

The Sum Scores and Discretization of Variables Under the Linear Normal One-Factor Model



Rudy Ligtoet

Abstract The sum score is often used in practical test applications and the joining of outcomes is common practice, when preparing response data for analysis. Yet, many models for response data are not designed for this kind of handling of data. Research on the use of the sum score for stochastic inferences and the discretization of response variables is extended to the linear normal one-factor. It is shown that the model implies a stochastic ordering on the latent factor by the sum of the observed variables, but that this property no longer needs to hold when variables are discretized prior to taking the sum score. The implications of this result are discussed.

Keywords Discretization · Linear normal one-factor model · Pólya frequency functions · Sum scores · Totally positive densities

1 Introduction

For test and questionnaire data, respondents are often assigned a latent value for the attribute that the test is aimed to measure, based on a model for the dependencies that exist between the test items. Because the latent variable is unobserved, it is convenient to consider the observable sum scores across the test items as a proxy for the latent values instead. For the use of the sum score, a desirable feature of a measurement model would then be that a higher sum score also corresponds to a higher expected latent value, so that the ordering of respondents by their sum scores *stochastically* agrees with the ordering by their latent values. The use of the sum score for making such ordinal inferences has been studied for various item response theory models (Hemker et al., 1996, 1997; Ligtoet, 2012, 2015), based on a *monotone likelihood ratio* (MLR) ordering of the latent variable by the sum scores.

R. Ligtoet (✉)
University of Cologne, Cologne, Germany
e-mail: rligtoe@uni-koeln.de

For these models, Hemker (2001) also looked at the effect of joining response categories. The results show that the MLR property is not implied by all models, and that for some models the model may be invalidated when (adjacent) outcomes are joined (cf. Andrich, 1995a, 1995b; Roskam, 1995). With the omnipresent use of the sum (or average) score across many test applications and the common practice of collapsing outcomes (e.g., median split), there thus seems to be a mismatch between the models proposed for response data and the way these data are handled in practice. The present chapter looks at the use of the sum score and the discretization of response data under the *linear normal one-factor* (LNF) model.

1.1 The Linear Normal One-Factor

Factor analysis provides a framework to account for dependencies that exist between the multiple item response variables of a test, whereby item response variables serve as indicators of the common factors that the test aims to measure. The LNF model proposes a single latent variable or factor Z to account for the covariances between the item variables X_1, \dots, X_n by the linear relationships

$$X_i = a_i Z + U_i, \text{ with } a_i > 0,$$

where a_i denotes the i th factor loading and U_i is the i th residual or unique factor. It is assumed that U_1, \dots, U_n, Z are independent and normally distributed, with zero means (centered), and (non-negative) variances

$$\text{Var}(U_i) = \sigma_i^2 \text{ and } \text{Var}(Z) = \sigma^2.$$

Further, assume that $\text{Cov}(U_i, U_j) = 0$ (for $i \neq j$) and $\text{Cov}(U_i, Z) = 0$ (Jöreskog, 1971; Lord & Novick, 1968). Hence, under the LNF model, the variables X_1, \dots, X_n are *conditionally independent* (CI), given $Z = z$.

1.2 Monotone Transformations

In this chapter, two monotone (non-decreasing) transformations are considered that are often used on X_1, \dots, X_n in practice. Let $\mathbf{X} = (X_1, \dots, X_n)$ denote the random vector containing the variables X_i , with realizations $\mathbf{x} \in \mathbb{R}^n$. Then, a function $\phi(\mathbf{x})$ is said to be monotone, whenever $\mathbf{x} < \mathbf{y}$ (element-wise) implies that $\phi(\mathbf{x}) \leq \phi(\mathbf{y})$.

The Sum Score The first transformation that is considered is the sum score $S = X_1 + \dots + X_n$ often used in practice as a proxy for Z (McNeish & Wolf, 2020), with $\phi(\mathbf{x})$ representing a mapping of many-to-one or aggregation; i.e., $\phi : \mathbb{R}^n \rightarrow \mathbb{R}$. In practice, a LNF model is often fitted to the data in order to assess the validity

(factorial structure) of the test, whereas in subsequent analysis the sum or average score is used to ascertain the validity indices (e.g., predictive validity) of classical test theory. A minimal requirement of the LNF model for such practice is that the model implies a stochastic ordering of the factor scores by the sum scores, such that higher factor scores are expected for higher sum scores. Let $f(s, z)$ denote the joint density of S and Z , then the stochastic ordering by the sum scores is satisfied, whenever the density $f(s, z)$ is *totally positive of order 2*; that is

$$f(s_1, z_1)f(s_2, z_2) \geq f(s_1, z_2)f(s_2, z_1), \quad (\text{TP}_2)$$

for all $s_1 < s_2$ and $z_1 < z_2$ (Karlin, 1968, cf. the MLR property).

Discretization As a second transformation, consider the discretization of the variables X_1, \dots, X_n . The LNF model is often used for the analysis of discrete ordinal response data (Jöreskog and Moustaki, 2001), where X_1, \dots, X_n are taken as the ghosts underlying the observable discrete variables V_1, \dots, V_n . Here, $\phi(\mathbf{x})$ takes on the form $(\phi_1(x_1), \dots, \phi_n(x_n))$, with

$$V_i = \phi_i(X_i; b_1, \dots, b_{m_i}) \text{ and } b_1 < \dots < b_{m_i},$$

where

$$V_i = v_i, \text{ whenever } b_v \leq x_i < b_{v+1}$$

($b_{m_i+1} = -b_0 = \infty$ by definition). In words, each discretization ϕ_i proposes m_i ordered thresholds, where v_i denote the largest threshold passed by the outcome of X_i . With $v_i \in \{0, 1, \dots, m_i\}$, the outcomes of V_i are said to be *equidistant* (Andrich, 1995a). Sijtsma and Van der Ark (2017) discuss the use of an equidistant scoring rule in relationship with the use of the sum score in the context of Mokken's *monotone homogeneity* (MH) model (Mokken, 1971; Molenaar, 1997). In addition to CI and the *unidimensionality* assumption, the MH model assumes that the tail distributions $1 - F(x_i|z)$ are non-decreasing in z (Holland & Rosenbaum, 1986). The LNF model satisfies the MH model assumptions, also after applying a discretization to X_1, \dots, X_n . In case $m_i = m = 1$, the transformation $\phi(\mathbf{x})$ corresponds to a dichotomization of the response variables.

For later reference, the concept of *Pólya frequency functions of order 2* (PF₂) is introduced (e.g., Efron, 1965; Schoenberg, 1951).

Definition 1 The density $f(x)$ is said to be PF₂, if for all $x_1 < x_2$ and $y_1 < y_2$

$$f(x_1 - y_1)f(x_2 - y_2) \geq f(x_1 - y_2)f(x_2 - y_1). \quad (1)$$

Ellis (2015) showed that the (monotone higher-order) one-factor model, with residuals having PF₂ densities (e.g., normally distributed), implies that $f(\mathbf{v})$ is

multivariate TP_2 (Karlin & Rinott, 1980). This in turn implies that each $f(v_i, v_j)$ is TP_2 .

Strictly speaking, the normality requirements of the LNF model does not hold for the discrete variables V_1, \dots, V_n . However, the LNF may still provide an adequate approximation of discrete response data (Rhemtulla et al., 2012).

Chapter Overview The purpose of this chapter is to investigate the effect the discretization of the indicators X_1, \dots, X_n has on the use of the sum score for the stochastic ordering on Z . In the next section, it is shown that the LNF model implies a stochastic ordering on Z by the sum score S . However, it is also shown that the stochastic ordering property does not generally hold when the sum score is used after discretizing the indicators X_1, \dots, X_n , except in the special case of a dichotomization. These results and their implications are further discussed in Sect. 3.

2 The LNF Model and the Sum Score

In this section, it is shown that the LNF model implies a stochastic ordering on Z by the sum $S = X_1 + \dots + X_n$. However, after discretization of the variables X_1, \dots, X_n obtained under the LNF model, the sum $R = V_1 + \dots + V_n$ of the newly obtained variables V_1, \dots, V_n no longer needs to provide a stochastic ordering on the factor Z . That is, $f(s, z)$ is TP_2 does not imply that $f(r, z)$ is also TP_2 .

2.1 Preliminaries

In order to show that the LNF model implies a stochastic ordering of Z by $S = X_1 + \dots + X_n$, it is convenient to express the model in terms of the more general properties TP_2 and PF_2 . To this end, consider the joint and conditional densities $f(x_i, z)$ and $f(x_i|z)$, respectively. First, assuming CI. Then, with $E(X_i) = E(Z) = 0$, the covariance between X_i and Z equals

$$E(X_i Z) = a_i E(Z^2) + E(U_i Z) = a_i \sigma^2 > 0.$$

This implies that $f(x_i, z)$ is TP_2 (Karlin & Rinott, 1983). Second, because U_i is normally distributed, so is the conditional density $f(x_i|z)$. Consequently, $f(x_i|z)$ has a PF_2 density (Efron, 1965). Note that, for strictly positive densities, if $f(x_i, z)$ is TP_2 , then $f(x_i|z)$ is TP_2 as a function of (x_i, z) (Holland & Rosenbaum, 1986). Because the TP_2 property implies that the tail distribution $1 - F(x_i|z)$ is non-decreasing in z , it thus follows that the LNF model satisfies the assumptions (i.e., is a special case) of the MH model (Holland & Rosenbaum, 1986).

The following observation is useful for the proof of Theorem 1 below. Assume that $f(x, y|z) > 0$ (as implied by the LNF model). Then, for $z_1 < z_2$, the inequality

$$f(x_2, y_2|z_2)f(x_1, y_1|z_1) \geq f(x_2, y_2|z_1)f(x_1, y_1|z_2) \quad (2)$$

holds, whenever

$$\frac{f(x_2, y_2|z_2)}{f(x_2, y_2|z_1)} \geq \frac{f(x_2, y_1|z_2)}{f(x_2, y_1|z_1)} \geq \frac{f(x_1, y_1|z_2)}{f(x_1, y_1|z_1)}.$$

Hence, (2) holds, if both (a) $f(x, y|z)$ is TP₂ as a function of (y, z) , with $y_1 < y_2$ and $X = x_2$ (fixed), and (b) $f(x, y|z)$ is TP₂ in (x, y) , with $x_1 < x_2$ and $Y = y_1$.

The next result is proven in Ligotvoet (2021), but here adapted to the LNF model.

Theorem 1 *The LNF model implies that $f(s, z)$ is TP₂.*

Proof Suppose that Theorem 1 holds for $n = 2$, with $X_1 = X$ and $X_2 = Y$. The proof of Theorem 1 then follows, by sequentially taking $X = X_1 + \dots + X_{i-1}$ and $Y = X_i$, for $i = 2, \dots, n$. Hence, it is sufficient to show that $f(s, z)$ is TP₂, for $S = X + Y$.

Due to CI, the conditional density of S is given by the convolution $f(s|z) = \int g(x|z)h(s-x|z)dx$. Then, $f(s, z)$ is TP₂, if for any $s_1 < s_2$ and $z_1 < z_2$ it holds that $f(s_2|z_2)f(s_1|z_1) \geq f(s_2|z_1)f(s_1|z_2)$, which yields

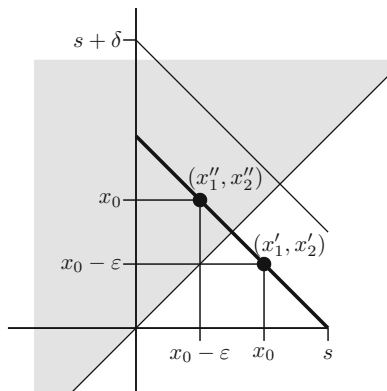
$$\int_{-\infty}^{s_1} \int_{-\infty}^{s_2} g(x_2|z_2)h(s_2-x_2|z_2)g(x_1|z_1)h(s_1-x_1|z_1) \\ - g(x_2|z_1)h(s_2-x_2|z_1)g(x_1|z_2)h(s_1-x_1|z_2) dx_2dx_1 \geq 0. \quad (3)$$

Note that (3) has the form of (2), with $f(x, y|z) = g(x|z)h(s-x|z)$. So, it is sufficient to show that (3) holds for the case that X is constant between z_1 and z_2 , and the case that Y is constant between z_1 and z_2 . Because both cases are symmetric in their arguments, we'll only consider taking Y to be constant at $Z = z$ (for $z_1 \leq z \leq z_2$). For (3), this yields

$$\int_{-\infty}^{s_1} \int_{-\infty}^{s_2} h(s_2-x_2|z)h(s_1-x_1|z) \cdot \\ (g(x_2|z_2)g(x_1|z_1) - g(x_2|z_1)g(x_1|z_2)) dx_2dx_1 \geq 0. \quad (4)$$

Because $g(x, z)$ is TP₂, the function within the integral of (4) has a positive outcome for $x_1 < x_2$ and negative values for $x_1 > x_2$. What remains to be shown is that the density of the area of the integral spanning all $x'_1 > x'_2$ is smaller than the area spanning all $x''_1 < x''_2$ (see Fig. 1 for illustration). Let $x'_1 = x_0$ and $x'_2 = x_0 - \varepsilon$, with $\varepsilon > 0$, and accordingly $x''_1 = x_0 - \varepsilon$ and $x''_2 = x_0$. This yields a one-to-one (injective) mapping of each pair (x'_1, x'_2) that yields negative values in (4) to (x''_1, x''_2) , as shown in Fig. 1. Also, let $s_1 = s$ and $s_2 = s + \delta$, with $\delta > 0$. Then, it is

Fig. 1 Plot of x_1 and x_2 , with the point (x'_1, x'_2) in the gray areas above the identity yielding positive values of (4)



sufficient to show that for all values $x_0, \varepsilon, s, \delta$, and $z_1 < z_2$,

$$(g(x_0|z_2)g(x_0 - \varepsilon|z_1) - g(x_0|z_1)g(x_0 - \varepsilon|z_2)) \cdot$$

$$(h(s - x_0 + \varepsilon|z)h(s - x_0 + \delta|z) - h(s - x_0 + \varepsilon + \delta|z)h(s - x_0|z)) \geq 0.$$

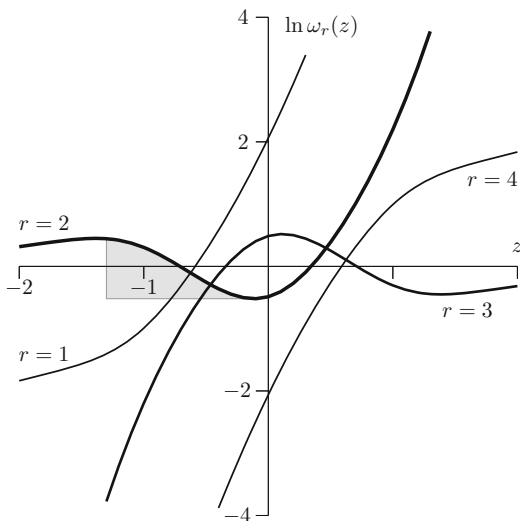
The first part of (5) in parentheses is non-negative, because $g(x, z)$ is TP₂. The second part in parentheses is also non-negative, because $h(y|z)$ is PF₂. This can be seen by taking $x_1 = s - x_0, x_2 = s - x_0 + \varepsilon, y_1 = -\delta$, and $y_2 = 0$ in (1). □

2.2 The Sum Score After Discretization of Variables

Theorem 1 shows that under the LNF model $f(s, z)$ is TP₂ for the sum score $S = X_1 + \dots + X_n$. Next, a discretization is considered, whereby we denote the sum of the discretized variables as $R = V_1 + \dots + V_n$. As mentioned earlier, the LNF model satisfies the MH model assumptions. For the special case when a dichotomization is applied to all variables (i.e., $m_i = m = 1$), the LNF reduces to Mokken’s MH model for binary variables, which has been shown to imply a stochastic ordering of the latent variable by the sum score (Ghurye & Wallace, 1959; Grayson, 1988; Huynh, 1994; Ünlü, 2008). For $m_i > 2$, however, the MH model does not imply a stochastic ordering of the latent variable by the sum score (Hemker et al., 1996, 1997). The next example also shows that $f(r, z)$ need not be TP₂, after discretizing the variables X_1, \dots, X_n obtained under the LNF model.

Example 1 Consider the LNF for $n = 2$ variables, with $a_1 = a_2 = 1, \sigma^2 = 1$, and $\sigma_1^2 = 2/5$ and $\sigma_2^2 = 5/2$. Also, let $V_i = \phi_i(X_i; -1/2, 1/2)$, for $i = 1, 2$ (i.e., $m_i = m = 2$), and $R = V_1 + V_2$. Further, define the log-odds $\ln \omega_r(z) = \ln f(r|z) - \ln f(r - 1|z)$, for $r = 1, \dots, 4$, which are non-decreasing in z , whenever $f(r, z)$ is TP₂. Figure 2 shows that for $r = 2, 3$, the log-odds are decreasing (in violation of the stochastic property). For $r = 2$, the gray area in Fig. 2 shows the decrease in

Fig. 2 Log-odds for the sum scores as a function of z , showing a violation (decrease; e.g., gray area) of the stochastic ordering of Z by $R = V_1 + V_2$



log-odds between $z = -1.37$ and $z = -0.10$, indicating that for two subjects with these factor scores, the subject with the (higher) factor score $z = -0.10$ is about 2.5 times more likely to obtain the lower sum score than the subject that has a factor score that is more than one standard deviation lower (i.e., a substantial violation).

3 Discussion

The property $f(s, z)$ is TP_2 was proposed as a minimal requirement for the use of the sum $S = X_1 + \dots + X_n$. This property is less restrictive than the tau-equivalence requirement from classical test theory (Lord & Novick, 1968; McNeish & Wolf, 2020), but is also limited to ordinal inferences. Theorem 1 implies that the confirmation of the LNF model justifies ordinal inferences about the latent factor based on the sum score. For applications that require more than mere ordinal inferences, the use of the estimated factor score may be more advantageous.

The discretization of variables obtained under the LNF model, not only jeopardizes the normality assumption, but also has implications for the practical use of the sum score. The extent to which the stochastic ordering property by the sum score is violated, in a practical sense, will depend on the number of items variables of the test, as well as the number of categories resulting from the discretization. Simulation studies may further address this issue.

Instead of assuming a normal distribution to underlie the observed discrete ordinal response data, alternative approaches for analyzing these data impose restrictions on the cumulative distributions similar to Samejima’s (1969) *graded*

response model. Jöreskog and Moustaki (2001) and Takane and De Leeuw (1987) showed that the normal ogive model for graded responses is formally equivalent to the LNF model that assumes a normal distribution to underlie the ordinal responses. The difference between these models is that, for the graded response model, the conditional response distributions is discretized prior to taking the marginal across the latent factor, whereas for the LNF model the discretization is applied (afterwards) to the marginal distribution (Takane & De Leeuw, 1987, p. 397). In the latter case, the discretization may invalidate the property $f(s, z)$ is TP₂, when applying the LNF to discrete ordinal response data. That the graded response model does not imply this property was already shown by Hemker et al. (1996, 1997).

To conclude, the use of the sum score, albeit practical, is not what most models are designed for. The applied researcher should realize that an ordering on a latent variable by the sum score is not something that can be simply assumed to hold. If the applied researcher has a model that accurately describes the response data, it might generally be best to rely on the model estimates, rather than using the sum scores. And if a transformation of the data is deemed necessary, the validity of the model will need to be reassessed.

References

- Andrich, D. (1995a). Models for measurement, precision, and the nondichotomization of graded responses. *Psychometrika*, *60*(1), 7–26.
- Andrich, D. (1995b). Further remarks on nondichotomization of graded responses. *Psychometrika*, *60*(1), 37–46.
- Efron, B. (1965). Increasing properties of Pólya frequency functions. *The Annals of Mathematical Statistics*, *36*(1), 272–279.
- Ellis, J. L. (2015). MTP2 and partial correlations in monotone higher-order factor models. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, W. C. Wang (Eds.), *Quantitative psychology research* (pp. 261–272). Berlin: Springer.
- Ghurye, S. G., & Wallace, D. L. (1959). A convolutive class of monotone likelihood ratio families. *The Annals of Mathematical Statistics*, *30*(4), 1158–1164.
- Grayson, D. A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. *Psychometrika*, *53*(3), 383–392.
- Hemker, B. T. (2001). Reversibility revisited and other comparisons of three types of polytomous IRT models. In A. Boomsma, M. A. J. Van Duijn, T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 277–296). Berlin: Springer.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1996). Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, *61*(4), 679–693.
- Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*(3), 331–347.
- Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *The Annals of Statistics*, *14*(4), 1523–1543.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. *Psychometrika*, *59*(1), 77–79.
- Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133.

- Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387.
- Karlin, S. (1968). *Total positivity*. Redwood City: Stanford University Press.
- Karlin, S., & Rinott, Y. (1980). Classes of orderings of measures and related correlation inequalities, I. Multivariate totally positive distributions. *Journal of Multivariate Analysis*, *10*(4), 467–498.
- Karlin, S., & Rinott, Y. (1983). M-matrices as covariance matrices of multinormal distributions. *Linear Algebra and its Applications*, *52–53*, 419–438.
- Ligtvoet, R. (2012). An isotonic partial credit model for ordering subjects on the basis of their sum scores. *Psychometrika*, *77*(3), 479–494.
- Ligtvoet, R. (2015). A test for using the sum score to obtain a stochastic ordering of subjects. *Journal of Multivariate Analysis*, *133*, 136–139.
- Ligtvoet, R. (2021). Conditional TP₂ distributions of sums for latent ordinal inferences. Manuscript submitted for publication.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Boston: Addison-Wesley.
- McNeish, D., & Wolf, M. G. (2020). Thinking twice about sum scores. *Behavior Research Methods*, *52*(6), 2287–2305.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.
- Molenaar, I. W. (1997). Nonparametric models for polytomous responses. In *Handbook of modern item response theory* (pp. 369–380). Berlin: Springer.
- Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.
- Roskam, E. E. (1995). Graded responses and joining categories: A rejoinder to Andrich's "Models for measurement, precision and non-dichotomization of graded responses". *Psychometrika*, *60*(1), 27–35.
- Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika*, *34*, 1–97
- Schoenberg, I. J. (1951). On pólya frequency functions. I. The totally positive functions and their Laplace transforms. *Journal d'Analyse Mathématique*, *1*(1), 331–374.
- Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 137–158.
- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
- Ünlü, A. (2008). A note on monotone likelihood ratio of the total score variable in unidimensional item response theory. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 179–187.