# Sum scores in questionnaires, some asymptotic results ~~and partial identification calculations~~

Steffen Grønneberg

BI Norwegian Business School

December 4th, 2023

# Contents

# Contents

- Illustration: The five factor model of personality.

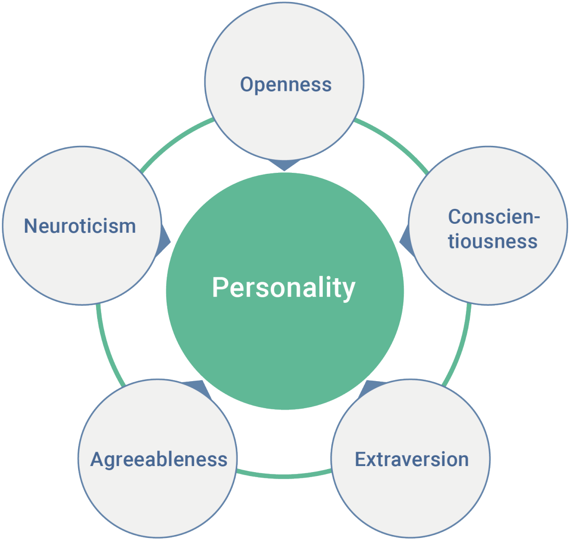



Figure: Big Five Personality model

| | Disagree | | Neutral | | Agree |
|---|---|---|---|---|---|
| I am the life of the party. | | | | | |
| I feel little concern for others. | | | | | |
| I am always prepared. | | | | | |
| I get stressed out easily. | | | | | |
| I have a rich vocabulary. | | | | | |
| I don't talk a lot. | | | | | |
| I am interested in people. | | | | | |
| I leave my belongings around. | | | | | |
| I am relaxed most of the time. | | | | | |
| I have difficulty understanding abstract ideas. | | | | | |
| I feel comfortable around people. | | | | | |
| I insult people. | | | | | |
| I pay attention to details. | | | | | |

Figure: Extract from a big five questionnaire

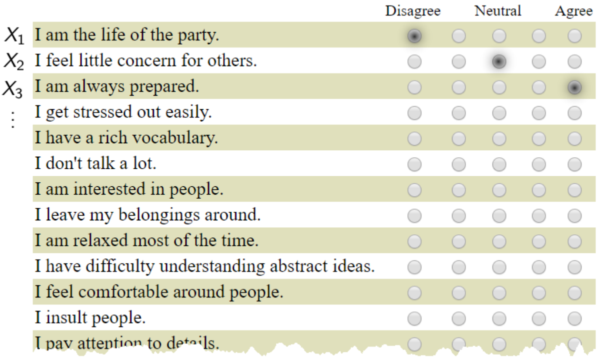



Figure: Extract from a big five questionnaire

- Usual to integer encode questions. The first three answers are therefore:

$$X_1 = 1, \qquad X_2 = 3, \qquad X_3 = 5, \qquad \dots$$

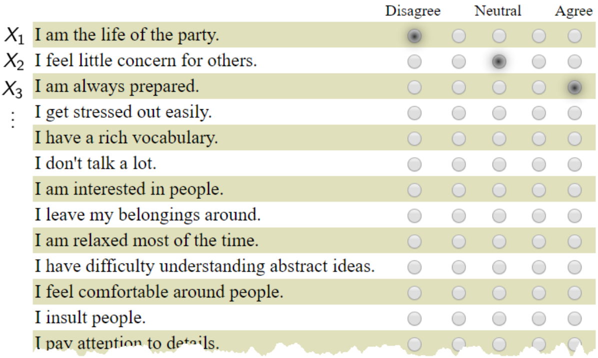



Figure: Extract from a big five questionnaire

- Usual to integer encode questions. The first three answers are therefore:

$$X_1 = 1, \qquad X_2 = 3, \qquad X_3 = 5, \qquad \ldots$$

- This is really $X_{i,1}, X_{i,2}, X_{i,3}, \ldots$. We only consider one person in the notation.

- Ordinal methods exists, but they make strong distributional assumptions which cannot easily be weakened (Moss & Grønneberg, 2023).
- In practical work, two dominant ways:
  1. Treat the integer encoded data as continuous.
  2. Take sum scores (today's topic)
- The consensus appears to be that this works well, with few assumptions and well-developed tools (e.g. goodness of fit tests).

- Ordinal methods exists, but they make strong distributional assumptions which cannot easily be weakened (Moss & Grønneberg, 2023).
- In practical work, two dominant ways:
  1. Treat the integer encoded data as continuous.
  2. Take sum scores (today's topic)
- The consensus appears to be that this works well, with few assumptions and well-developed tools (e.g. goodness of fit tests).
- However:
- Under very special cases, (1) can work but often does not, and is usually inconsistent (Foldnes & Grønneberg, 2021; Grønneberg & Foldnes, 2022).
- Today's conclusion: Also (2) can work as intended in special cases, but usually not.

A question is called an item.

Each item is designed to measure just one of the five factors (e.g. "I am the life of the party" measures extraversion)

Some of the items measure *Openness*. Jointly, they form a *scale* for the *latent variable* openness.

A question is called an item.

Each item is designed to measure just one of the five factors (e.g. "I am the life of the party" measures extraversion)

Some of the items measure *Openness*. Jointly, they form a *scale* for the *latent variable* openness.

The sum of the integer encoded items is your openness-score.

When analyzing sum scores, their empirically standardized versions are supposed to approximate the latent variable measured by the scale.

- Consider an ordinal scale $X = (X_1, \ldots, X_d)'$ influenced by a latent variable $\xi$ (e.g. openness). $\xi$ is never observed, only $X$
- For notational simplicity: Each item is binary (Outcome: agree/disagree or right/wrong)
- Assumption (A non-parametric (NP) factor structure): Conditional on $\xi$, the items $X_j$ are independent.

- Consider an ordinal scale $X = (X_1, \ldots, X_d)'$ influenced by a latent variable $\xi$ (e.g. openness). $\xi$ is never observed, only $X$
- For notational simplicity: Each item is binary (Outcome: agree/disagree or right/wrong)
- Assumption (A non-parametric (NP) factor structure): Conditional on $\xi$, the items $X_j$ are independent.

**Theorem 1**

*For a binary scale with d items that follows a NP factor structure,*

$$\bar{X} = \bar{\pi}_d(\xi) + R_d, \qquad R_d = o_P(1) \quad \text{as } d \to \infty.$$

*where* $\bar{\pi}_d(\xi) = d^{-1} \sum_{j=1}^{d} P(X_j = 1|\xi)$

- Consider an ordinal scale $X = (X_1, \ldots, X_d)'$ influenced by a latent variable $\xi$ (e.g. openness). $\xi$ is never observed, only $X$
- For notational simplicity: Each item is binary (Outcome: agree/disagree or right/wrong)
- Assumption (A non-parametric (NP) factor structure): Conditional on $\xi$, the items $X_j$ are independent.

**Theorem 1**

*For a binary scale with $d$ items that follows a NP factor structure,*

$$\bar{X} = \bar{\pi}_d(\xi) + R_d, \qquad R_d = o_P(1) \quad \text{as } d \to \infty.$$

*where* $\bar{\pi}_d(\xi) = d^{-1} \sum_{j=1}^{d} P(X_j = 1|\xi)$

- Unless $\bar{\pi}$ is linear (with positive slope), standardized sum scores will not approximate the standardized $\xi$.
- $\bar{\pi}_d(\xi)$ need not even converge without more assumptions.
- We now prove Theorem 1 through a simple probability argument.

## Lemma 1 (A stochastic representation)

*Let $U_1, \ldots, U_d$ be IID $U[0,1]$ and independent of $\xi$.*
*For a binary scale $X_1, \ldots, X_d$ with a NP factor structure, we have that $X$ has the same distribution as if*

$$X_j = I\{U_j \leq \pi_j(\xi)\}, \qquad \pi_j(\xi) := P(X_j = 1\}|\xi),$$

## Lemma 1 (A stochastic representation)

*Let $U_1, \ldots, U_d$ be IID $U[0,1]$ and independent of $\xi$.*
*For a binary scale $X_1, \ldots, X_d$ with a NP factor structure, we have that $X$ has the same distribution as if*

$$X_j = I\{U_j \leq \pi_j(\xi)\}, \qquad \pi_j(\xi) := P(X_j = 1\}|\xi),$$

## Proof.

- Recall: Conditional on $\xi$, the binary items $X_j$ are independent. For $x_1, \ldots, x_d \in \{0,1\}$ we have

$$P(\cap_{j=1}^d \{X_j = x_j\}) = \mathbb{E}P(\cap_{j=1}^d \{X_j = x_j\}|\xi) = \mathbb{E}\prod_{j=1}^d P(X_j = x_j|\xi)$$

$$= \mathbb{E}\prod_{j=1}^d \pi_j(\xi)^{x_j}(1 - \pi_j(\xi))^{1-x_j}.$$

## Lemma 1 (A stochastic representation)

*Let $U_1, \ldots, U_d$ be IID $U[0,1]$ and independent of $\xi$.*
*For a binary scale $X_1, \ldots, X_d$ with a NP factor structure, we have that $X$ has the same distribution as if*

$$X_j = I\{U_j \leq \pi_j(\xi)\}, \qquad \pi_j(\xi) := P(X_j = 1\}|\xi),$$

## Proof.

- Recall: Conditional on $\xi$, the binary items $X_j$ are independent. For $x_1, \ldots, x_d \in \{0,1\}$ we have

$$P(\cap_{j=1}^d \{X_j = x_j\}) = \mathbb{E}P(\cap_{j=1}^d \{X_j = x_j\}|\xi) = \mathbb{E}\prod_{j=1}^d P(X_j = x_j|\xi)$$
$$= \mathbb{E}\prod_{j=1}^d \pi_j(\xi)^{x_j}(1-\pi_j(\xi))^{1-x_j}.$$

- If $X_j = I\{U_j \leq \pi_j(\xi)\}$ conditional independence holds, and $P(\{X_j = x_j\}|\xi) = \pi_j(\xi)^{x_j}(1-\pi_j(\xi))^{1-x_j}$ as required.

$\square$

- Now for the proof of Theorem 1: Recall

$$X_j = I\{U_j \le \pi_j(\xi)\}, \qquad \pi_j(\xi) = P(X_j = 1|\xi) \qquad i = 1, 2, \ldots, d.$$

where $U_1, \ldots, U_d$ IID $U[0,1]$ and independent to $\xi$.

- Then

$$\bar{X} = \frac{1}{d}\sum_{j=1}^{d} I\{U_j \le \pi_j(\xi)\}$$

is an average over independent variables except for the non-varying $\xi$.

- Now for the proof of Theorem 1: Recall

$$X_j = I\{U_j \le \pi_j(\xi)\}, \qquad \pi_j(\xi) = P(X_j = 1|\xi) \qquad i = 1, 2, \ldots, d.$$

  where $U_1, \ldots, U_d$ IID $U[0,1]$ and independent to $\xi$.
- Then

$$\bar{X} = \frac{1}{d} \sum_{j=1}^{d} I\{U_j \le \pi_j(\xi)\}$$

  is an average over independent variables except for the non-varying $\xi$.
- The average over the independent variables ought to be less and less random, and $\bar{X}$ ought to approximate $\mathbb{E}[\bar{X}|\xi]$, where

$$\mathbb{E}[\bar{X}|\xi] = \frac{1}{d} \sum_{j=1}^{d} \mathbb{E}[I\{U_j \le \pi_j(\xi)\}|\xi] = \frac{1}{d} \sum_{j=1}^{d} \pi_j(\xi) = \bar{\pi}_d(\xi)$$

  which also equals $\mathbb{E}_U \bar{X}$ (expectation with respect only to $U_1, \ldots, U_d$).

- For $\epsilon > 0$, Chebyshev's inequality gives

$$
\begin{aligned}
&P(|\bar{X} - \bar{\pi}_d(\xi)| > \epsilon) = \mathbb{E}P(|\bar{X} - \bar{\pi}_d(\xi)| > \epsilon|\xi) \\
&\overset{(a)}{=} \mathbb{E}_\xi P_U(P(|\bar{X} - \bar{\pi}_d(\xi)| > \epsilon) \\
&\leq \mathbb{E}_\xi \epsilon^{-2} \mathrm{Var}_U \bar{X} \overset{(b)}{=} \epsilon^{-2} \mathbb{E}_\xi d^{-2} \sum_{j=1}^{d} \mathrm{Var}_U I\{U_j \leq \pi_j(\xi)\} \\
&\leq \epsilon^{-2} \mathbb{E}_\xi d^{-2} \sum_{j=1}^{d} 1/4 \\
&= \epsilon^{-2} d^{-1}/4 \to 0.
\end{aligned}
$$

*(a)* $U$ is independent to $\xi$. *(b)* $U_1, \ldots, U_d$ is IID

- Therefore, $\bar{X} = \bar{\pi}_d(\xi) + R_d$ where $R_d = o_P(1)$ as $d$ increases.

# Contents

- Work-horse model in psychometrics: Confirmatory factor models (CFA). For $p$ factors $\xi = (\xi_1, \ldots, \xi_p)'$ (here: $p = 5$ for big five), and $d$ questions ($d > p$), we observe <u>for each person</u>

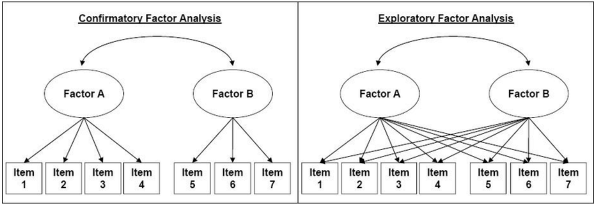
$$X = (X_1, \ldots, X_d)' = \mu + \underbrace{\Lambda}_{d \times p} \underbrace{\xi}_{p \times 1} + \epsilon$$

- Basic assumptions: $\xi, \epsilon$ are uncorrelated, $\mathbb{E}\epsilon = 0$.

- Work-horse model in psychometrics: Confirmatory factor models (CFA). For $p$ factors $\xi = (\xi_1, \dots, \xi_p)'$ (here: $p = 5$ for big five), and $d$ questions ($d > p$), we observe for each person

$$X = (X_1, \dots, X_d)' = \mu + \underbrace{\Lambda}_{d \times p} \underbrace{\xi}_{p \times 1} + \epsilon$$

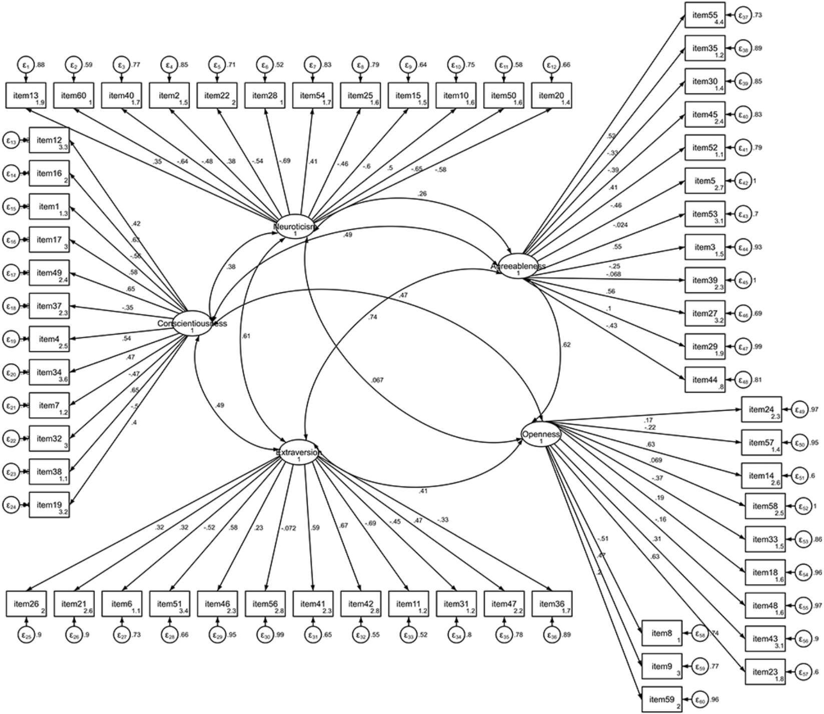
- Basic assumptions: $\xi, \epsilon$ are uncorrelated, $\mathbb{E}\epsilon = 0$.
- Confirmatory factor models: $\Lambda$ is an identified parameter from fixing many elements to zero. Typically, each item $X_j$ is influenced by just one factor, say, $X_j = \mu_j + \lambda_j \xi_1 + \epsilon_j$




- Some elements of $\epsilon$ may be correlated, but not "too many", as we otherwise loose identification.

- CFAs were developed for *continuous data*.
- Historically, sum scores were taken as a foundational data-point, and inputted into CFAs.
- This makes sense:
    1. With "enough" items ($d$), the sum scores are "close to continuous".
    2. Sum scores were formulated using substantive knowledge in psychology. The critique of this talk then does not apply.

- CFAs were developed for *continuous data*.
- Historically, sum scores were taken as a foundational data-point, and inputted into CFAs.
- This makes sense:
  1. With "enough" items ($d$), the sum scores are "close to continuous".
  2. Sum scores were formulated using substantive knowledge in psychology. The critique of this talk then does not apply.
- Ordinal scales are now developed using CFAs on the item level (the ordinal observations).
- Under a CFA, sum scores are well behaved, as we shortly see.
- But ordinal data, except very under limited circumstances, will not follow a CFA, invalidating this argument.



Neuroticism
Conscientiousness
Agreeableness
Extraversion
Openness

- If $X_1, \ldots, X_K$ follows a one-factor model ("unidimensional" factor model), then

$$X_j = \mu_j + \lambda_j \xi_1 + \epsilon_j,$$

where $\mathbb{E}\epsilon_j = 0, \mathrm{Cov}(\epsilon_j, \xi_1) = 0$, and where $\mathrm{Cov}(\epsilon_j, \epsilon_k) = 0$ for "most" pairs $k \neq j$.

- If $X_1, \ldots, X_K$ follows a one-factor model ("unidimensional" factor model), then
$$X_j = \mu_j + \lambda_j \xi_1 + \epsilon_j,$$
where $\mathbb{E}\epsilon_j = 0, \mathrm{Cov}(\epsilon_j, \xi_1) = 0$, and where $\mathrm{Cov}(\epsilon_j, \epsilon_k) = 0$ for "most" pairs $k \neq j$.
- The mean score is
$$\bar{X} = \frac{1}{K} \sum_{j=1}^{K} X_j = \bar{\mu} + \bar{\lambda}\xi + \bar{\epsilon},$$
- Therefore
$$\bar{X} \approx \bar{\mu} + \bar{\lambda}\xi$$
given reasonable bounds on $\mathrm{Cov}(\epsilon_j, \epsilon_k)$.
- In the ordinal case, we have in contrast seen $\bar{X} \approx \bar{\pi}_d(\xi)$. So what goes wrong?

- Recall

$$X_j = I\{U_j \leq \pi_j(\xi)\}, \qquad \pi_j(\xi) = P(X_j = 1|\xi) \qquad i = 1, 2, \ldots, d,$$

where $U_1, \ldots, U_d$ IID $U[0,1]$ and all independent to $\xi$.

- Recall

$$X_j = I\{U_j \le \pi_j(\xi)\}, \qquad \pi_j(\xi) = P(X_j = 1|\xi) \qquad i = 1, 2, \dots, d,$$

where $U_1, \dots, U_d$ IID $U[0,1]$ and all independent to $\xi$.

- Suppose $\xi$ is univariate (one factor). Let $\lambda_j = \text{Cov}(\xi, X_j)(Var\xi)^{-1}$ and $\mu_j = \mathbb{E}X_j - \lambda_j \mathbb{E}\xi$. Then

$$\epsilon_j := X_j - (\mu_j + \lambda_j \xi) \quad \text{fulfills } \mathbb{E}\epsilon = 0, \text{Cov}(\epsilon, \xi) = 0.$$

- Hence $X_j = \mu_j + \lambda_j \xi + \epsilon_j$ fulfills a confirmatory factor model of sorts. However, notice $\mathbb{E}[X_j|\xi] = \pi_j(\xi)$ is not assumed to be linear.
- Can show: $\epsilon_1, \dots, \epsilon_d$ can all be correlated. Then the confirmatory factor model is not identified.
- Ordinal variables will then not follow a confirmatory factor model (except when $\pi_j$ is linear!).

# Contents

- There are also factor models designed specifically for ordinal data.
- For a one-factor model, all such models are equivalent to threshold type models originating from Pearson (1900):

$$X_j = I\{\lambda_j \xi + \epsilon_j \geq \tau_j\}, \qquad \tau_j \text{ a number, } \epsilon_j \text{ independent to } \xi.$$

  It follows a NP factor model.
- Gives $\pi_j(\xi) = P_\epsilon(\lambda_j \xi + \epsilon_j \geq \tau_j) = 1 - F_{\epsilon_j}(\tau_j - \lambda_j \xi)$.

- There are also factor models designed specifically for ordinal data.
- For a one-factor model, all such models are equivalent to threshold type models originating from Pearson (1900):

$$X_j = I\{\lambda_j \xi + \epsilon_j \geq \tau_j\}, \qquad \tau_j \text{ a number, } \epsilon_j \text{ independent to } \xi.$$

  It follows a NP factor model.
- Gives $\pi_j(\xi) = P_\epsilon(\lambda_j \xi + \epsilon_j \geq \tau_j) = 1 - F_{\epsilon_j}(\tau_j - \lambda_j \xi)$.
- If e.g. $\epsilon_j \sim N(0, \psi_j^2)$, then

$$\bar{\pi}_d(x) = 1 - d^{-1} \sum_{j=1}^{d} \Phi((\tau_j - \lambda_j x)/\psi_j),$$

  which is not linear.
- If $\lambda_j > 0$, then $\bar{\pi}_d$ is invertible. If the parameters are identified, $\hat{\bar{\pi}}_d^{-1}(\bar{X}) \approx \xi$. (Appears to be a new ordinal factor score)

- There are also factor models designed specifically for ordinal data.
- For a one-factor model, all such models are equivalent to threshold type models originating from Pearson (1900):

$$X_j = I\{\lambda_j \xi + \epsilon_j \geq \tau_j\}, \qquad \tau_j \text{ a number, } \epsilon_j \text{ independent to } \xi.$$

  It follows a NP factor model.
- Gives $\pi_j(\xi) = P_\epsilon(\lambda_j \xi + \epsilon_j \geq \tau_j) = 1 - F_{\epsilon_j}(\tau_j - \lambda_j \xi)$.
- If e.g. $\epsilon_j \sim N(0, \psi_j^2)$, then

$$\bar{\pi}_d(x) = 1 - d^{-1} \sum_{j=1}^{d} \Phi((\tau_j - \lambda_j x)/\psi_j),$$

  which is not linear.
- If $\lambda_j > 0$, then $\bar{\pi}_d$ is invertible. If the parameters are identified, $\hat{\bar{\pi}}_d^{-1}(\bar{X}) \approx \xi$. (Appears to be a new ordinal factor score)
- To justify current empirical practice, we require linearity of $\bar{\pi}_d$.
- This is implied by the linearity of $\pi_j(x) = P(X_j = 1|\xi)$.
- (Notice $\pi_j(\xi) = 1 - F_{\epsilon_j}(\tau_j - \lambda_j \xi)$ is linear if $\epsilon_j$ uniform.)

If $\xi$ is a random variable, and $\pi_j(x) = \mu_j + \lambda_j x$ for $\lambda_j > 0$, let's say the NP factor structure is unidimensional and linear.

## Lemma 2

*Suppose given a binary scale $X$ following a unidimensional linear NP factor structure. Then $P(\xi \in [\max_j l_j, \min_j u_j]) = 1$ where $l_j = -\mu_j/\lambda_j, u_j = (1-\mu_j)/\lambda_j$, and*

$$X_j = I\{U_j \leq \mu_j + \lambda_j \xi\}$$

*where $U_1, \ldots, U_d$ are IID $U[0,1]$ and independent to $\xi$.*

## Proof.

- Notice that $\mu_j + \lambda_j x = \pi_j(x) = P(X_j = 1 | \xi = x) \in [0,1]$ for all $x$ attainable by $\xi$. Therefore, the support of $\xi$ is contained in $\cap_{j=1}^d \{x : 0 \leq \mu_j + \lambda_j x \leq 1\} = \cap_{j=1}^d \{x : -\mu_j \leq x \leq (1-\mu_j)/\lambda_j\} = [\max_j(-\mu_j), \min_j(1-\mu_j)/\lambda_j]$.
- The stochastic representation then gives $X_j = I\{U_j \leq \pi_j(\xi)\} = I\{U_j \leq \mu_j + \lambda_j \xi\}$

$\square$

## Theorem 2

*A binary scale $X$ following a unidimensional linear NP factor structure also follows a unidimensional confirmatory factor structure.*

## Proof.

- by Lemma 2, $\mathbb{E}[X_j|\xi] = \mathbb{E}[I\{U_j \leq \mu_j + \lambda_j\xi\}|\xi] = \mu_j + \lambda_j\xi$. Therefore,

$$X_j = \mu_j + \lambda_j\xi + \epsilon_j, \qquad \epsilon_j := X_j - \mathbb{E}[X_j|\xi].$$

- Clearly $\mathbb{E}\epsilon_j = 0, \mathrm{Cov}(\epsilon_j, \xi) = 0$.
- Let $i \neq j$. Then $\epsilon_j = I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi]$ and $\epsilon_j = I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi]$ are conditionally independent and conditionally zero mean given $\xi$. Gives $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

$\square$

From the "continuous argument": Sum scores of CFAs also follow a CFA: $\bar{X} = \bar{\mu} + \bar{\lambda}\xi + \bar{\epsilon}$. So also sum scores of the whole or *parts of the scale* follow a CFA.

## Corollary 1

*Suppose $X$ is a binary scale following a unidimensional linear NP factor structure. Then $\bar{X} = \bar{\mu} + \bar{\lambda}\xi + r_d$ where for any $c > 0$, $P(|r_d| > c) \leq 4\exp(1 - 2dc^2)$.*

Consistency follows from Theorem 1. Corollary 1 gives a concentration bound with fixed constants.

## Proof.

- Notice $r_d = d^{-1}\sum_{j=1}^{d}\epsilon_j = d^{-1}\sum_{j=1}^{d} I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi] = d^{-1}\sum_{j=1}^{d}[I\{(U_j - \mu_j)/\lambda_j \leq \xi\} - P_U((U_j - \mu_j)/\lambda_j \leq \xi)] = \mathbb{F}_d(\xi) - \bar{F}_d(\xi)$ where $\mathbb{F}_d$ is the empirical distribution of the independent sequence $((U_j - \mu_j)/\lambda_j)$, and $\bar{F}_d(x) = d^{-1}\sum_{j=1}^{d} P_U((U_j - \mu_j)/\lambda_j \leq x)$.

## Corollary 1

*Suppose $X$ is a binary scale following a unidimensional linear NP factor structure. Then $\bar{X} = \bar{\mu} + \bar{\lambda}\xi + r_d$ where for any $c > 0$, $P(|r_d| > c) \leq 4\exp(1 - 2dc^2)$.*

Consistency follows from Theorem 1. Corollary 1 gives a concentration bound with fixed constants.

## Proof.

- Notice $r_d = d^{-1}\sum_{j=1}^{d}\epsilon_j = d^{-1}\sum_{j=1}^{d} I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi] = d^{-1}\sum_{j=1}^{d}[I\{(U_j - \mu_j)/\lambda_j \leq \xi\} - P_U((U_j - \mu_j)/\lambda_j \leq \xi)] = \mathbb{F}_d(\xi) - \bar{F}_d(\xi)$ where $\mathbb{F}_d$ is the empirical distribution of the independent sequence $((U_j - \mu_j)/\lambda_j)$, and $\bar{F}_d(x) = d^{-1}\sum_{j=1}^{d} P_U((U_j - \mu_j)/\lambda_j \leq x)$.
- By independence, $P(|r_d| > c) = P(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c) = \mathbb{E}_\xi P_U(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c)$

## Corollary 1

*Suppose $X$ is a binary scale following a unidimensional linear NP factor structure. Then $\bar{X} = \bar{\mu} + \bar{\lambda}\xi + r_d$ where for any $c > 0$, $P(|r_d| > c) \leq 4\exp(1 - 2dc^2)$.*

Consistency follows from Theorem 1. Corollary 1 gives a concentration bound with fixed constants.

## Proof.

- Notice $r_d = d^{-1}\sum_{j=1}^{d}\epsilon_j = d^{-1}\sum_{j=1}^{d} I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi] = d^{-1}\sum_{j=1}^{d}[I\{(U_j - \mu_j)/\lambda_j \leq \xi\} - P_U((U_j - \mu_j)/\lambda_j \leq \xi)] = \mathbb{F}_d(\xi) - \bar{F}_d(\xi)$ where $\mathbb{F}_d$ is the empirical distribution of the independent sequence $((U_j - \mu_j)/\lambda_j)$, and $\bar{F}_d(x) = d^{-1}\sum_{j=1}^{d} P_U((U_j - \mu_j)/\lambda_j \leq x)$.
- By independence, $P(|r_d| > c) = P(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c) = \mathbb{E}_\xi P_U(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c) \leq \mathbb{E}_\xi P_U(\sup_x |\mathbb{F}_d(x) - \bar{F}_d(x)| > c) = P_U(\sup_x |\mathbb{F}_d(x) - \bar{F}_d(x)| > c)$

## Corollary 1

*Suppose $X$ is a binary scale following a unidimensional linear NP factor structure. Then $\bar{X} = \bar{\mu} + \bar{\lambda}\xi + r_d$ where for any $c > 0$, $P(|r_d| > c) \leq 4\exp(1 - 2dc^2)$.*

Consistency follows from Theorem 1. Corollary 1 gives a concentration bound with fixed constants.

## Proof.

- Notice $r_d = d^{-1}\sum_{j=1}^d \epsilon_j = d^{-1}\sum_{j=1}^d I\{U_j \leq \mu_j + \lambda_j\xi\} - \mathbb{E}[X_j|\xi] = d^{-1}\sum_{j=1}^d [I\{(U_j - \mu_j)/\lambda_j \leq \xi\} - P_U((U_j - \mu_j)/\lambda_j \leq \xi)] = \mathbb{F}_d(\xi) - \bar{F}_d(\xi)$ where $\mathbb{F}_d$ is the empirical distribution of the independent sequence $((U_j - \mu_j)/\lambda_j)$, and $\bar{F}_d(x) = d^{-1}\sum_{j=1}^d P_U((U_j - \mu_j)/\lambda_j \leq x)$.
- By independence, $P(|r_d| > c) = P(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c) = \mathbb{E}_\xi P_U(|\mathbb{F}_d(\xi) - \bar{F}_d(\xi)| > c) \leq \mathbb{E}_\xi P_U(\sup_x |\mathbb{F}_d(x) - \bar{F}_d(x)| > c) = P_U(\sup_x |\mathbb{F}_d(x) - \bar{F}_d(x)| > c) \leq 4\exp(1 - 2dc^2)$ by Inequality 2 in Chapter 25 in Shorack & Wellner (2009) and Massart (1990).

□

- A binary linear NP one-factor model:

$$X_j = I\{U_j \leq \mu_j + \lambda_j \xi\}$$

is also a binary threshold one-factor model (with highly non-traditional distributional assumptions): $X_j = I\{\tau_j \leq \lambda_j \xi + \epsilon_j\}$ with $\mu = -\tau_j$ and $\epsilon_j = -U_j$.

- A binary linear NP one-factor model:

$$X_j = I\{U_j \leq \mu_j + \lambda_j \xi\}$$

  is also a binary threshold one-factor model (with highly non-traditional distributional assumptions): $X_j = I\{\tau_j \leq \lambda_j \xi + \epsilon_j\}$ with $\mu = -\tau_j$ and $\epsilon_j = -U_j$.
- Traditionally, parameters of such models are identified only under very strong assumptions, such as joint normality.
- Here, parameter identification follows from Theorem 2 (a binary one-factor NP linear model is a confirmatory factor model) using CFA results, as long as $d$ is at least 3.
- Also if we have at least 3 variables measuring $\eta$ such as

$$Y_j = I\{V_j \leq \nu_j + \kappa_j \eta\}$$

  these will jointly form a confirmatory factor model, enabling estimating e.g. the correlation of $\xi$ and $\eta$.
- This is surprising, as identification is unusual under weak assumptions in very similar models.

# Contents

- Likely, the identified assumption set for linearity is never/rarely fulfilled in practical settings, and likely, no test can be made to check this against all alternatives.
- A non-parametric and reasonable assumption is that $\pi_j(x) = P(X_j = 1|\xi = x)$ are all strictly increasing.

- Likely, the identified assumption set for linearity is never/rarely fulfilled in practical settings, and likely, no test can be made to check this against all alternatives.
- A non-parametric and reasonable assumption is that $\pi_j(x) = P(X_j = 1|\xi = x)$ are all strictly increasing.
- Then, for two scales $X, Y$ that follows NP factor structures measuring $\xi$ and $\eta$ respectively, we have

$$\bar{X} = \bar{\pi}_d^X(\xi) + o_P(1), \qquad \bar{Y} = \bar{\pi}_d^Y(\eta) + o_P(1)$$

  approximate strictly increasing marginal transformations of $\xi, \eta$.
- Usually, $\bar{\pi}_d^X, \bar{\pi}_d^Y$ are not identified, meaning the marginals of $\xi, \eta$ will not be identified.
- But copula of $(\bar{\pi}_d^X(\xi), \bar{\pi}_d^Y(\eta))$ equals the copula of $(\xi, \eta)$, and can therefore be estimated non-parametrically.
- This is asymptotic in $d$. For fixed $d$, we can investigate the partial identification question: Which copulas are compatible with the distributions of $X$, $Y$?

Bibliography:

FOLDNES, N. & GRØNNEBERG, S. (2021). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods* .

GRØNNEBERG, S. & FOLDNES, N. (2022). Factor analyzing ordinal items requires substantive knowledge of response marginals. *Psychological Methods* .

MASSART, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability* , 1269–1283.

MOSS, J. & GRØNNEBERG, S. (2023). Partial identification of latent correlations with ordinal data. *Psychometrika* **88**, 241–252.

PEARSON, K. (1900). Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philos. Trans. R. Soc. SA* **196**, 1–47.

SHORACK, G. R. & WELLNER, J. A. (2009). *Empirical processes with applications to statistics*, vol. 59. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM). Originally published in 1986 by Wiley, New York.