A relationship between Likert-type responses and hypothetical latent variables

Seminar for Computational Statistical Methods

Marko Lalović

November 2018

Abstract

Researchers frequently collect responses to questionnaires in the Likert item format. A *Likert item* is a statement that the respondent is asked to evaluate by giving it a quantitative value on some dimension, with level of agreement being the dimension most commonly used. Here the *agreement* is defined as an underlying continuum that is unobserved and ranges from complete disagreement to complete agreement with the statement. Although researchers may be interested in responses for their own sake, they are often more interested in the analysis of the underlying agreement. Therefore the methods of analysis must involve a model for the relationship between Likert-type responses and the attitudes underlying them.

Accordingly, we introduce a model with which we remain in the context of the assumptions of the classical theory of the treatment of Likert-type responses. The response to a question is the observed *manifest variable*, denoted with $\hat{X}$, and we adopt the assumption that there exist a corresponding *latent variable*, denoted with X. Another assumption is that all questions are equivalent instruments for measuring the agreement.

We investigate the relationship between the variables X and $\hat{X}$, particularly the optimal relationship in the ideal case. We derive a mapping from X to $\hat{X}$, called *discretization*. Using optimal discretization and skew-normal distribution we show how we can simulate asymmetrically distributed Likert-type responses while maintaining the relationship between variables with nice mathematical properties.

We also derive a mapping from $\hat{X}$ to X, called *reconstruction* that avoids arbitrary mapping of responses to numbers and the result allows the use of classical statistical methods. We investigate the simple problem of comparison of two means. We show that the common testing procedure is biased in general, since it doesn't take into account the discrete nature of Likert-type responses. Our proposed reconstruction procedure is based on the proportion of responses, as this provides minimal sufficient statistic for this problem. We illustrate our proposed testing procedure on a concrete example.

Keywords: Likert, Likert scale, Nonlinearity, Latent variable models, Statistical methodology

Contents

1 Introduction

The diverse research community recognizes that the classical theory of analyzing the Likert-type responses requires improvement [1]. The problems associated with the classical approach to analyzing Likert-type responses have already been discussed [2]. Several alternative approaches have been proposed, even based on a methodology developed by quantum physicist [3], [4]. Our approach may be closer to item response theory [5], [6], but we did not follow this theory and remained in the context of classical theory. Similar way of discretization to simulate Likert-type responses was used in [7]. They considered an additional condition and proposed a different algorithm that also allows for biased discretization. In the context of ordinal variables, optimal scaling using the nonlinear principal components method (NLPCA) is described in [8] and [9]. The introduction to NLPCA is found in [10]. We also do not assume a normal distribution of agreement in our model and instead model it as a skew-normal distribution similar to [11], which also includes a normal distribution. Rice, for example, stated [12] as early as 1938, that there is no obvious a priori reason for assuming a normal distribution when modeling agreement.

In Section 2 we introduce a model where the distribution of agreement is described with a continuous random variable X and the distribution of Likert-type responses with a discrete random variable $\hat{X}$. We are mainly trying to solve the following problem. We are interested in the relationship between the variables X and $\hat{X}$. Our goal is to find the optimal mapping $\mathcal{D}$:

$$\mathcal{D} : X \to \hat{X} \tag{1}$$

and the optimal mapping $\mathcal{R}$:

$$\mathcal{R} : \hat{X} \to X. \tag{2}$$

This problem can also be described in the language of NLPCA as follows. Find the mappings $\mathcal{D}$ and $\mathcal{R}$ such that:

$$\hat{X} = (\mathcal{R} \circ \mathcal{D})(X) + \epsilon, \tag{3}$$

and minimize the expected value of the error $\left\|\epsilon^2\right\|$. If we were to solve this problem using a neural network, we call such mappings *auto-associative neural network* or simply *autoencoder*. The optimal discretization is closely related to optimal scaling. Here we solve this problem using numerical methods, using the Lloyd-Max algorithm for optimal discretization and the adaptive Gauss-Newton method for optimal reconstruction.

The mapping $\mathcal{D}$ is called *discretization*. We describe it in Section 3 and show how we can use it to simulate Likert-type responses in Section 4. The mapping $\mathcal{R}$ is called *reconstruction*. A procedure to reconstruct latent variables from Likert-type responses is in Section 5. In Section 6 we investigate the simple problem of comparison of two means in the context of our model. We first present a common approach, then we show it is biased in general. And finally illustrate our proposed testing procedure on a concrete example. We do not suggest using this approach in the case of small samples, as it fails on very small samples. In addition, we did not perform a theoretical analysis of the impact of sampling.

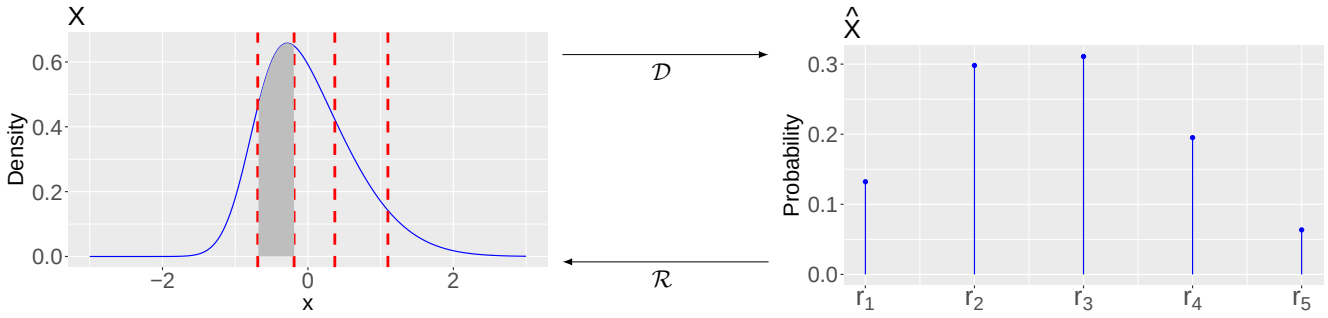


Figure 1: Latent variable $X \sim \mathcal{SN}(-0.76, 1, 3)$ and the corresponding manifest variable $\hat{X}$.

2 Model

The Likert-type response is the observed discrete random variable we call *manifest variable* and denote with $\hat{X}$. Suppose that there are more than two distinct possible responses: $K > 2$, this means $\hat{X}$ is a *polytomous variable*. We denote the possible responses by r_k, $k = 1, ..., K$ and denote the probability that the value of $\hat{X}$ is r_k with p_k. The distribution of the manifest variable $\hat{X}$ can be represented by the Table 1, where in the first row are the values r_k on which $\hat{X}$ is defined and in the second row are the probabilities p_k that the value of $\hat{X}$ is r_k, for $k = 1, ..., K$.

r_k	r_1	r_2	...	r_{K-1}	r_K
$\mathrm{P}(\hat{X} = r_k)$	p_1	p_2	...	p_{K-1}	p_K

Table 1: Distribution of a manifest variable $\hat{X}$.

We adopt the generally accepted assumption [13] that for each observed manifest variable $\hat{X}$ there exist a corresponding underlying continuum which lies in the minds of the subjects. We model this continuum with a continuous random variable we call *latent variable* and denote with X. Suppose the density of latent variable X is defined on the set of real numbers and denote this density f_X.

Suppose the relation between X and $\hat{X}$ is monotone and define the relation similarly to [7]:

$$\hat{X} = r_k, \qquad \text{if} \quad x_k < X \leq x_{k+1} \quad k = 1, ..., K+1, \tag{4}$$

where the real numbers x_k are called *thresholds*. The thresholds are defined in the domain of the corresponding latent variable X and satisfy the constraints:

$$-\infty = x_1 < x_2 < ... < x_K < x_{K+1} = \infty. \tag{5}$$

For given thresholds x_k, $k = 1, ..., K+1$ denote the intervals $\mathcal{R}_k$:

$$\mathcal{R}_1 = (-\infty, x_2], \quad ..., \quad \mathcal{R}_k = (x_k, x_{k+1}], \quad ..., \quad \mathcal{R}_K = (x_K, \infty) \tag{6}$$

into which the thresholds x_k partition the domain of the latent variable X.

The possible responses r_k are called *representatives* of the intervals $\mathcal{R}_k$. The following relationship follows from the definition:

$$p_k = \mathrm{P}(\hat{X} = r_k) = f_X(x \in \mathcal{R}_k), \quad k = 1, ..., K. \tag{7}$$

We say that the random variable $\hat{X}$ is a *discretization* of the random variable X and denote the mapping:

$$\mathcal{D} : X \to \hat{X}. \tag{8}$$

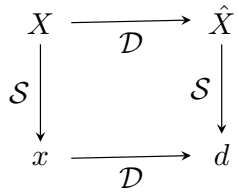
We say that the random variable X is a *reconstruction* of the random variable $\hat{X}$ and denote the mapping:

$$\mathcal{R} : \hat{X} \to X. \tag{9}$$

2.1 Sampling

By sampling X and then discretizing the sample, we get the same result as by discretizing X and then sampling $\hat{X}$.

Proposition 1. *Denote by $\mathcal{S}$ the sampling from the random variable X or $\hat{X}$. Denote by $\mathcal{D}$ the discretization of the continuous random variable X or the sample $\mathbf{x}$ from the continuous random variable X. We claim that the diagram:*



is commutative. Or in other words, this equation holds:

$$\mathcal{D}(\mathcal{S}(X)) = \mathcal{S}(\mathcal{D}(X)).$$

Proof. Let $\mathbf{d} = \{d_i\}$ be the sample obtained by first sampling X to obtain a sample $\mathbf{x} = \{x_i\}$ and then discretizing the sample $\mathbf{x}$ using thresholds x_k, $k \in \{1, ..., K+1\}$, as defined by Eq. 4. For $k \in \{1, ..., K\}$ it holds:

$$r_k \in d, \qquad \text{if there exists} \quad x_i \in \mathbf{x}, \quad \text{such that} \quad x_k < x_i \leq x_{k+1} \tag{10}$$

Or in other words, let

$$\mathbf{d} = \mathcal{D}(\mathcal{S}(X)).$$

The probability that r_k is in sample $\mathbf{d}$ is

$$\mathrm{P}(x_k < X \leq x_{k+1})$$

which is exactly p_k in the definition of $\hat{X} = \mathcal{D}(X)$. □

Let us also determine the distribution of the sample $\mathbf{d}$ of size n. Denote by b_k the number of responses with the value r_k in the sample $\mathbf{d}$. The sample $\mathbf{d}$ has a multinomial distribution:

$$\mathrm{P}(\{b_k \mid k = 1, ..., K\}) = \binom{n}{b_1, ..., b_K} \prod_{k=1}^{K} p_k^{b_k}, \tag{11}$$

for $b_k \in \{0, ..., n\}$ and where $\sum_{k=1}^{K} b_k = n$.

2.2 Skew normal distribution

We model the latent variable X, similar to [11], by skew-normal distribution $\mathcal{SN}(\xi, \omega, \alpha)$. Where ξ is the location parameter, ω is the scaling parameter, and α is asymmetry parameter. We present the skew-normal distribution in more detail.

The random variable X has a *standard skew-normal distribution* [14] with parameter α, if it's probability density function is:

$$f(x, \alpha) = 2 \cdot \phi(x) \cdot \Phi(\alpha x), \tag{12}$$

where α and x are real numbers, $\phi(\cdot)$ density, and $\Phi(\cdot)$ cumulative distribution function of the standard normal distribution. Denote by $X \sim \mathcal{SN}(\alpha)$. It is easy to check:

- When $\alpha = 0$ asymmetry disappears and we get the density of the normal distribution.
- The distribution is left asymmetric when $\alpha > 0$ and right asymmetric when $\alpha < 0$.
- As $|\alpha|$ increases, the asymmetry increases and converges to the so-called *half normal density*.

Let $X \sim \mathcal{SN}(\alpha)$ and use the linear transformation $Y = \omega X + \xi$, where ξ is called the *location parameter* and ω *scaling parameter*. Then Y is distributed *skew-normal* with density:

$$f_Y(y; \xi, \omega, \alpha) = \frac{2}{\omega} \cdot \phi\left(\frac{y-\xi}{\omega}\right) \cdot \Phi\left(\alpha\left(\frac{y-\xi}{\omega}\right)\right), \tag{13}$$

where ξ is a real number and ω is a positive real number. Denote by $Y \sim \mathcal{SN}(\xi, \omega, \alpha)$.

We assume the latent variable has a skew-normal distribution:

$$X \sim \mathcal{SN}(\xi, \omega, \alpha) \tag{14}$$

for some parameter values ξ, ω and α.

In Figure 1 there is an example of the latent variable $X \sim \mathcal{SN}(-0.76, 1, 3)$ and the corresponding manifest random variable $\hat{X}$ for $K = 5$. The thresholds x_k, $k = 2, ..., 5$ are drawn in dashes and the shaded area equals to the probability:

$$p_2 = \mathrm{P}(\hat{X} = r_2) = f_X(x \in \mathcal{R}_2).$$

3 Discretization

In general the discretization problem can be described as follows. For a given K, we wish to find a tessellation, which determines the partition of the continuous continuum of the domain of the continuous random variable X into K subsets on which we can define a probability distribution and obtain a discrete random variable $\hat{X}$ defined on some set r_k, $k = 1, ..., K$. In our case, where the domain of X is one-dimensional set of real numbers, we wish to find the thresholds:

$$x_k, \quad k = 1, ..., K+1,$$

which divide the interval on which X is defined into K subintervals $\mathcal{R}_k$, on which we can define the probability distribution. At the same time, we require that the thresholds satisfy the Condition 5, so we can already set the edge thresholds to:

$$x_1 = -\infty, x_{K+1} = \infty.$$

3.1 Equal width discretization

One of the most commonly used [15] techniques of discretization is the use of equally wide intervals, the so-called 'Equal Width Interval Binning' (EW) discretization. The reason why this method is so popular is probably that it is easy to understand and the implementation is trivial.

Description of the equal width discretization procedure. In the case where we have a continuous random variable X defined on an unbounded interval, we have to select a subinterval on which the majority of the density f_X is defined. The popular choice is the interval $[a, b]$ for the random variable X, where the boundaries are:

$$a = \mathrm{E}[X] - 3 \cdot \mathrm{sd}(X),$$
$$b = \mathrm{E}[X] + 3 \cdot \mathrm{sd}(X).$$

In the case of $X \sim N(0, 1)$, the interval is $[-3, 3]$. Then we divide the interval $[a, b]$ into K equally wide intervals with thresholds:

$$x_2, \quad x_3, \quad ..., \quad x_K,$$

which define the middle thresholds of EW discretization. The result we get with the described procedure satisfies the Condition 5 and we have a necessary and sufficient condition for the probability distribution on some set of responses r_k, $k = 1, .., K$.

3.2 Optimal discretization

Let us find the optimal thresholds x_k and optimal representatives r_k, so that the discretization $\hat{X}$ will be the best approximation of X by some criterion.

For this purpose, define the function $q : \mathbb{R} \to \mathbb{R}$:

$$q(x) = r_k, \quad \text{if} \quad x \in \mathcal{R}_k, \tag{15}$$

which assigns to each $x \in \mathbb{R}$ a representative r_k. The function q is called *quantization* [16]. The function q is defined with $K-1$ by thresholds $x_2, ..., x_K$ and K representatives $r_1, ..., r_K$.

Description of the optimal discretization procedure. For a given K, we need to find the thresholds x_k and representatives r_k of the intervals $\mathcal{R}_k$ such that they minimize the mean square error (MSE):

$$\epsilon = \int_{-\infty}^{\infty} (x - q(x))^2 \cdot f_X(x)dx \tag{16}$$

$$= \sum_{k=1}^{K} \int_{x_k}^{x_{k+1}} (x - r_k)^2 \cdot f_X(x)dx. \tag{17}$$

To find the minimum of ϵ with respect to x_k and r_k we set the equations:

$$\frac{\partial \epsilon}{\partial x_k} = 0 \implies (x_k - r_{k-1})^2 \cdot f_X(x_k) - (x_k - r_k)^2 \cdot f_X(x_k) = 0 \tag{18}$$

$$\frac{\partial \epsilon}{\partial r_k} = 0 \implies 2 \cdot \int_{x_k}^{x_{k+1}} (x - r_k) \cdot f_X(x)dx = 0, \quad k = 1, ..., K. \tag{19}$$

From the obtained equations we express x_k and r_k to derive the conditions:

$$x_k = \frac{r_{k-1} + r_k}{2}, \quad k = 2, ..., K \tag{20}$$

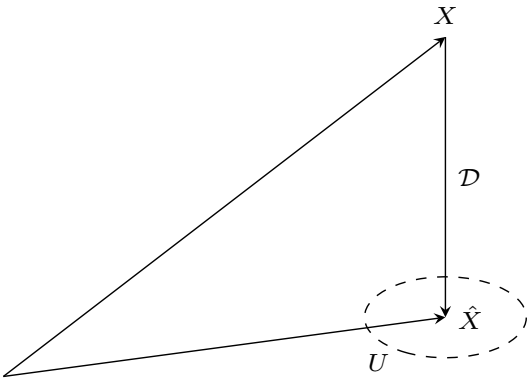
$$r_k = \frac{\int_{x_k}^{x_{k+1}} x \cdot f_X(x)dx}{\int_{x_k}^{x_{k+1}} f_X(x)dx}, \quad k = 1, ..., K. \tag{21}$$

We can also check the sufficiency of the derived conditions. If all the second order partial derivatives 18 and 19 with respect to x_k and r_k exist, then the critical point determined by the Conditions 20 and 21 is really the minimum, if the matrix of second order partial derivatives is positive definite.

Let's interpret the result. Optimal thresholds x_k lie between the optimal representatives r_k and optimal representatives r_k are centroids of the density f_X between the successive thresholds x_k. This resulting tessellation of X is called *weighted centroid Voronoi tessellation*, because it is weighted with respect to the density f_X and centroid, because the representative of each cell is the centroid of that cell.

For optimal discretization we can use the Lloyd-Max [17] algorithm. This algorithm is very similar to the k-means algorithm, it repeatedly finds the centroid of each set and then rearranges the boundaries of the sets, but the input to the algorithm is a continuous rather than a discrete set. In our examples when using skew-normal distribution, the convergence is fast and we stop after 10 iterations.

We can imagine the optimal discretization as an orthogonal projection $\mathcal{D}$ of a continuous random variable X from latent space to manifest space:



When using Lloyd-Max algorithm we allow some error. Thus, the resulting discretization is a discrete random variable $\hat{X}$ from the neighborhood U of the orthogonal projection $\mathcal{D}$ of X from latent space to manifest space.

3.3 Properties of optimal discretization

We list some important properties of the optimal discretization $\hat{X}$ of a continuous random variable X.

Proposition 2. *Optimal discretization is unbiased:*

$$\mathrm{E}[X] = \mathrm{E}[\hat{X}] \tag{22}$$

Proof. It follows from the Condition 21 for r_k:

$$\begin{aligned}
\mathrm{E}[\hat{X}] &= \sum_{k=1}^{K} r_k \int_{x_k}^{x_{k+1}} f_X(x)dx \\
&= \sum_{k=1}^{K} \int_{x_k}^{x_{k+1}} x \cdot f_X(x)dx \\
&= \int_{x_1}^{x_{K+1}} x \cdot f_X(x)dx \\
&= \mathrm{E}[X].
\end{aligned}$$

□

Proposition 3. *The error we make with optimal discretization is orthogonal or decorrelated with a random variable $\hat{X}$:*

$$\mathrm{E}\left[\left(X - \hat{X}\right) \cdot \hat{X}\right] = 0. \tag{23}$$

Proof. Equivalently we must show

$$\mathrm{E}[X \cdot \hat{X}] = \mathrm{E}[\hat{X}^2].$$

Formally $\hat{X}$ must be treated as a continuous random variable with density $f_{\hat{X}}$, which is a step function defined on the set of real numbers:

$$f_{\hat{X}}(\hat{x}) = \sum_{k=1}^{K} p_k \cdot \mathbb{1}_{\mathcal{R}_k}(\hat{x}), \quad \hat{x} \in \mathbb{R} \tag{24}$$

where $\mathbb{1}_{\mathcal{R}_k}$ is an indicator function:

$$\mathbb{1}_{\mathcal{R}_k}(\hat{x}) = \begin{cases} 1, & \text{if} \quad \hat{x} \in \mathcal{R}_k \\ 0, & \text{if} \quad \hat{x} \notin \mathcal{R}_k. \end{cases}$$

Correlation between X and $\hat{X}$ is then:

$$\mathrm{E}[X \cdot \hat{X}] = \sum_{k=1}^{K} \int_{x_k}^{x_{k+1}} x \cdot \hat{x} \cdot f_{X,\hat{X}}(x, \hat{x} = r_k)dx.$$

When $x \in \mathcal{R}_k$, the total density can be written as

$$f_{\hat{X}}(\hat{x} = r_k \mid \hat{x} \in \mathcal{R}_k) \cdot f_X(x \in \mathcal{R}_k) = f_X(x \in \mathcal{R}_k).$$

That is why

$$\mathrm{E}[X \cdot \hat{X}] = \sum_{k=1}^{K} r_k \int_{x_k}^{x_{k+1}} x \cdot f_X(x) dx.$$

From Condition 21 and Proposition 2 follows:

$$\begin{aligned}
\mathrm{E}[X \cdot \hat{X}] &= \sum_{k=1}^{K} r_k \int_{x_k}^{x_{k+1}} x \cdot f_X(x) dx \\
&= \sum_{k=1}^{K} r_k^2 \cdot f_{\hat{X}}(\hat{x} = r_k) \\
&= \mathrm{E}[\hat{X}^2].
\end{aligned}$$

□

The consequence of the Propositions 2 and 3 is as follows.

Proposition 4. *Variance of optimal discretization $\hat{X}$ is:*

$$\mathrm{var}(\hat{X}) = \mathrm{var}(\mathrm{X}) + \epsilon, \tag{25}$$

where $\epsilon \in \mathbb{R}$ is minimal for a given random variable X.

Proof. Similar to the proof of Proposition 3, we must treat $\hat{X}$ as a continuous random variable. Using Propositions 2, 3:

$$\begin{aligned}
\epsilon &= \mathrm{E}[\Big(X - \hat{X}\Big)^2] \\
&= \mathrm{E}[X^2 - 2 \cdot X \cdot \hat{X} + \hat{X}^2] \\
&= \mathrm{E}[X^2] - 2 \cdot \mathrm{E}[\hat{X}^2] + \mathrm{E}[\hat{X}^2] \\
&= \mathrm{E}[X^2] - \mathrm{E}[\hat{X}^2] \\
&= \big(\mathrm{E}[X^2] - \mathrm{E}[X]^2\big) - \Big(\mathrm{E}[\hat{X}^2] - \mathrm{E}[\hat{X}]^2\Big) \\
&= \mathrm{var}(X) - \mathrm{var}(\hat{X}).
\end{aligned}$$

□

4 Simulation of Likert-type responses

4.1 Common approach

One of the most commonly used [15] approaches of simulating Likert-type item responses is to choose a normal distribution with some mean and variance for the underlying latent variable X. And then sample from this distribution to get a random sample $\mathbf{x}$. Then discretize $\mathbf{x}$ using using equal width discretization to simulate a random sample $\hat{\mathbf{x}}$ from manifest variable $\hat{X}$.

To simulate responses that are distributed asymmetrically can be achieved by adjusting the widths of the intervals of EW discretization. For example for the right asymmetry, we can set each interval to be half as narrow as the previous one.

The problem with this approach is the lack of control. There are no guarantees of what the result would be. If we do not take into account the relationship between X and $\hat{X}$, then evaluating the statistical methods with such a simulation process can be questionable at least.

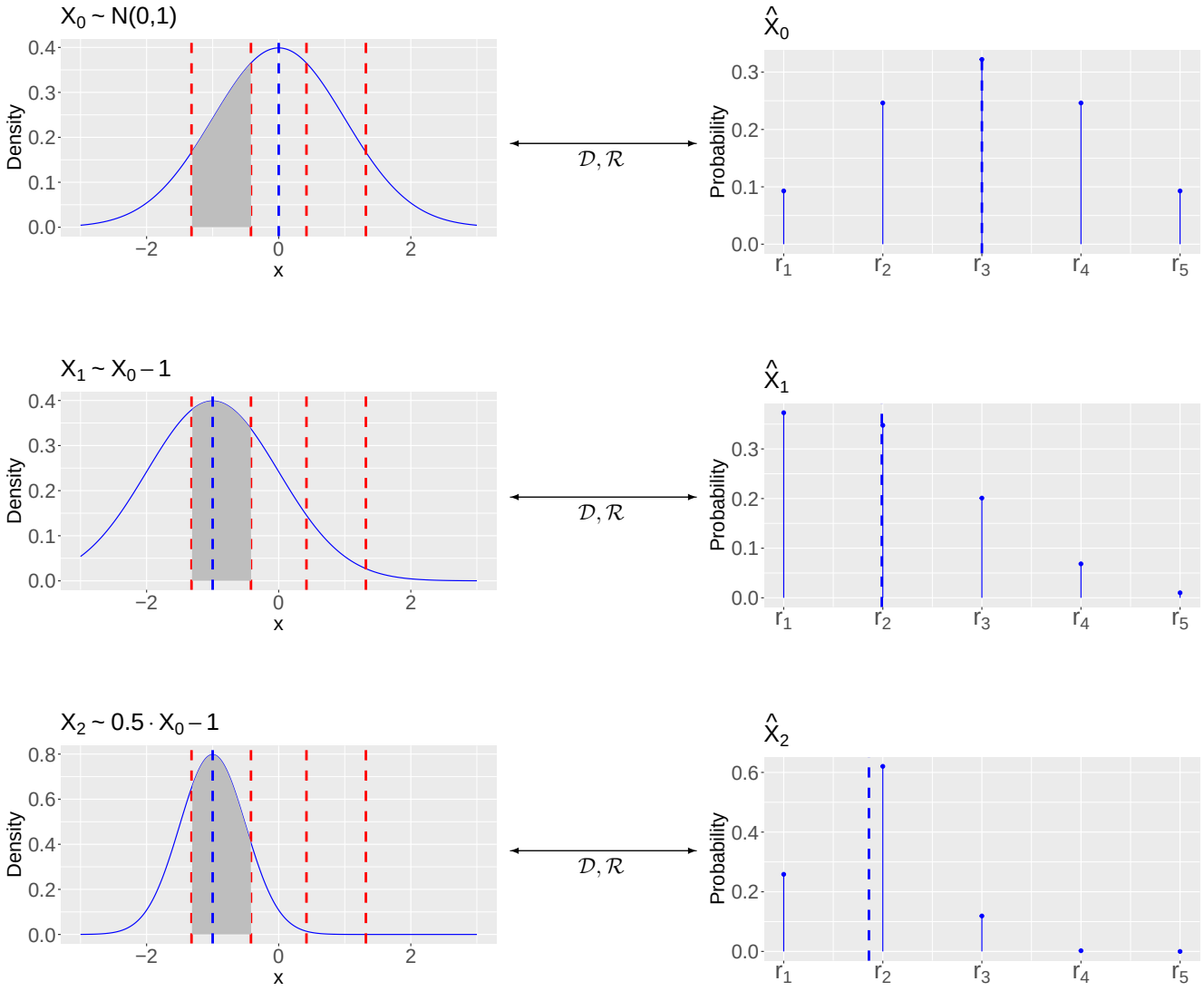


Figure 2: Examples of pairs X_i, $\hat{X}_i$.

4.2 Proposed approach

Let there be K possible responses r_k, $k = 1, ..., K$, which measure say people's agreement with some abstract idea. Imagine a hypothetical population of people with an underlying latent variable X_0 with expected value:

$$\mathrm{E}[X_0] = 0 \tag{26}$$

and variance:

$$\mathrm{var}(X_0) = \sigma^2. \tag{27}$$

We call this population *neutral*, because we can say they are on average neutral about this abstract idea. Let $\hat{X}_0$ be the manifest variable, obtained by using the thresholds x_k and defined on responses r_k. The probability that $\hat{X}_0$ equals r_k is by definition:

$$\mathrm{P}(\hat{X}_0 = r_k) = f_X(x \in \mathcal{R}_k), \quad k = 1, ..., K. \tag{28}$$

The expected value of the manifest variable is then:

$$\mathrm{E}[\hat{X}_0] = 0 \tag{29}$$

and variance is:

$$\mathrm{var}(\hat{X}_0) = \mathrm{var}(X_0) + \epsilon, \tag{30}$$

where $\epsilon \in \mathbb{R}$ is minimal for a given latent variable X_0 when the manifest variable is the optimal discretization of X_0. That is when the responses r_k are the optimal representatives of the intervals $\mathcal{R}_k$.

Imagine some other hypothetical population with underlying latent variable X_i:

$$X_i \sim \omega_i \cdot X_0 + \xi_i \tag{31}$$

for some real number ξ_i and for some positive real number ω_i. For example, for $\xi_i < 0$, this population on average agrees for ξ_i less with this abstract idea than neutral population, and for $\omega_i < 1$, the agreement of this population is less dispersed than in neutral population. The same responses r_k, $k = 1, ..., K$ are available to this population. Therefore the probability that $\hat{X}_i$ is r_k equals:

$$\mathrm{P}(\hat{X}_i = r_k) = f_{X_i}(x \in \mathcal{R}_k), \quad k = 1, ..., K. \tag{32}$$

Examples of pairs: X_i, $\hat{X}_i$ for $i = 0, 1, 2$ are in Figure 2, where $X_0 \sim N(0, 1)$, $\xi_1 = \xi_2 = -1$, $\omega_1 = 1$ and $\omega_2 = 0.5$. In the Figure, the density f_{X_i} and the distribution $\hat{X}_i$ is shown in blue, the thresholds x_k, $k = 2, ..., 5$ dashed in red, area $p_2 = \mathrm{P}(\hat{X}_i = 2)$ with grey, and the expected values are dashed in blue.

To simulate asymmetric manifest variables, we can choose $X_0 \sim \mathcal{SN}(\alpha)$, for some $\alpha \neq 0$.

This simulation approach can be very useful for having more control, as the relationship between pairs X_i, $\hat{X}_i$ is preserved, as we did not use biased discretization.

5 A procedure to reconstruct latent variables

Optimal estimator of p_k for the manifest random variable $\hat{X}$ of some population is:

$$\hat{p}_k = \frac{b_k}{n}, \tag{33}$$

where b_k is the number of responses with the value r_k and n is the number of all responses. This follows from Eq. 11. In this reconstruction procedure we assume optimal thresholds. Responses r_k are arbitrary, they can be: totally disagree, ..., completely agree.

Based on the sample from $\hat{X}$, we would like to find the values of parameters α, ξ, ω and thus reconstruct the latent variable, which we assume follows the skew-normal distribution:

$$X \sim \omega \cdot X_0(\alpha) + \xi \tag{34}$$

Let the random variable X_0 be the latent variable of the neutral population:

$$X_0(\alpha) \sim X - \mu \tag{35}$$

where $X \sim \mathcal{SN}(\alpha)$ and $\mu = \mathrm{E}(X)$. The parameters α, ξ, ω describe the whole family of skew-normal distributed random variables. We see that $\hat{X}$ is well defined because we have assumed the thresholds $x_k(\alpha)$ which discretize $X_0(\alpha)$ for each value of α.

5.1 Estimation of parameters ξ and ω

For a given $\alpha = \alpha_0$:

$$p_k = F_{\omega\cdot(X-\mu)+\xi}(x_{k+1}) - F_{\omega\cdot(X-\mu)+\xi}(x_k) \tag{36}$$

$$= F_X(\frac{x_{k+1} - \xi}{\omega} + \mu) - F_X(\frac{x_k - \xi}{\omega} + \mu), \qquad k = 1, ..., K. \tag{37}$$

From this we obtain an overdetermined system of K nonlinear equations for parameters ξ and ω:

$$g_k(\xi, \omega) = F_X(\frac{x_{k+1} - \xi}{\omega} + \mu) - F_X(\frac{x_k - \xi}{\omega} + \mu) - p_k, \qquad k = 1, ..., K. \tag{38}$$

This is a nonlinear least squares problem that can only be solved numerically because F_X is not expressed by elementary functions. For greater stability, we reparametrize it:

$$u = \xi \tag{39}$$
$$v = 1/\omega, \tag{40}$$

and obtain a new system of equations:

$$h_k(u,v) = F_X(v \cdot x_{k+1} - v \cdot u + \mu) - F_X(v \cdot x_k - v \cdot u + \mu) - p_k, \qquad k = 1, ..., K, \tag{41}$$

which can be written in matrix form:

$$H(u,v) = \begin{pmatrix} h_1(u,v) \\ . \\ . \\ . \\ h_K(u,v) \end{pmatrix}.$$

And compute partial derivatives:

$$\frac{\partial h_k}{\partial u} = f_X(v \cdot x_{k+1} - v \cdot u + \mu) \cdot (-v) - f_X(v \cdot x_k - v \cdot u + \mu) \cdot (-v)$$
$$\frac{\partial h_k}{\partial v} = f_X(v \cdot x_{k+1} - v \cdot u + \mu) \cdot (x_{k+1} - u) - f_X(v \cdot x_k - v \cdot u + \mu) \cdot (x_k - u)$$

For the edge cases where $x_1 = -\infty$ or $x_K = \infty$, the limit $f_X(x)$, when x goes to $-\infty$ or ∞, equals to 0 and the terms with x_1 and x_K can be omitted. We obtain the Jacobian matrix:

$$J(u,v) = \begin{pmatrix} \frac{\partial h_1}{\partial u} & \frac{\partial h_1}{\partial v} \\ . & . \\ . & . \\ \frac{\partial h_K}{\partial u} & \frac{\partial h_K}{\partial v} \end{pmatrix} \mid (u,v)$$

and solve the system using the Gauss-Newton method, starting from an initial guess:

$$(u_0, v_0) = (0, 1)$$

and finding (u_*, v_*) which minimizes the second norm:

$$\|H(u_*, v_*)\|_2 . \tag{42}$$

The Gauss-Newton method was implemented as follows. At each step, we evaluate the Jacobian matrix

$$A = J(u_i, v_i)$$

and decompose into singular values:

$$A = UDV.$$

The correction d for (u_i, v_i) is calculated by solving the linear least squares problem, obtaining:

$$d = V' D^{-1} U' H(u_i, v_i).$$

and finding the next approximation:

$$(u_{i+1}, v_{i+1}) = (u_i, v_i) + d_*.$$

Implementation is adaptive in the following way. We adjust the correction d at each step i as needed. While the value of correction d yields: $v_{i+1} < 0$, we halve the correction d:

$$d := d/2.$$

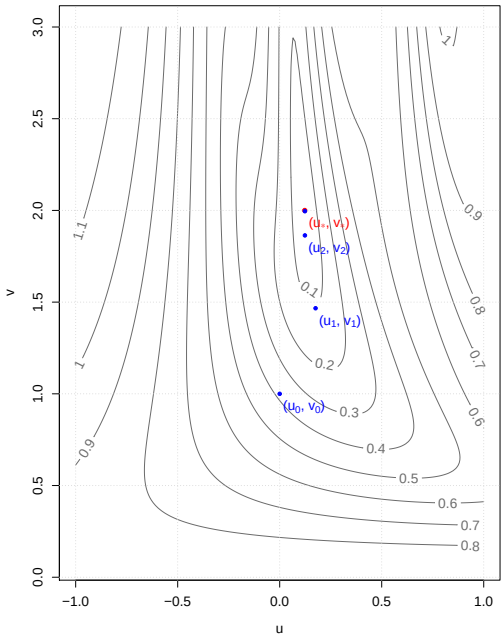


Figure 3: Trace of approximations using the adaptive Gauss-Newton method.

This way we reduce the value d exponentially, until we get the first appropriate value of correction d. The value of v_i must not be less than 0, because $\omega > 0$.

An example is in Figure 3, which shows a trace of approximations obtained by the adaptive Gauss-Newton method when reconstructing the parameters:

$$\xi = 0.2 \cdot \mathrm{sd}(X) \approx 0.123$$
$$\omega = 0.5$$

and $X \sim \mathcal{SN}(\alpha)$ where $\alpha = 6.3$. In this case $v = 2$ and $u = \xi$.

5.2 Estimation of parameter α

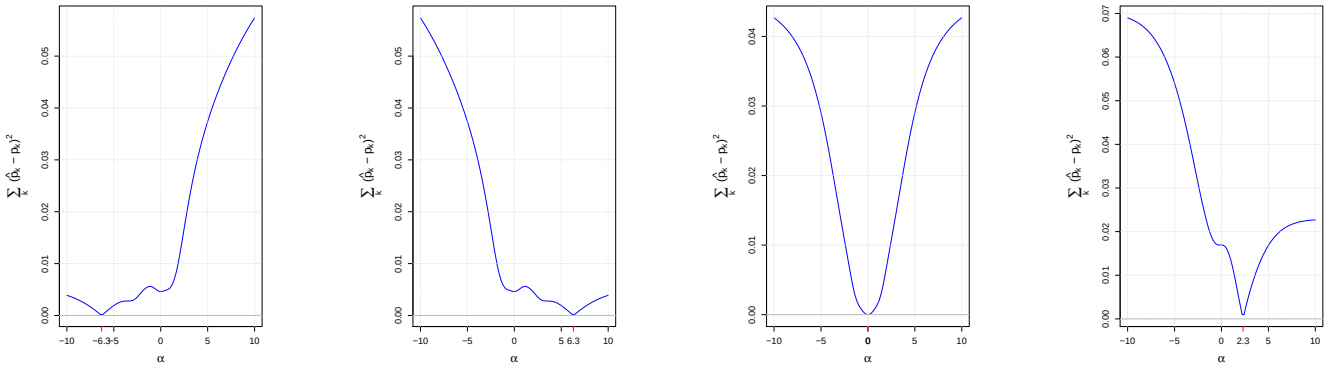
The sample from asymmetric manifest variable $\hat{X}$ may look symmetric ($\alpha = 0$) when in fact comes from an asymmetric distribution ($\alpha \neq 0$). Therefore, the value of α is estimated as follows.

For each value α_i of parameter α we calculate the corresponding values of ξ_i and ω_i using the procedure already described. This procedure reconstructs some continuous random variable X_i. We discretize it using thresholds $x_k(\alpha_i)$ to get $\hat{X}_i$, which has probabilities $\hat{p}_i$.

If the value α_i of parameter α is far from the actual value, then $\hat{p}_k$ are far from the actual p_k. We can calculate the error:

$$\epsilon = \sum_{k=1}^{K} (\hat{p}_k - p_k)^2. \tag{43}$$

We choose the value of α, which minimizes the error ϵ.



(a) $\alpha = -6.3$. (b) $\alpha = 6.3$. (c) $\alpha = 0$. (d) $\alpha = 2.3$.

Figure 4: Estimation of the value of parameter α.

Examples of reconstructions are in Figure 4 for different actual parameter values. For example, if the actual value of α is negative, the error ϵ in terms of the value of α is shown in Figure **??**. The actual values of parameters ξ and ω only affect the shape of the function of error ϵ in terms of α and do not effect at which value of α the minimum error ϵ is reached.

5.3 Estimation of parameters from a sample

As the value of $|\alpha|$ increases, the density of the random variable $X \sim \mathcal{SN}(\alpha, \cdot, \cdot)$ converges to half normal density. For values $|\alpha|$ greater than 10, the differences between the distributions are negligible. Therefore, we limit ourselves to the interval $[-10, 10]$ on which we look for the value of α and choose a denser division near $\alpha = 0$, where the differences between the distribution are greatest. From a small sample sizes, we can get estimates that are very different from the actual values of parameters. However, for larger sample sizes, the estimates are very close to the true values. For example, for the sample size $n = 1000$ and actual parameters:

$$\begin{aligned} \alpha &= 0 \\ \xi &= -1 \\ \omega &= 1 \end{aligned}$$

we get estimates:

$$\begin{aligned} \alpha &= -0.2000000 \\ \xi &= -0.9893231 \\ \omega &= 1.0012696 \end{aligned}$$

The reconstruction is shown in Figure 5. On small sample sizes the shape of the error function ϵ in terms of α can even become concave. We select the value of α, where the first select the value of the alpha parameter where extreme value of error function ϵ is reached.

6 Comparison of two means

Say we use a survey and gather the answers of a random sample of people to questions by which we measure agreement with some abstract idea. For example, we measure agreement with non-compliance with rules with a question: “How much do you agree with: To request state aid for something which you are not entitled” with 5 possible responses: completely disagree, ..., completely agree. And say we only want to find out if there is a difference between the mean values of agreement between two populations, for example between male and female populations. We would like to reject the null hypothesis that there is no difference and conclude, for example, that men on average agree more with non-compliance than women.

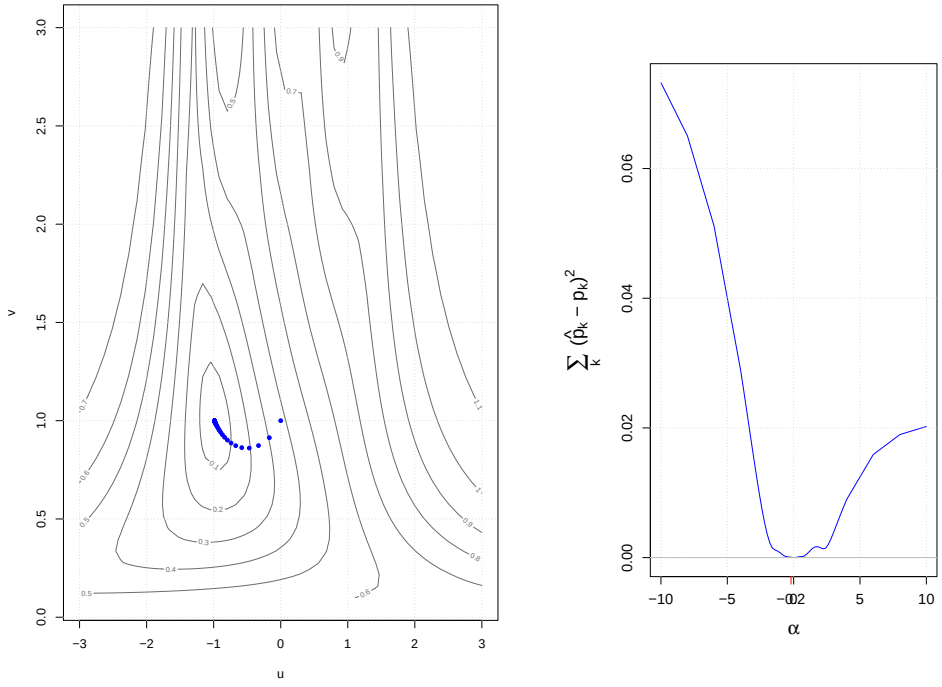


Figure 5: Reconstruction from a sample.

In this context, a nonparametric test Mann-Whitney is often proposed [18], however this tests different null hypothesis, namely whether the samples are coming from populations with same distribution. We can also try to use a model for analyzing ordinal data, the so-called proportional odds model, and then test the influence of dichotomous covariant variable (e.g. gender) on the ordinal response [19]. However, we obtain the same test statistic as with the Mann-Whitney test, proof in [20]. We can also compare the response rates between the groups with χ^2-square test. If we only want to find out if there is a difference between the mean values, this tests are not suitable. One of the most commonly used approach is to apply t-test on encoded responses directly.

6.1 Common testing procedure

The sum of responses r_k is not defined in general. For example: completely disagree + ... + completely agree = ? Therefore the common practice is to first *encode* the responses. This means to map the responses to some set of numbers. In the case of K possible responses, usually into the first K natural numbers, e.g.: completely disagree $\mapsto$ 1, ..., completely agree $\mapsto K$.

Let r_k represent encoded responses:

$$r_k \in \mathbb{R}, \qquad k = 1, ..., K \tag{44}$$

and let $\hat{X}$ $k = 1, ..., K$ be a manifest variable defined on the encoded responses. Denote the expected value of $\hat{X}$:

$$\mathrm{E}[\hat{X}] = \hat{\mu}$$

and standard deviation:

$$\mathrm{sd}(\hat{X}) = \hat{\sigma}.$$

Denote by $\hat{X}_{i,j}$ the response of the individual i to question j. Responses $\{\hat{X}_{i,j}\}$ for sample size n from some population to m questions can be represented by table:

	1	...	m
1	$\hat{X}_{1,1}$	...	$\hat{X}_{1,m}$
.	.		.
.	.		.
.	.		.
n	$\hat{X}_{n,1}$	...	$\hat{X}_{n,m}$

Common practice is to compute the averages of encoded responses by individuals. Denote this average by $\hat{X}_i$:

$$\hat{X}_i = \frac{1}{m}\sum_{j=1}^{m}\hat{X}_{i,j}. \tag{45}$$

It holds:

$$\mathrm{E}[\frac{1}{n}\sum_{i=1}^{n}\hat{X}_i] = \frac{1}{n\cdot m}\sum_{i=1}^{n}\sum_{j=1}^{m}\hat{X}_{i,j} = \hat{\mu} \tag{46}$$

and

$$\begin{aligned}\mathrm{var}(\hat{X}_i) &= \mathrm{var}\left(\frac{1}{m}\sum_{j=1}^{m}\hat{X}_{i,j}\right)\\ &= \frac{1}{m^2}\left(\sum_{j=1}^{m}\mathrm{var}(\hat{X}_{i,j}) + \sum_{j_1\neq j_2}\mathrm{cov}(X_{i,j_1}, X_{i,j_2})\right)\end{aligned}$$

In general, we should model responses to different questions with different latent variables and covariances between pairs of these latent variables. We simplify by accepting the generally accepted assumption [1] that these questions are equivalent instruments for measuring, e.g. agreement with some abstract idea. Therefore, for each pair j_1, j_2, the following holds:

$$\mathrm{cov}(\hat{X}_{i,j_1}, \hat{X}_{i,j_2}) = \mathrm{var}(\hat{X}_{i,j}) \tag{47}$$

and $\mathrm{var}(\hat{X}_i)$ simplifies to:

$$\begin{aligned}\mathrm{var}(\hat{X}_i) &= \frac{1}{m^2}\left(\sum_{j=1}^{m}\mathrm{var}(\hat{X}_{i,j}) + \sum_{j_1\neq j_2}\mathrm{var}(\hat{X}_{i,j})\right)\\ &= \frac{1}{m^2}\left(m\cdot\hat{\sigma}^2 + m\cdot(m-1)\cdot\hat{\sigma}^2\right)\\ &= \hat{\sigma}^2.\end{aligned}$$

Let $\hat{X}_1$ and $\hat{X}_2$ be two manifest variables defined on encoded responses of two different populations to the same questions. Denote the expected values $\hat{\mu}_1$, $\hat{\mu}_2$ and standard deviations $\hat{\sigma}_1$, $\hat{\sigma}_2$ of the manifest variables $\hat{X}_1$, $\hat{X}_2$ respectively.

Common practice is to take random samples from manifest variable $\hat{X}_i$ of size n_i, estimate the expected value $\hat{\mu}_i$ and standard deviation $\hat{\sigma}_i$, for $i = 1, 2$. Then to use for example Welch t-test and estimate the ratio:

$$\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \tag{48}$$

and finally calculate the p-value from t-distribution, where the degrees of freedom can be calculated using Satterhaite-Welch correction.

This procedure is a serious simplification because the value of the ratio σ_1^2/σ_1^2 is unknown. We must instead use a bit more complicated test statistic for which the distribution is not known in closed-form. By using the likelihood ratio test we also obtain equations that do not have a closed-form solution and must use an algorithm to solve them.

6.2 Problem with the common testing procedure

We show that the common testing procedure is not suitable for mean comparison when variances of latent variables differ. It is not suitable even in the best case when we assume optimal discretization. We also show this problem persist if we use the bootstrap test or permutation test this time only by a simulation study at the end.

Let's look at a simple example where we obtain biased test statistics using the common testing procedure. Pairs of latent and manifest variables for this example are shown in Figure 2.

Imagine again a neutral population with standard normal latent variable $X_0 \sim N(0,1)$ and suppose the corresponding manifest variable is optimal discretization $\hat{X_0}$ of X_0 with thresholds x_k and representatives r_k for $k = 1, ..., K$.

Now suppose we have a population with latent variable $X_1 \sim X_0 - 1$, we can imagine that, on average, they agree less with this abstract idea then the neutral population. And suppose we have another population with latent variable $X_2 \sim 0.5 \cdot X_0 - 1$, we can imagine that, on average, they agree the same as the second population, but their agreement is less dispersed.

It holds:

$$\mathrm{E}[X_1] - \mathrm{E}[X_2] = 0 \tag{49}$$

but

$$\mathrm{E}[\hat{X_1}] - \mathrm{E}[\hat{X_2}] \approx 0.13. \tag{50}$$

In this case, we used optimal discretization. The difference is of course not equal to 0 if we use the EW discretization.

Proposition 5. *The test statistic in common testing procedure is biased in general.*

Proof. Let there be two populations with latent variables X_1, X_2 and corresponding manifest variables $\hat{X_1}$, $\hat{X_2}$. Assume that to both populations are available the same responses on the questionnaire. Assume the representatives r_k, on which manifest variables are defined, are real numbers or encoded responses from the questionnaire. Then

$$\mathrm{E}[\hat{X_i}] = \mathrm{E}[X_i] + \delta_i \tag{51}$$

for some $\delta_i \in \mathcal{R}$ and

$$\mathrm{var}(\hat{X_i}) = \mathrm{var}(X_i) + \epsilon_i \tag{52}$$

for positive real number ϵ_i that depends on X_i for $i = 1, 2$.

With the common procedure, based on sample from manifest variable of size n_i, we estimate the test statistic:

$$T = \frac{\mathrm{E}[\hat{X_1}] - \mathrm{E}[\hat{X_2}]}{\sqrt{\frac{\mathrm{var}(\hat{X_1})}{n_1} + \frac{\mathrm{var}(\hat{X_2})}{n_2}}} \tag{53}$$

$$= \frac{\mathrm{E}[X_1] + \delta_1 - \mathrm{E}[X_2] - \delta_2}{\sqrt{\frac{\mathrm{var}(X_1)+\epsilon_1}{n_1} + \frac{\mathrm{var}(X_2)+\epsilon_2}{n_2}}} \tag{54}$$

Now suppose that:

$$\mathrm{var}(X_1) \neq \mathrm{var}(X_2) \tag{55}$$

and:

$$\mathrm{E}[X_1] = \mathrm{E}[X_2]. \tag{56}$$

The same answers on the questionnaire are available to both populations. Threrefore, the manifest variables $\hat{X_1}$, $\hat{X_2}$ are defined on the same domain r_k, $k = 1, ..., K$. Then $\hat{X_1}$ and $\hat{X_2}$ cannot be both optimal discretizations of latent variables X_1 and X_2. Therefore, δ_1 and δ_2 cannot both be zero. So the difference $\mathrm{E}[\hat{X_1}] - \mathrm{E}[\hat{X_2}]$ can be arbitrarily large. □

6.3 Proposed testing procedure on a concrete example

We illustrate our proposed testing procedure and common testing procedure on a concrete example. Data is from the World Values Survey from Sweden in 2011 [21]. The responses are values from 1 to 10. A value of 1 means the least agreement and a value of 10 the most agreement with how justifiable the claim is. The surveyed Swedes responded to the following claims:

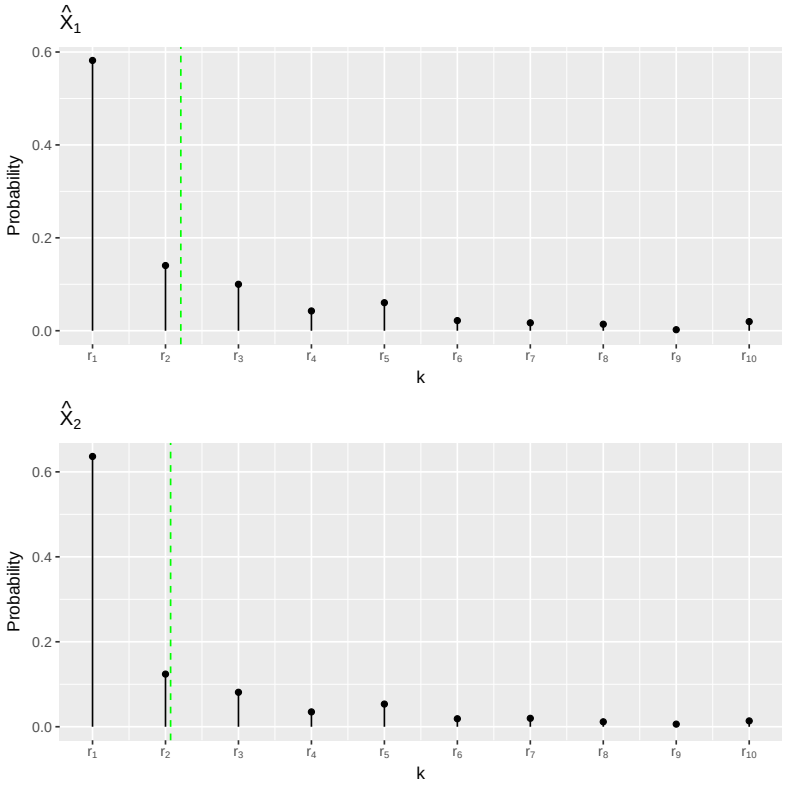


Figure 6: Estimated manifest variables for male and female populations.

- Claiming government benefits to which you are not entitled.
- Avoiding a fare on public transport.
- Stealing property.
- Cheating on taxes if you have a chance.
- Someone accepting a bribe in the course of their duties.

We suppose all the questions are equivalent instruments for measuring agreement with some abstract idea in this case the idea can be described as non-compliance with the rules.

We have a sample of $n_1 = 540$ male responses and a sample of $n_2 = 584$ female responses to these 5 claims. With common testing procedure we get p-value ≈ 0.10341.

Let's illustrate our testing procedure for comparing means described in 5. We assume that agreement follows skew-normal distribution and compute estimates for p_k of manifest random variable $\hat{X}_1$ for male population, by dividing the number of male responses with the number of questions 5: $540 \cdot 5$ and similar for female responses. We obtain estimates of the manifest variables shown in Figure 6.

Let X_1, X_2 be latent variables for male and female populations and assume optimal thresholds so we can reconstruct both latent variables. For both populations, we obtain $\alpha_1 = \alpha_2 = -1.2$. The trace of the adaptive Gauss-Newton method and the function of error ϵ in terms of α based on male responses is shown in Figure 7. Estimates μ_i and standard deviation σ_i for X_i, $i = 1, 2$ is:

$$\begin{aligned}
\hat{\mu}_1 &= \xi_1 \approx -3.33, \\
\hat{\sigma}_1 &= \omega_1 \cdot \mathrm{sd}(X_0) \approx 2.97, \\
\hat{\mu}_2 &= \xi_2 \approx -2.99, \\
\hat{\sigma}_1 &= \omega_1 \cdot \mathrm{sd}(X_0) \approx 3.25,
\end{aligned}$$

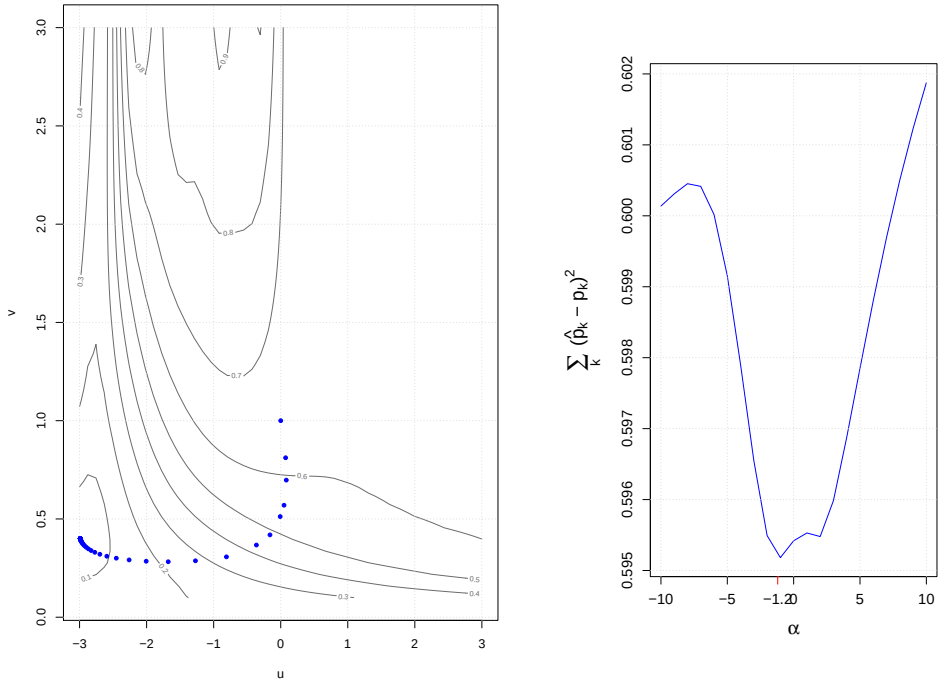


Figure 7: Trace of the Gauss-Newton method and function of error ϵ in terms of α based on male responses.

where $X_0 \sim SN(-1.2)$.

Finally we calculate test statistics:

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}} \tag{57}$$

and p-value from t-distribution, where we calculate the degrees of freedom with Satterthaite-Welch correction and obtain p-value ≈ 0.06449033.

7 Conclusions

We introduce a model to explore the relationship between Likert-type responses and underlying hypothetical latent variables. Then derive the optimal relationship in the form of discretization and reconstruction procedures. We use this discretization procedure to simulate Likert-type responses. We then show the problem with the common approach to comparison of two means in the context of our model. And propose an alternative approach where we illustrate the use of our reconstruction procedure. Source code is available at:

`https://github.com/markolalovic/latent-variable-reconstruction`

References

[1] A. Van Alphen, R. Halfens, A. Hasman, and T. Imbos, "Likert or rasch? nothing is more applicable than good theory," *Journal of Advanced Nursing. Vol. 20(196–201)*, 1994.

[2] M. R. Harwell and G. G. Gatti, "Rescaling ordinal data to interval data in educational research," *Review of Educational Research. Vol. 71(105–131)*, 2001.

[3] J. Camparo, "A geometrical approach to the ordinal data of likert scaling and attitude measurements: The density matrix in psychology," *Journal of Mathematical Psychology. Vol. 57(29–42)*, 2013.

[4] J. Camparo and B. Camparo, Lorinda, "The analysis of likert scales using state multipoles an application of quantum methods to behavioral sciences data," *Journal of Educational and Behavioral Statistics: SAGE Journals*, 2013.

[5] R. Jabrayilov, W. H. M. Emons, and K. Sijtsma, "Comparison of classical test theory and item response theory in individual change assessment," *Applied Psychological Measurement. Vol. 40(8)*, 2016.

[6] C. Magno, "Demonstrating the difference between classical test theory and item response theory using derived test data," *The International Journal of Educational and Psychological Assessment. Vol. 1(1)*, 2009.

[7] G. Boari and M. Ruscone, "A procedure simulating likert scale item responses," *Electronic Journal of Applied Statistical Analysis. Vol. 8(3)*, 2015.

[8] W. Young, F., Y. Takane, and J. De Leeuw, "The principal components of mixed measurement level multivariate data," *Psychometrika. Vol. 43(279)*, 1978.

[9] A. Gifi, *Nonlinear Multivariate Analysis*. Wiley, 1990.

[10] J. de Leeuw, *Visualization and Verbalization of Data; History of Nonlinear Principal Component Analysis*. CRC Press, 2005.

[11] C. Eijk and J. Rose, "Risky business: Factor analysis of survey data – assessing the probability of incorrect dimensionalisation," *PLOS ONE, Vol. 10(3)*, 2015.

[12] S. Rice, "Quantitative methods in politics," *Journal of the American Statistical Association. Vol. 33*, 1938.

[13] K. A. Bollen, *Structural equations with latent variables*. John Wiley, 2014.

[14] A. Azzalini, "A class of distributions which includes the normal ones," *Scandinavian Journal of Statistics. Vol. 12*, 1985.

[15] H. Wu and S.-O. Leung, "Can likert scales be treated as interval scales?—a simulation study," *Journal of Social Service Research. Vol. 43(4)*, 2017.

[16] R. Gonzalez and R. Woods, *Digital Image Processing*. Prentice-Hall, 2008.

[17] J. Max, "Quantizing for minimum distorsion," *IRE Transactions on Information Theory. Vol. 6(1)*, 1960.

[18] D. Winter, F. Joost, C., and D. Dodou, "Five-point likert items: t test versus mann-whitney-wilcoxon," *Practical Assessment Research I& Evaluation. Vol. 15(11)*, 2010.

[19] M. P., "Regression models for ordinal data (with discussion)," *Journal of the Royal Statistical Society B. Vol. 42(2)*, 1980.

[20] S. Natarajan, R. Lipsitz, S., M. Fitzmaurice, G., and et al., "An extension of the wilcoxon rank-sum test for complex sample survey data," *Journal of the Royal Statistical Society C. Vol. 61(4)*, 2012.

[21] "World values survey." `http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`. Accessed: 2018-10-24.