4 z_0) + \beta_3 z_j + \beta_5 z_{j(q)}, \tag{10}$$

where z_{mpt}, z_{lnr}, z_0 and z_j are independently generated standard Gaussian variables responsible for the different types of correlations considered and $z_{j(q)}$ are independent standard Gaussian variables responsible for specific variances.

The loadings associated, respectively, with correlation between midpoints, log-ranges, midpoints and log-ranges of the same variables and midpoints and log-ranges of different variables,

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

when not null by design, have been set to

$$\beta_{1(1)} = \beta_{2(2)} = \sqrt{0.6}; \quad \beta_{1(2)} = \beta_{2(1)} = 0.0; \quad \beta_3 = \sqrt{0.8}; \quad \beta_4 = \sqrt{0.4}.$$

The loading associated with specific variances was set to

$$\beta_5 = \sqrt{\text{Max}\{0.1, 1 - \beta_{1(1)}^2 - \beta_{2(2)}^2 - \beta_3^2 - \beta_4^2\}}.$$

Table 3. Summary results for one Interval Variable.

Sample size	p	Pop. config.	Midpoints separation	Log-Ranges separation	Normal config. 1	Normal config. 5	Skew Normal config. 1
Normal, configuration 1 results							
60	1	Norm 1	None	None	0.0700	0.1025	0.2300
60	1	Norm 1	Bad	Bad	0.1025	0.1475	0.2400
60	1	Norm 1	Good	Bad	0.2700	0.2475	0.4300
60	1	Norm 1	Bad	Good	0.3175	0.2425	0.4325
60	1	Norm 1	Good	Good	0.2375	0.4225	0.3700
300	1	Norm 1	None	None	0.0425	0.0925	0.0550
300	1	Norm 1	Bad	Bad	0.2075	0.3300	0.2025
300	1	Norm 1	Good	Bad	0.9425	0.7875	0.9375
300	1	Norm 1	Bad	Good	0.9375	0.8200	0.9300
300	1	Norm 1	Good	Good	0.8675	0.9550	0.8400
600	1	Norm 1	None	None	0.0425	0.0850	0.0550
600	1	Norm 1	Bad	Bad	0.3450	0.5500	0.3600
600	1	Norm 1	Good	Bad	1.0000	0.9825	1.0000
600	1	Norm 1	Bad	Good	1.0000	0.9850	1.0000
600	1	Norm 1	Good	Good	0.9950	1.0000	0.9875
Normal, configuration 5 results							
60	1	Norm 5	None	None	0.0425	0.0375	0.2050
60	1	Norm 5	Bad	Bad	0.1150	0.1200	0.2775
60	1	Norm 5	Good	Bad	0.2825	0.2775	0.4050
60	1	Norm 5	Bad	Good	0.2825	0.2825	0.4200
60	1	Norm 5	Good	Good	0.4075	0.4200	0.4850
300	1	Norm 5	None	None	0.0425	0.0450	0.0825
300	1	Norm 5	Bad	Bad	0.3250	0.3250	0.3375
300	1	Norm 5	Good	Bad	0.8925	0.8900	0.8700
300	1	Norm 5	Bad	Good	0.8700	0.8700	0.8650
300	1	Norm 5	Good	Good	0.9850	0.9875	0.9800
600	1	Norm 5	None	None	0.0425	0.0450	0.0550
600	1	Norm 5	Bad	Bad	0.5650	0.5675	0.6025
600	1	Norm 5	Good	Bad	0.9975	0.9975	0.9975
600	1	Norm 5	Bad	Good	1.0000	1.0000	1.0000
600	1	Norm 5	Good	Good	1.0000	1.0000	1.0000
Skew Normal, configuration 1 results							
60	1	SkN 1	None	None	0.0475	0.0725	0.1225
60	1	SkN 1	Bad	Bad	0.0550	0.0900	0.1925
60	1	SkN 1	Good	Bad	0.1375	0.1475	0.3125
60	1	SkN 1	Bad	Good	0.1525	0.1575	0.2775
60	1	SkN 1	Good	Good	0.1425	0.2550	0.3200
300	1	SkN 1	None	None	0.0525	0.0800	0.0325
300	1	SkN 1	Bad	Bad	0.0950	0.1850	0.1125
300	1	SkN 1	Good	Bad	0.5675	0.4650	0.6075
300	1	SkN 1	Bad	Good	0.5325	0.5100	0.5875
300	1	SkN 1	Good	Good	0.5525	0.7025	0.7550
600	1	SkN 1	None	None	0.0525	0.0775	0.0150
600	1	SkN 1	Bad	Bad	0.2325	0.3050	0.2500
600	1	SkN 1	Good	Bad	0.8575	0.8225	0.8975
600	1	SkN 1	Bad	Good	0.8450	0.7925	0.9275
600	1	SkN 1	Good	Good	0.8600	0.9550	0.9825

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 4. Summary results for five interval variables.

Sample size	p	Pop. config.	Midpoints separation	Log-Ranges separation	Normal config. 1	Normal config. 3	Normal config. 4	Normal config. 5	Skew Normal config. 1
Normal, configuration 1 results									
60	5	Norm 1	None	None	0.1275	0.1675	0.1375	0.1150	0.3750
60	5	Norm 1	Bad	Bad	0.1550	0.1975	0.1800	0.2175	0.4800
60	5	Norm 1	Good	Bad	0.2725	0.3650	0.2650	0.3850	0.4725
60	5	Norm 1	Bad	Good	0.2075	0.3975	0.2050	0.3575	0.4800
60	5	Norm 1	Good	Good	0.2400	0.4325	0.3125	0.5250	0.5250
300	5	Norm 1	None	None	0.0625	0.1175	0.1075	0.1000	0.0650
300	5	Norm 1	Bad	Bad	0.1325	0.3525	0.2375	0.4700	0.5675
300	5	Norm 1	Good	Bad	0.5650	0.8725	0.6075	0.8900	0.6050
300	5	Norm 1	Bad	Good	0.6300	0.9325	0.6175	0.9150	0.5675
300	5	Norm 1	Good	Good	0.7150	0.9450	0.8450	0.9825	0.7025
600	5	Norm 1	None	None	0.0325	0.1125	0.0800	0.1225	0.0575
600	5	Norm 1	Bad	Bad	0.2625	0.5225	0.4100	0.7150	0.8775
600	5	Norm 1	Good	Bad	0.9275	0.9975	0.9100	0.9975	0.9150
600	5	Norm 1	Bad	Good	0.9325	0.9975	0.9100	0.9975	0.8775
600	5	Norm 1	Good	Good	0.9800	1.0000	0.9900	1.0000	0.9750
Normal, configuration 3 results									
60	5	Norm 3	None	None	0.1200	0.0725	0.1250	0.1175	0.3175
60	5	Norm 3	Bad	Bad	0.2175	0.1525	0.2875	0.2650	0.8925
60	5	Norm 3	Good	Bad	0.7500	0.7275	0.5950	0.5625	0.8725
60	5	Norm 3	Bad	Good	0.7625	0.7550	0.5800	0.5800	0.8925
60	5	Norm 3	Good	Good	0.6825	0.6350	0.8400	0.8525	0.7875
300	5	Norm 3	None	None	0.0450	0.0275	0.0925	0.0825	0.0925
300	5	Norm 3	Bad	Bad	0.4550	0.4425	0.7200	0.7150	1.0000
300	5	Norm 3	Good	Bad	1.0000	1.0000	1.0000	1.0000	1.0000
300	5	Norm 3	Bad	Good	1.0000	1.0000	1.0000	1.0000	1.0000
300	5	Norm 3	Good	Good	1.0000	1.0000	1.0000	1.0000	0.9975
600	5	Norm 3	None	None	0.0600	0.0625	0.1050	0.1025	0.1050
600	5	Norm 3	Bad	Bad	0.8425	0.8475	0.9625	0.9675	1.0000
600	5	Norm 3	Good	Bad	1.0000	1.0000	1.0000	1.0000	1.0000
600	5	Norm 3	Bad	Good	1.0000	1.0000	1.0000	1.0000	1.0000
600	5	Norm 3	Good	Good	1.0000	1.0000	1.0000	1.0000	1.0000
Normal, configuration 4 results									
60	5	Norm 4	None	None	0.1250	0.1525	0.1100	0.1475	0.2950
60	5	Norm 4	Bad	Bad	0.1400	0.2450	0.1250	0.2450	0.5150
60	5	Norm 4	Good	Bad	0.2500	0.5375	0.2125	0.5425	0.4825
60	5	Norm 4	Bad	Good	0.2825	0.5375	0.2400	0.5250	0.5150
60	5	Norm 4	Good	Good	0.4475	0.7400	0.4025	0.7575	0.6375
300	5	Norm 4	None	None	0.0450	0.0950	0.0400	0.0925	0.1200
300	5	Norm 4	Bad	Bad	0.2275	0.6200	0.2375	0.6325	0.8875
300	5	Norm 4	Good	Bad	0.7950	0.9950	0.8025	0.9950	0.8500
300	5	Norm 4	Bad	Good	0.8675	0.9925	0.8575	0.9925	0.8875
300	5	Norm 4	Good	Good	0.9900	1.0000	0.9875	1.0000	0.9925
600	5	Norm 4	None	None	0.0800	0.1250	0.0775	0.1200	0.0450
600	5	Norm 4	Bad	Bad	0.4975	0.8925	0.4825	0.8925	1.0000
600	5	Norm 4	Good	Bad	0.9975	1.0000	0.9975	1.0000	0.9975
600	5	Norm 4	Bad	Good	0.9925	1.0000	0.9925	1.0000	1.0000
600	5	Norm 4	Good	Good	1.0000	1.0000	1.0000	1.0000	1.0000
Normal, configuration 5 results									
60	5	Norm 5	None	None	0.1325	0.0800	0.1000	0.0850	0.3100
60	5	Norm 5	Bad	Bad	0.2900	0.2225	0.2500	0.2250	0.7725
60	5	Norm 5	Good	Bad	0.6550	0.6100	0.6175	0.6100	0.7675
60	5	Norm 5	Bad	Good	0.6350	0.6000	0.6025	0.6075	0.7725
60	5	Norm 5	Good	Good	0.8975	0.8975	0.8725	0.8975	0.9000
300	5	Norm 5	None	None	0.0650	0.0650	0.0700	0.0675	0.1075
300	5	Norm 5	Bad	Bad	0.7850	0.7775	0.7850	0.7900	1.0000
300	5	Norm 5	Good	Bad	1.0000	1.0000	1.0000	1.0000	0.9975

(*Continued*)

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 4. Continued

Sample size	p	Pop. config.	Midpoints separation	Log-Ranges separation	Normal config. 1	Normal config. 3	Normal config. 4	Normal config. 5	Skew Normal config. 1
300	5	Norm 5	Bad	Good	1.0000	1.0000	1.0000	1.0000	1.0000
300	5	Norm 5	Good	Good	1.0000	1.0000	1.0000	1.0000	1.0000
600	5	Norm 5	None	None	0.0550	0.0575	0.0550	0.0550	0.0700
600	5	Norm 5	Bad	Bad	0.9775	0.9775	0.9800	0.9775	1.0000
600	5	Norm 5	Good	Bad	1.0000	1.0000	1.0000	1.0000	1.0000
600	5	Norm 5	Bad	Good	1.0000	1.0000	1.0000	1.0000	1.0000
600	5	Norm 5	Good	Good	1.0000	1.0000	1.0000	1.0000	1.0000
Skew Normal, Configuration 1 results:									
60	5	SkN 1	None	None	0.1225	0.1800	0.1525	0.1650	0.2400
60	5	SkN 1	Bad	Bad	0.1575	0.2050	0.1900	0.2350	0.3050
60	5	SkN 1	Good	Bad	0.2050	0.4025	0.2050	0.3725	0.4325
60	5	SkN 1	Bad	Good	0.2325	0.4325	0.2125	0.4150	0.4050
60	5	SkN 1	Good	Good	0.2425	0.3700	0.3075	0.5500	0.4250
300	5	SkN 1	None	None	0.0575	0.1000	0.1050	0.1300	0.1200
300	5	SkN 1	Bad	Bad	0.1525	0.3500	0.2300	0.4275	0.2525
300	5	SkN 1	Good	Bad	0.6050	0.9000	0.5900	0.8950	0.7500
300	5	SkN 1	Bad	Good	0.6200	0.9000	0.6075	0.9225	0.7400
300	5	SkN 1	Good	Good	0.7800	0.9525	0.8750	0.9875	0.8700
600	5	SkN 1	None	None	0.0600	0.1375	0.1000	0.1275	0.0200
600	5	SkN 1	Bad	Bad	0.2650	0.5625	0.4225	0.7300	0.2900
600	5	SkN 1	Good	Bad	0.9425	1.0000	0.9450	0.9975	0.9725
600	5	SkN 1	Bad	Good	0.9325	0.9975	0.9275	0.9925	0.9900
600	5	SkN 1	Good	Good	0.9875	1.0000	0.9975	1.0000	1.0000

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

We compared the methods according to the following criteria, based on 400 independent replications:

- Size of the test: comparing true significance level with a nominal significance level set at $\alpha = 5\%$.
- Power of the test: frequency of H_0 rejections when groups are not identical.

When there is only one interval variable, only configurations 1 and 5 apply.

The results presented in Tables 3 and 4 show the frequency of null-hypothesis rejections across the 400 replications. In the case of no-separation this percentage is an estimate of the size of the test; in the remaining cases it is an estimate of its power. The analysis of the simulation results in Tables 3 and 4 leads to the following conclusions.

For small sample sizes most methods have unacceptably high values for the test size, with the exception of the method assuming a Normal distribution with configuration 1 when there is only one interval variable, and the methods assuming a Normal distribution with configurations 3 and 5 when the constraints they assume are true.

For medium and large samples, when the data are Normal, the method assuming a Normal distribution with configuration 1 is always reasonable in terms of size (with estimated values varying between 0.0325 and 0.0800) and always better or comparable in terms of power with all the methods for which the estimated size does not differ from the nominal 0.05 by more than 0.03.

When the data follow a non-restricted Skew-Normal distribution the corresponding method produces low estimates of size for medium and large samples when there is only one interval variable and for large samples when there are five interval variables. In all these cases, this method produces better results in terms of power than all other methods with estimated sizes below 0.10, with the exception of the method assuming a Normal distribution with configuration 5 when there is only interval variable and both midpoints and log-ranges are badly separated.

6. Applications

6.1 *Real data: analysis of China temperatures*

In the first application, we study temperatures measured in meteorological stations in northern China. The analysis is based on data consisting of the intervals of observed temperatures (Celsius scale) in each of the four quarters, Q_1–Q_4, of the years 1974–1988 in 22 stations. Table 5 reproduces the original data for some stations and years. The full table comprises $n = 22 \times 15 = 330$ rows and 4 columns. The 22 meteorological stations belong to three different regions in China (North, Northwest, Northeast), which therefore define a partition of the 330 stations-year combinations. A MANOVA is performed to assess whether the regions are different as concerns the observed temperature intervals in the given periods. To control for possible temporal autocorrelation, the global yearly average temperature was subtracted from the corresponding original values.

We performed a preliminary analysis to assess deviations from normality using Q–Q plots and the Kolmogorov–Smirnov goodness-of-fit test. The Q–Q plots did not reveal any strong deviations from normality although for a few variables and classes normality was rejected by the Kolmogorov–Smirnov test, what was to be expected, given the relatively large sample sizes.

Likelihood ratio tests were performed to test multivariate normality against the Skew-Normal distribution for each configuration and the five different considered configurations among themselves for each model (see Table 2). The results are shown in Table 6 and reveal that the most parsimonious models are always rejected against the more general ones, except in the comparision of configurations 2 and 4 for the Skew-Normal model. Therefore, for this data set, there is strong evidence in favour of the non-restricted Skew-Normal model. Given the results obtained (Table 6), the MANOVA analysis should be based on the Skew-Normal model with configuration 1. However, for the sake of completeness, we have decided to present the results for all models. For all the models the global MANOVA results (see Table 7) show unambiguously that the three regions are different. Then separate ANOVA's for each interval variable were performed. In this case, only configurations 1 and 5 apply; both were considered for the Normal and the Skew-Normal models. The analysis shows that regions are distinguishable in terms of each individual interval variable (see Tables 8–11).

6.2 *Illustration: credit card data*

In a second application, we used a data set obtained from survey data included as a sample in the SPSS package (named "customer.dbase"). Among the large set of variables available, we focused on credit card usage variables: *debt to income ratio* ($\times 100$), *credit card debt* (*in thousands*) and *amount spent on primary card last month*. Individual observations have been aggregated on the basis of gender, age category (18–24, 25–34, 35–49, 50–64, more than 65 years old), level of education (did not complete high school, high-school degree, some college, college degree, post-undergraduate degree), and designation of primary credit card (none, gold, platinum, other) leading to 192 groups described by the intervals bounded by the minimum and maximum observed values on the three credit card usage variables.

Table 5. China temperatures interval data.

Station	Region	$Q1$	$Q2$	$Q3$	$Q4$
Beijing-1974	North	[−9.5, 10.6]	[6.5, 29.8]	[12.6, 29.6]	[−10.44, 9.06]
Beijing-1975	North	[−8.6, 12.9]	[7.9, 30.2]	[15.0, 31.6]	[−7.0, 19.2]
...	...	...	...	...	...
ZhangYe-1988	Northwest	[−15.4, 7.2]	[2.3, 26.4]	[8.6, 30.2]	[−12.0, 15.1]

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 6. China temperatures – tests for comparing configurations.

Test	$2\ln\lambda$	DF	P-value
NORM 2–NORM 1	2.992×10^{2}	8	$< 1\times10^{-10}$
NORM 3–NORM 1	1.754×10^{3}	24	$< 1\times10^{-10}$
NORM 3–NORM 2	1.45×10^{3}	16	$< 1\times10^{-10}$
NORM 4–NORM 1	3.217×10^{2}	16	$< 1\times10^{-10}$
NORM 4–NORM 2	2.25×10^{1}	8	0.0040
NORM 5–NORM 1	1.953×10^{3}	28	$< 1\times10^{-10}$
NORM 5–NORM 2	1.65×10^{3}	20	$< 1\times10^{-10}$
NORM 5–NORM 3	1.99×10^{3}	4	$< 1\times10^{-10}$
NORM 5–NORM 4	1.63×10^{3}	12	$< 1\times10^{-10}$
SkN 2–SkN 1	$\mathbf{3.606\times10^{2}}$	**8**	$\mathbf{< 1\times10^{-10}}$
SkN 3–SkN 1	1.783×10^{3}	24	$< 1\times10^{-10}$
SkN 3–SkN 2	1.42×10^{3}	16	$< 1\times10^{-10}$
SkN 4–SkN 1	3.734×10^{2}	16	$< 1\times10^{-10}$
SkN 4–SkN 2	1.27×10^{1}	8	0.1216
SkN 5–SkN 1	1.983×10^{3}	28	$< 1\times10^{-10}$
SkN 5–SkN 2	1.62×10^{3}	20	$< 1\times10^{-10}$
SkN 5–SkN 3	2.01×10^{2}	4	$< 1\times10^{-10}$
SkN 5–SkN 4	1.61×10^{3}	12	$< 1\times10^{-10}$
SkN 1 - NORM 1	$\mathbf{8.273\times10^{1}}$	**8**	$\mathbf{< 1\times10^{-10}}$
SkN 2–NORM 2	2.126×10^{1}	8	0.0065
SkN 3–NORM 3	5.371×10^{1}	8	7.865×10^{-9}
SkN 4–NORM 4	3.108×10^{1}	8	0.0001
SkN 5–NORM 5	5.246×10^{1}	8	1.370×10^{-8}

Table 7. China temperatures – global MANOVA.

Model	$2\ln\lambda$	DF	P-value
NORM 1	480.2475	16	$< 1\times10^{-10}$
NORM 2	527.4521	16	$< 1\times10^{-10}$
NORM 3	989.9340	16	$< 1\times10^{-10}$
NORM 4	529.2541	16	$< 1\times10^{-10}$
NORM 5	1057.9210	16	$< 1\times10^{-10}$
SkN 1	**447.4244**	**16**	$\mathbf{< 1\times10^{-10}}$
SkN 2	518.1720	16	$< 1\times10^{-10}$
SkN 3	974.0240	16	$< 1\times10^{-10}$
SkN 4	530.3980	16	$< 1\times10^{-10}$
SkN 5	1110.3840	16	$< 1\times10^{-10}$

Table 8. China temperatures – ANOVA for the 1st Quarter.

Model	$2\ln\lambda$	DF	P-value
NORM 1	307.9338	4	$< 1\times10^{-10}$
NORM 5	350.8622	4	$< 1\times10^{-10}$
SkN 1	286.6883	4	$< 1\times10^{-10}$
SkN 5	315.4922	4	$< 1\times10^{-10}$

The credit card designation defines a partition in four classes of the gender/age-category/education-level groups. The aim is to investigate whether the three credit card usage variables differ with designation of card, among the considered groups.

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 9. China temperatures – ANOVA for the 2nd Quarter.

Model	$2\ln\lambda$	DF	P-value
NORM 1	203.8863	4	$< 1 \times 10^{-10}$
NORM 5	258.7401	4	$< 1 \times 10^{-10}$
SkN 1	198.0769	4	$< 1 \times 10^{-10}$
SkN 5	252.4611	4	$< 1 \times 10^{-10}$

Table 10. China temperatures – ANOVA for the 3rd Quarter.

Model	$2\ln\lambda$	DF	P-value
NORM 1	115.9361	4	$< 1 \times 10^{-10}$
NORM 5	128.0717	4	$< 1 \times 10^{-10}$
SkN 1	115.4760	4	$< 1 \times 10^{-10}$
SkN 5	127.0646	4	$< 1 \times 10^{-10}$

Table 11. China temperatures – ANOVA for the 4th Quarter.

Model	$2\ln\lambda$	DF	P-value
NORM 1	362.1778	4	$< 1 \times 10^{-10}$
NORM 5	320.2474	4	$< 1 \times 10^{-10}$
SkN 1	306.5288	4	$< 1 \times 10^{-10}$
SkN 5	259.8980	4	$< 1 \times 10^{-10}$

Table 12. Credit card data – tests for comparing configurations.

Test	$2\ln\lambda$	DF	P-value
NORM 2–NORM 1	8.481×10^{1}	8	$< 1 \times 10^{-10}$
NORM 3–NORM 1	1.270×10^{2}	24	$< 1 \times 10^{-10}$
NORM 3–NORM 2	4.21×10^{1}	16	0.0004
NORM 4–NORM 1	7.928×10^{2}	16	$< 1 \times 10^{-10}$
NORM 4–NORM 2	7.08×10^{2}	8	$< 1 \times 10^{-10}$
NORM 5–NORM 1	9.457×10^{2}	28	$< 1 \times 10^{-10}$
NORM 5–NORM 2	8.61×10^{2}	20	$< 1 \times 10^{-10}$
NORM 5–NORM 3	8.19×10^{2}	4	$< 1 \times 10^{-10}$
NORM 5–NORM 4	1.53×10^{2}	12	$< 1 \times 10^{-10}$
SkN 2-SkN 1	$\mathbf{1.089 \times 10^{2}}$	**8**	$\mathbf{< 1 \times 10^{-10}}$
SkN 3–SkN 1	1.309×10^{2}	24	$< 1 \times 10^{-10}$
SkN 3–SkN 2	2.20×10^{1}	16	0.1442
SkN 4–SkN 1	9.871×10^{2}	16	$< 1 \times 10^{-10}$
SkN 4–SkN 2	8.78×10^{2}	8	$< 1 \times 10^{-10}$
SkN 5–SkN 1	1.141×10^{3}	28	$< 1 \times 10^{-10}$
SkN 5–SkN 2	1.03×10^{3}	20	$< 1 \times 10^{-10}$
SkN 5–SkN 3	1.01×10^{3}	4	$< 1 \times 10^{-10}$
SkN 5–SkN 4	1.54×10^{2}	12	$< 1 \times 10^{-10}$
SkN 1–NORM 1	$\mathbf{2.240 \times 10^{2}}$	**8**	$\mathbf{< 1 \times 10^{-10}}$
SkN 2–NORM 2	1.999×10^{2}	8	$< 1 \times 10^{-10}$
SkN 3–NORM 3	2.201×10^{2}	8	$< 1 \times 10^{-10}$
SkN 4–NORM 4	2.967×10^{1}	8	0.0002
SkN 5–NORM 5	2.879×10^{1}	8	0.0003

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Table 13. Credit card data – Global MANOVA.

Model	$2 \ln \lambda$	DF	P-value
NORM 1	15.81713	18	0.605
NORM 2	13.86005	18	0.738
NORM 3	13.92881	18	0.734
NORM 4	18.37095	18	0.431
NORM 5	12.97615	18	0.793
SkN 1	**6.52897**	**18**	**0.994**
SkN 2	10.38462	18	0.919
SkN 3	6.27445	18	0.995
SkN 4	11.42168	18	0.876
SkN 5	11.44587	18	0.874

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Again, preliminary analysis did not reveal any strong deviations from normality, although formal statistical tests did reject normality in some cases.

As in the former example, likelihood ratio tests reveal (see Table 12) that, for this data set, the more parsimonious models are almost always rejected against the more general ones; the Skew-Normal model with configuration 1 is again preferred to the alternatives. MANOVA analysis was performed for all models, results are shown in Table 13. For all the models the global MANOVA results do not show differences among groups.

7. Conclusions and perspectives

In this paper, a parametric modelling for interval data is proposed. We model the midpoints and log-ranges of the interval variables by multivariate Normal or Skew-Normal distributions. The model assumes intrinsic interval data for all variables and cases, i.e., with no degenerate intervals. Mixed situations where degenerate intervals are allowed require different modellings, that we leave for further research.

The intrinsic nature of the interval variables naturally leads to special structures of the variance–covariance matrix, which are represented by five different possible configurations. Maximum likelihood estimation for the different cases is studied. This estimation relies on closed analytical expressions in some cases but requires numerical methods in others. To implement these methods we derive analytical expressions for log-likelihood gradients under constraints. The proposed modelling is then considered in the context of (M)ANOVA testing.

To assess the behaviour of the proposed methodology, a simulation study is performed. This shows that, as long as the sample sizes are not too small, tests have good power and their true significance level approaches nominal levels when the constraints assumed for the model are respected. Applications to meteorological stations in three different regions, based on observed data for 15 years, and to credit card data for different card designations, illustrate the proposed methodology.

Other modellings along the same lines may also be considered. For instance, in [6] models based on the Gamma distribution were proposed. On the other hand, more general distributions than the Skew-Normal may be considered, like the Skew-t, or general skew-elliptical distributions (see [2]).

The framework presented here may now be extended to other statistical methodologies, opening the way to inference approaches for symbolic data.

References

[1] A. Azzalini, *A class of distributions which includes the Normal ones*, Scand. J. Statist. 12 (1985), pp. 171–178.

[2] A. Azzalini, *The Skew-Normal distribution and related multivariate families*, Scand. J. Statist. 32 (2005), pp. 159–188.

[3] A. Azzalini and A. Capitanio, *Statistical applications of the multivariate Skew-Normal distribution*, J. R. Statist. Soc. B 61(3) (1999), pp. 579–602.
[4] A. Azzalini and A. Dalla Valle, *The multivariate Skew-Normal distribution*, Biometrika 83(4) (1996), pp. 715–726.
[5] L. Billard and E. Diday, *From the statistics of data to the statistics of knowledge: Symbolic data analysis*, J. Amer. Statist. Assoc. 98(462) (2003), pp. 470–487.
[6] H.H. Bock, *Probabilistic modeling for symbolic data*, in *COMPSTAT – Proceedings in Computational Statistics*, P. Brito, ed., Springer, Heidelberg, 2008, pp. 55–65.
[7] H.H. Bock and E. Diday, *Analysis of Symbolic Data, Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, Heidelberg, 2000.
[8] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound constrained optimization*, SIAM J. Scient. Comput. 16 (1995), pp. 1190–1208.
[9] J. Case, *Interval arithmetic and analysis*, Coll. Math. J. 30(2) (1999), pp. 106–111.
[10] A. Colubi, *ANCOVA for interval data: A bootstrap approach*, in *Proceedings of ISI-2007*, Lisbon, 2007, CD-Rom.
[11] E. Diday and M. Noirhomme-Fraiture, *Symbolic Data Analysis and the SODAS Software*, Wiley, Chichester, 2008.
[12] M.A. Gil, G. González Rodríguez, A. Colubi, and M. Montenegro, *Testing linear independence in linear models with interval-valued data*, Comput. Statist. Data Anal. 51 (2007), pp. 3002–3015.
[13] D.A. Harville, *Matrix Algebra from a Statistician's Perspective*, Springer, Berlin, Heidelberg, 1999.
[14] V. Krätschmer, *Least-squares estimation in linear regression models with vague concepts*, Fuzzy Sets and Systems 157 (2006), pp. 2579–2592.
[15] R.E. Moore, *Interval Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1966.
[16] D.F. Morrison, *Multivariate Statistical Methods*, 3rd ed., McGraw-Hill, New York, 1990.
[17] G.A.F. Seber, *Multivariate Observations*, Wiley, New York, 1984.

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

Appendix

In this appendix, we give the derivatives used in the computation of the gradient of the log-likelihood function for the Skew-Normal model,

$$l = \text{constant} - \frac{1}{2} n \ln|\Omega| - \frac{n}{2} \operatorname{tr}(\Omega^{-1} V) + \sum_i \zeta_0(\alpha^{\mathrm{t}} \omega^{-1}(x_i - \xi)).$$

These formulae follow from known results concerning matrix differentiation of quadratic and bilinear forms (see [13]).

A.1. Derivatives in the non-restricted case (configuration 1)

A.1.1. Derivatives with respect to the location parameters, $\xi = (\xi_1, \ldots, \xi_{2p})^{\mathrm{t}}$

$$\frac{\partial l}{\partial \xi_j} = -\frac{1}{2} \operatorname{tr}[\Omega^{-1}(M_j + M_j^{\mathrm{t}})] - \sum_{i=1}^{n} \frac{(\alpha_j/\omega_j)\phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))}{\Phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))},$$

where M_j is a $2p \times 2p$ matrix with null entries everywhere except for the jth column, which is equal to $\sum_{i=1}^{n}(\xi - x_i)$.

A.1.2. Derivatives with respect to the skewness parameters, $\alpha = (\alpha_1, \ldots, \alpha_{2p})^{\mathrm{t}}$

$$\frac{\partial l}{\partial \alpha_j} = \sum_{i=1}^{n} \frac{\phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))(1/\omega_j)(x_{ij} - \xi_j))}{\Phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))}.$$

A.1.3. Derivatives with respect to the dispersion parameters, $\Omega(j, j)$, $j = 1, \ldots, 2p$

$$\frac{\partial l}{\partial \Omega(j, j)} = -\frac{n}{2}\Omega^{-1}(j, j) + \frac{n}{2}\operatorname{tr}(\Omega_j^{-1}\Omega_j^{-1\mathrm{t}}V) + \sum_{i=1}^{n}\frac{2\phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))(\alpha_j/\omega_j)(x_{ij} - \xi_j)}{\Phi(\alpha^{\mathrm{t}}\omega^{-1}(x_i - \xi))}$$

where Ω_j^{-1} denotes the jth column of the matrix Ω^{-1}.

A.1.4. Derivatives with respect to the association parameters, $\Omega(j, j')$ $j, j' = 1, \ldots, 2p$, $j \neq j'$

$$\frac{\partial l}{\partial \Omega(j, j')} = -n\Omega^{-1}(j, j') + n\operatorname{tr}(\Omega_j^{-1}\Omega_{j'}^{-1\mathrm{t}}V).$$

A.2. Derivatives in the restricted cases (configurations 2–5)

In this case, the derivatives with respect to the skewness, dispersion and free association parameters are computed using the chain rule and the implicit function theorem, taking into account the restrictions given by Equation (9); the derivatives with respect to the location parameters remain unchanged.

The application of the implicit function theorem requires the differentiation of the covariance

$$\begin{aligned}\Sigma(j, j') = g_{j,j'}(\Omega, \alpha) &= \Omega(j, j') - \frac{2}{\pi}\frac{1}{1 + \alpha^{\mathrm{t}}\omega^{-1}\Omega\omega^{-1}\alpha}\Omega_j^{\mathrm{t}}\omega^{-1}\alpha\alpha^{\mathrm{t}}\omega^{-1}\Omega_{j'} \\ &= \Omega(j, j') - \frac{2}{\pi}A^{-1}B_{j,j'},\end{aligned}$$

where the scalars A and $B_{j,j'}$ are given by

$$A = 1 + \alpha^{\mathrm{t}}\omega^{-1}\Omega\omega^{-1}\alpha, \quad B_{j,j'} = \Omega_j^{\mathrm{t}}\omega^{-1}\alpha\alpha^{\mathrm{t}}\omega^{-1}\Omega_{j'}.$$

A.2.1. Covariance derivatives with respect to the skewness parameters, $\alpha = (\alpha_1, \ldots, \alpha_{2p})^{\mathrm{t}}$

$$\frac{\partial g_{j,j'}}{\partial \alpha_\ell} = \frac{2}{\pi}\frac{1}{\omega_\ell}[2A^{-2}\alpha^{\mathrm{t}}\omega^{-1}\Omega_\ell B_{j,j'} - A^{-1}\alpha^{\mathrm{t}}\omega^{-1}(\Omega(\ell, j)\Omega_{j'} + (\Omega(\ell, j')\Omega_j).$$

A.2.2. Covariance derivatives with respect to the dispersion parameters, $\Omega(\ell, \ell)$

$\ell \neq j, \ell \neq j'$

$$\begin{aligned}\frac{\partial g_{j,j'}}{\partial \Omega(\ell, \ell)} = \frac{2}{\pi}\frac{\alpha_\ell}{\omega_\ell^3}\Bigg[&-A^{-2}\alpha_{(-\ell)}^{\mathrm{t}}\omega_{(-\ell)}^{-1}\Omega_{(-\ell)j}B_{j,j'} + \frac{A^{-1}}{2}(\Omega(\ell, j)\alpha^{\mathrm{t}}\omega^{-1}\Omega_{j'} \\ &+ \Omega(\ell, j')\alpha^{\mathrm{t}}\omega^{-1}\Omega_j)\Bigg],\end{aligned}$$

where $\alpha_{(-\ell)}^{\mathrm{t}}$ and $\Omega_{(-\ell)j}$ denote the corresponding vectors without the ℓth element and likewise $\omega_{(-\ell)}^{-1}$ denotes the matrix ω^{-1} without the ℓth row and column.

$$\frac{\partial g_{j,j'}}{\partial \Omega(j, j)} = \frac{2}{\pi}\frac{\alpha_j}{\omega_j}\left[-A^{-2}\alpha_{(-j)}^{\mathrm{t}}\Omega_{(-j)j}\frac{B_{j,j'}}{\omega_j^2} - \frac{A^{-1}}{2}\alpha_{(-j)}^{\mathrm{t}}\Omega_{(-j)j'} + \frac{A^{-1}}{2}\frac{\Omega(j, j')}{\omega_j^2}\alpha_{(-j)}^{\mathrm{t}}\Omega_{(-j)j}\right].$$

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013

A.2.3. Covariance derivatives with respect to the association parameters, $\Omega(\ell, \ell')$

(i) $\ell, \ell' \neq j, j'$:

$$\frac{\partial g_{j,j'}}{\partial \Omega(\ell, \ell')} = \frac{2}{\pi} \frac{\alpha_\ell}{\omega_\ell} \frac{\alpha_{\ell'}}{\omega_{\ell'}} A^{-2} B_{j,j'}.$$

(ii) $\ell = j, \ell' \neq j, j'$:

$$\frac{\partial g_{j,j'}}{\partial \Omega(j, \ell')} = \frac{2}{\pi} \frac{\alpha_{\ell'}}{\omega_{\ell'}} \left[2A^{-2} \frac{\alpha_j}{\omega_j} B_{j,j'} - A^{-1} \alpha^{\mathrm{t}} \omega^{-1} \Omega_{j'} \right].$$

(iii) $\ell = j, \ell' = j', j \neq j'$:

$$\frac{\partial g_{j,j'}}{\partial \Omega(j, j')} = 1 + \frac{2}{\pi} \left[2A^{-2} \frac{\alpha_j}{\omega_j} \frac{\alpha'_j}{\omega'_j} B_{j,j'} - A^{-1} \frac{\alpha_j}{\omega_j} \alpha^{\mathrm{t}} \omega^{-1} \Omega_j - \frac{\alpha'_j}{\omega'_j} \alpha^{\mathrm{t}} \omega^{-1} \Omega_{j'} \right].$$

Downloaded by [Knihovna Univerzity Palackeho] at 06:06 06 May 2013