Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013)
International Conference, Mataró (Barcelona), Spain 25 - 28 June 2013

SGR modeling of fake ordinal data with correlational structures

Luigi Lombardi[1], Massimiliano Pastore[2],
Massimo Nucci[3], and Andrea Bobbio[4]

[1] Dipartimento di Psicologia e Scienze Cognitive, Università di Trento,
Corso Bettini, 31, I-38068 Rovereto (TN), Italy
(E-mail: luigi.lombardi@unitn.it)
[2] Dipatimento di Psicologia dello Sviluppo e della Socializzazione,
Università di Padova, via Venezia, 8, I-35131 Padova, Italy
(E-mail: massimiliano.pastore@unipd.it)
[3] Dipartimento di Psicologia Generale, Università di Padova,
via Venezia, 8, I-35131 Padova, Italy
(E-mail: massimo.nucci@unipd.it)
[4] Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata,
Università di Padova, via Venezia, 8, I-35131 Padova, Italy
(E-mail: andrea.bobbio@unipd.it)

Abstract. In many psychological inventories (i.e., personnel selection surveys and diagnostic tests) the collected samples often include fraudulent records. This confronts the researcher with the crucial problem of biases yielded by the usage of standard statistical models. In this paper we generalize a recent probabilistic perturbation procedure, called SGR - Sample Generation by Replacements - (Lombardi & Pastore[4]), to simulate fake data with correlational structures. To mimic these more complex faking data we proposed a novel extension of the SGR conditional replacement distribution which is based on a discrete version of the truncated multivariate normal distribution. We also applied the new procedure to real behavioral data on the role of perceived affective self-efficacy in social contexts.
Keywords: Sample Generation by Replacement, Fake-good data, Truncated multivariate normal distribution.

Many self-report measures of attitudes, beliefs, personality, and pathology include items that can be easily manipulated by respondents. For example, an individual may deliberately attempt to manipulate or distort responses to personality inventories and attitude tests to create positive impressions (e.g., Paulhus[7]; Zickar & Robie[8]). In other circumstances, some individuals may tend to malinger responses on a symptom checklist to simulate grossly exaggerated physical or psychological symptoms in order to reach specific goals such as, for example, obtaining financial compensation, avoiding being charged with a crime, avoiding military duty, or obtaining drugs (e.g., Hall & Hall[2]).

Sample Generation by Replacement (SGR) is a recent data simulation procedure to artificially generate samples of fake ordinal data (Lombardi & Pastore[4]). SGR is based on a two-stage sampling algorithm characterized by two distinct generative models: the model representing the process that generates the data prior to any fake perturbation (data generation process) and the

model representing the faking process to perturb the data (data replacement process). This approach has been recently applied to evaluate the impact of hypothetical faking good manipulations in ordinal data on the performances of a set of widely known and commonly used stand-alone SEM-based fit indices (Lombardi & Pastore[4]). SGR has also been used to study the sensitivity of reliability indices to fake perturbations in dichotomous and ordered data generated by factorial models (Pastore & Lombardi[6]).

Up to now the SGR approach has been limited to the modeling of conditionally independent fake data. In this contribution, we propose a novel generalization of the conditional replacement distribution that accounts for possible correlated structures in the simulated fake data.

1 Main features of the SGR approach

1.1 Data generation process

We think of the original data as being represented by an $n \times m$ matrix $\mathbf{D}$, that is to say, n observations (e.g., participants) each containing m elements (e.g., participant's responses). We assume that entry d_{ij} of $\mathbf{D}$ ($i = 1, \ldots, n; j = 1, \ldots, m$) takes values on a small ordinal range $1, 2, \ldots, Q$. In particular, let $\mathbf{d}_i$ be the $(1 \times m)$ array of $\mathbf{D}$ denoting the simulated pattern of responses of participant i. The response pattern $\mathbf{d}_i$ is a multidimensional ordinal random variable with probability distribution $p(\mathbf{d}_i|\theta)$, where θ indicates the vector of parameters of the generative probabilistic model of the data. Moreover, we assume that the simulated response patterns are independent and identically distributed (i.i.d.) observations.

1.2 Data replacement process

The basic principle of the SGR approach is to generate a new $n \times m$ ordinal data matrix $\mathbf{F}$, called the *fake data matrix* of $\mathbf{D}$, by manipulating each element d_{ij} in $\mathbf{D}$ according to a replacement probability distribution. Let $\mathbf{f}_i$ be the $(1 \times m)$ array of $\mathbf{F}$ denoting the hypothetical pattern of fake responses of participant i. The fake response pattern $\mathbf{f}_i$ is a multidimensional ordinal random variable with conditional replacement probability distribution $p(\mathbf{f}_i|\mathbf{d}_i, \theta_F)$ where θ_F indicates the vector of parameters of the probabilistic faking model. In general, θ_F represents hypothetical a priori knowledge about the distribution of faking (e.g., the chance of observing a fake observation in the data) or empirically based knowledge about the process of faking (e.g., the direction of faking - fake good vs. fake bad -). In the SGR framework the replacement distribution $p(\mathbf{f}_i|\mathbf{d}_i, \theta_F)$ is restricted to satisfy a *conditional independence assumption* (see Lombardi & Pastore[4]; Pastore & Lombardi[6]). More precisely, in the replacement distribution each fake response f_{ij} only depends on the corresponding data observation d_{ij} and the model parameter θ_F. Therefore, because the patterns of fake responses are also i.i.d. observations, the

simulated data array $(\mathbf{D}, \mathbf{F})$ is drawn from the joint probability distribution

$$p(\mathbf{D}, \mathbf{F}|\theta, \theta_F) = \prod_{i=1}^{n} p(\mathbf{d}_i|\theta) p(\mathbf{f}_i|\mathbf{d}_i, \theta_F) \quad (1)$$

$$= \prod_{i=1}^{n} p(\mathbf{d}_i|\theta) \prod_{j=1}^{m} p(f_{ij}|d_{ij}, \theta_F) \quad (2)$$

By repeatedly sampling data from the two generative models we can simulate the so called *fake data sample* (FDS). We can then study the distribution of some relevant statistics computed on this FDS.

1.3 The problem of the independence assumption

We recall that the SGR simulation procedure generates fake perturbations that are restricted to satisfy the conditional independence assumption. Unfortunately, this restriction clearly limits the range of empirical faking processes that can be mimicked by the SGR simulation procedure. In particular, because the replacement distribution acts as a perturbation process for the original data, the resulting fake data will always show correlations that are (on average) weaker than the ones observed for the original data, thus showing a sort of residual correlation effect. However, it is known that some empirical contexts may require different model assumptions about the faking process that cannot be captured by this simple framework. For example, different modulations of graded faking such as slight faking and extreme faking (e.g., Zickar & Robie[8]) are not consistent with the simple independence hypothesis.

To fill this gap, in this contribution we propose a novel conditional replacement distribution that allows to modulate different levels of correlational patterns in the simulated fake data. Because our new proposal is based on a discrete version of the truncated multivariate normal distribution, in what follows we present some relevant properties of this important distribution function before introducing the new perturbation model that does not hinge on the independence assumption.

2 A SGR framework for correlated fake-good data

The SGR approach offers an elegant way to simulate faking good scenarios. In general, faking good can be conceptualized as an individual's deliberate attempt to manipulate or distort responses to create a positive impression (Paulhus[7]; Zickar & Robie[8]). Notice that, the faking good (as well as the faking bad) scenario always entails a conditional replacement model in which the conditioning is a function of response polarity. This model represents a perturbation context in which responses are exclusively subject to positive feigning:

$$f_{ij} \geq d_{ij}; \qquad i = 1, \ldots, n; j = 1, \ldots, m$$

2.1 The truncated multivariate replacement distribution

As a kernel for the conditional replacement distribution we consider the truncated multivariate normal distribution $TN(\mu, \mathbf{\Sigma}, \mathbf{a}, \mathbf{b})$ (e.g., Horrace[3]). This distribution can be expressed as

$$f(\mathbf{x}|\mu, \mathbf{\Sigma}, \mathbf{a}, \mathbf{b}) = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\}}{\int_{\mathbf{a}}^{\mathbf{b}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right\} d\mathbf{x}} \tag{3}$$

for $\mathbf{a} \leq \mathbf{x} \leq \mathbf{b}$ and 0 otherwise. The $(1 \times m)$ vectors $\mathbf{a}$ and $\mathbf{b}$ are the lower and upper truncation points ($a_j < b_j$; $j = 1, \ldots, m$) for the multivariate normal distribution with m dimensions. Finally, μ and $\mathbf{\Sigma}$ are the location parameter vector and the covariance matrix of the (not truncated) multivariate normal distribution.

Now, let $\mathbf{f}_i = (k_1, \ldots, k_m)$ and $\mathbf{d}_i = (h_1, \ldots, h_m)$ be the replaced values and the original values for the i^{th} simulated observation, respectively. According to the Underlying Variable Approach (UVA; Muthén[5]) we can set

$$p(\mathbf{f}_i|\mathbf{d}_i, \theta_F) = \begin{cases} \int_{\alpha_{k_1-1}}^{\alpha_{k_1}} \cdots \int_{\alpha_{k_m-1}}^{\alpha_{k_m}} f(\mathbf{x}|\mathbf{0}, \mathbf{\Sigma}, \mathbf{a}^i, \mathbf{b}^i) d\mathbf{x}, & \forall j \ : \ 1 \leq h_j \leq k_j \leq Q \\ 0, & \exists j \ : \ k_j < h_j \end{cases}$$

In the replacement distribution $\mathbf{0}$ is the $1 \times m$ array of zeros representing the location parameter (that is to say $\mu = \mathbf{0}$), whereas the pair $(\alpha_{k_j-1}, \alpha_{k_j})$ are the thresholds corresponding to the discrete value k_j ($j = 1, \ldots, m$). Following the UVA framework we assume that there exists a continuous data matrix $\mathbf{F}^*$ underlying the fake ordinal data matrix $\mathbf{F}$. The connection between the ordinal variable f_{ij} and the underlying variable f_{ij}^* in $\mathbf{F}^*$ is given by

$$f_{ij} = k_j \quad \Leftrightarrow \quad \alpha_{k_j-1} < f_{ij}^* < \alpha_{k_j}; \quad i = 1, \ldots, n; j = 1, \ldots, m.$$

Recall that to represent an ordinal item with Q categories we need $Q + 1$ thresholds. Finally, the bounds $\mathbf{a}^i = (a_1^i, \ldots, a_m^i)$ and $\mathbf{b}^i = (b_1^i, \ldots, b_m^i)$ are set to

$$a_j^i = \alpha_{h_j-1} \qquad b_j^i = +\infty, \qquad j = 1, \ldots, m$$

where we recall that $(\alpha_{h_j-1}, \alpha_{h_j})$ is the pair of thresholds corresponding to the value h_j for the original variable d_{ij} in $\mathbf{d}_i$. Figure 1 shows an example for a one dimensional case. In sum, the parameter array for the faking model is given by

$$\theta_F = (\alpha, \mathbf{\Sigma}).$$

with α being the $m \times (Q-1)$ threshold matrix.

2.2 Some relevant properties of the new replacement distribution

It is important to point out two interesting properties of the novel replacement distribution for correlated fake data.

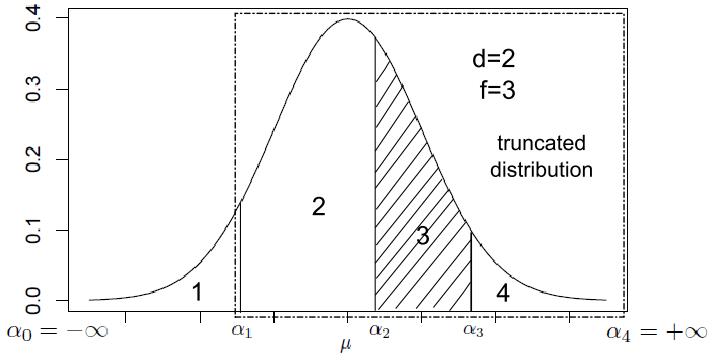

Fig. 1. Example of a truncated normal distribution (one dimensional) with $d = 2$ and $f = 3$. The mean μ of the distribution is 0, the bounds a and b are α_1 and $+\infty$, respectively. The conditional probability $p(3|2, \theta_F)$ is the shaded area between α_2 and α_3.

The first property is related to the probability of replacement π. According to the truncated multivariate replacement distribution the value $1 - \pi_i$ of the conditional probability of non-replacement

$$\mathbf{f}_i = \mathbf{d}_i = (h_1, \ldots, h_m)$$

is equivalent to

$$1 - \pi_i = \int_{\alpha_{h_1-1}}^{\alpha_{h_1}} \cdots \int_{\alpha_{h_m-1}}^{\alpha_{h_m}} f(\mathbf{x}|\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{a}^i, \mathbf{b}^i) d\mathbf{x} \tag{4}$$

and consequently, the probability of replacement π_i is

$$\pi_i = 1 - \left(\int_{\alpha_{h_1-1}}^{\alpha_{h_1}} \cdots \int_{\alpha_{h_m-1}}^{\alpha_{h_m}} f(\mathbf{x}|\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{a}^i, \mathbf{b}^i) d\mathbf{x} \right) \tag{5}$$

Note that in the original model of replacement (Lombardi & Pastore[6]) the probability of replacement π was an explicit parameter in the replacement distribution, whereas in this new proposal π is an implicit parameter that can be derived by taking the average across the n distincts π_i:

$$\bar{\pi} = \frac{1}{n} \sum_{i=1}^{n} \pi_i.$$

The second property is related to the correlational structure of the simulated fake data set $\mathbf{F}$. Unlike the standard model of replacement, in the new configuration we can directly represent correlations between the replaced values for the m ordinal variables. In particular, the correlation matrix $\mathbf{R}_f$ of $\mathbf{F}$ can be modulated by the covariance matrix $\boldsymbol{\Sigma}$ in the replacement model. Note, however, that $\boldsymbol{\Sigma}$ is the covariance matrix of the original (not truncated) multivariate normal distribution. In general, the computation of the first and

second moments is not trivial for the truncated case, since they are obviously not the same as μ and $\mathbf{\Sigma}$ from the parametrization of $TN(\mathbf{0}, \mathbf{\Sigma}, \mathbf{a}, \mathbf{b})$. In particular, it can be seen that truncation can significantly reduce the variance and change the covariance between variables. In the next section we will show some examples of how the resulting correlation matrix $\mathbf{R}_f$ can be affected from the interaction between different modulations of faking (represented by different configurations of threshold values α) and the structure of the covariance matrix $\mathbf{\Sigma}$.

3 Applicative example

The new replacement distribution is illustrated using data from a questionnaire about the role of perceived affective self-efficacy in personality and social psychology (Bandura, Caprara, Barbaranelli, Gerbino, and Pastorelli[1]). Participants were 463 undergraduate students (389 females) at the University of Padua (Italy). Ages ranged from 18 to 48, with a mean of 20.64 and a standard deviation of 2.71. Data consisted of the participants' responses to three of the 12 items of the AEP/A scale (Caprara, 2001) scored on a 4-point agree-disagree scale (value 1 denotes that a participant totally disagrees with the statement, whereas value 4 means total agreement with the statement). However, the 463 participants were divided into two groups. The first group ($n_1 = 231$) received a neutral set of instructions, whereas the second group ($n_2 = 232$) received ad lib faking instructions. The resulting responses were collected into two empirical data matrices: $\mathbf{D}_e$ for the neutral group and $\mathbf{F}_e$ for the ad lib faking group. As expected the observed responses for the ad lib faking group were affected by fake good observations. More precisely, the participants deliberately manipulated their responses using larger values of the scale to create better impressions. This hypothesis was partially supported by the moderate ceiling effects observed in the data (see Fig. 2). Because no additional items on social desirability was available in the AEP/A inventory, we decided to perform an SGR analysis on the basis of two hypothetical scenarios: slight faking and uniform faking.

3.1 Comparing faking models

An SGR analysis was used to evaluate the mimicking ability of four different faking models with respect to the empirical fake set $\mathbf{F}_e$. In particular, because $\mathbf{D}_e$ and $\mathbf{F}_e$ contained responses collected from two different groups of individuals, we decided to evaluate the simulated fake samples against the empirical marginal means of the three items as well as against their empirical correlations in $\mathbf{F}_e$. To that end, we defined four perturbation models derived by the combination of two factors with two levels each. The first factor defined two structures for the covariance matrix in the truncated replacement distribution: a) the identity matrix $\mathbf{I}$ (denoting an independent model) b) the empirical correlation matrix of $\mathbf{R}_e$ computed on the observed matrix $\mathbf{F}_e$ (representing the correlational structure in $\mathbf{F}_e$). The second factor represented

two different options for the thresholds in the replacement distribution: a) $\alpha_1 = -0.674, \alpha_2 = 0.0, \alpha_3 = 0.674$ denoting a *uniform support fake-good distribution* (Lombardi and Pastore[4]) b) $\alpha_1 = -1.591, \alpha_2 = 1.596, \alpha_3 = 3.332$ denoting a *slight fake-good distribution* (e.g., Zickar & Robie[8]). All the other components (parameters' values) were set to identical values in all the four faking models. In particular, the original data set $\mathbf{D}$ in the conditional replacement distribution was set equal to the empirical data set $\mathbf{D}_e$. The latter means that in this application we did not simulate the data $\mathbf{D}$, instead we directly used the observed data $\mathbf{D}_e$ in the replacement distribution equation. The four faking models were then used to simulated new FDSs.

The results of the SGR analysis are shown in Figures 2,3. Figure 2 represents the simulated marginal means of the fake-good data as a function of the two factors. The results showed that the slight faking model with mild dependency was preferred to the other three models. Similarly, figure 3 shows the simulated correlations of the fake-good data as a function of the two factors. Like for the marginal means, also for the correlations the slight faking model with mild dependency showed a better performance as compared to the other three models. In other words, the ad lib faking instructions seem to be more consistent with a mild *positive impression effect* that boosts the correlations between the items.

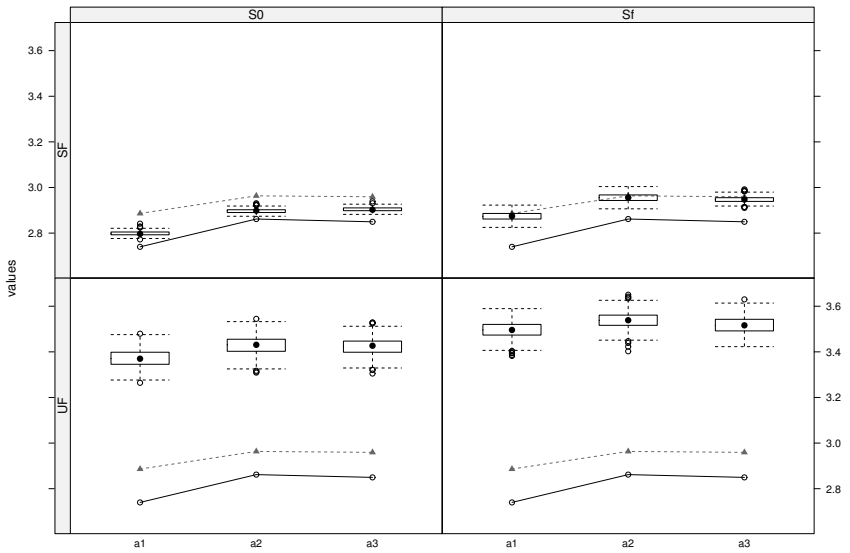

Fig. 2. Boxplots for the simulated marginal means of the fake-good data for the four models. The solid line denotes the observed pattern for the marginal means in $\mathbf{D}_e$. The dashed line indicates the observed pattern for the marginal means in $\mathbf{F}_e$. UF = uniform support fake-good distribution, SL = slight fake-good distribution; S0 = independent faking model, Sf = faking model with a correlational structure. a_1, a_2, and a_3 denote the three selected items of the AEP/A scale. The data represented in each boxplot were derived from 500 FDSs.

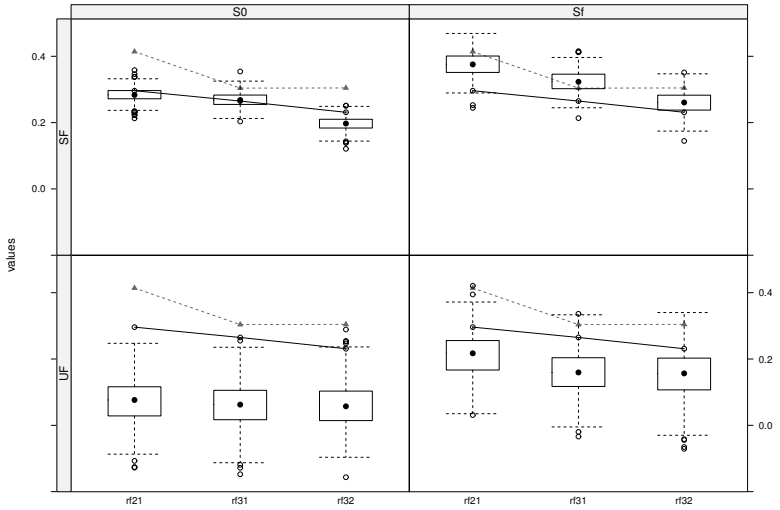

Fig. 3. Boxplots for the simulated correlations of the fake-good data for the four models. The solid line denotes the observed correlational pattern in $\mathbf{D}_e$. The dashed line indicates the observed correlational pattern in $\mathbf{F}_e$. UF = uniform support fake-good distribution, SL = slight fake-good distribution; S0 = independent faking model, Sf = faking model with a correlational structure. The variable $\text{rf}_{jj'}$ denotes the correlation between item a_j and item a'_j of the AEP/A scale. The data shown in each boxplot were derived from 500 FDSs.

References

1.A. Bandura, G.V. Caprara, C. Barbaranelli, M. Gerbino and C. Pastorelli. Role of affective and self-regulatory efficacy in diverse spheres of psychosocial functioning. *Child Development*, 74, 3, 769–782, 2003.
2.R. C. Hall and R. C. Hall. Detection of malingered PTSD: An overview of clinical, psychometric, and physiological assessment: Where do we stand? *Journal of Forensic Science*, 52, 717–725, 2007.
3.W.C. Horrace. Some results on the multivariate truncated normal distribution. *Journal of Multivariate Analysis*, 94, 209–221, 2005.
4.L. Lombardi and M. Pastore. Sensitivity of fit indices to fake perturbation of ordinal data: A sample by replacement approach. *Multivariate Behavioral Research*, 47, 519–546, 2012.
5.B. Muthén. A general structural equation model with dichotomous, ordered categorical and continuous latent variables indicators. *Psychometrika*, 49, 115–132, 1984.
6.M. Pastore and L. Lombardi. The impact of faking on Cronbach's Alpha for dichotomous and ordered rating scores. submitted paper, 2012.
7.D. L. Paulhus. Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609, 1984.
8.M. J. Zickar and C. Robie. Modeling faking good on personality items: An item-level analysis. *Journal of Applied Psychology*, 84, 551–563, 1999.