A Bayesian Semi-Parametric Copula Factor Model for Estimating Composite Measures of Quality

Stephen Salerno, MS [1] Lili Zhao, PhD [1] Yi Li, PhD [1]

[1]University of Michigan · Department of Biostatistics

Motivating Data Example

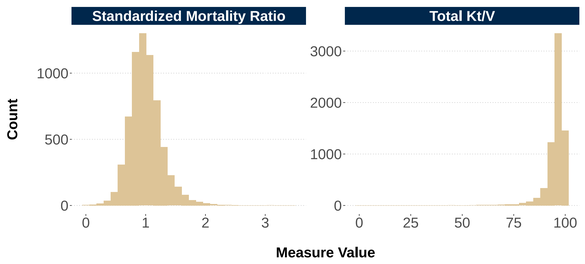
Data Source:

Clinical quality data from eight measures publicly reported on the Medicare Dialysis Facility Compare site in October 2018 for 7,164 Medicare-certified dialysis facilities.

Clinical Measure Domains:

1. Standardized Ratio Measures: Obs./exp. hospitalization, mortality, readmission, and transfusion events
2. Vascular Access Measures: Arteriovenous fistula utilization and long-term catheter utilization
3. Process Measures: Dialysis adequacy (Kt/V) and mineral and bone disorder management (Hypercalcemia)

Figure 1: Example Clinical Quality Measures



Challenges

Publicly reported composite ratings are meant to provide a global summary of quality to help patients and caregivers make informed decisions. However:

- They are often comprised of a set of mixed continuous and discrete clinical quality of care metrics
- The individual measures that make up such ratings have a complex correlation structure
- There is often no measure of uncertainty associated with the estimated ratings
- Potential biases may arise due to missing data in the observed (manifest) variables

Overview of Our Proposal

Factor analysis is a natural choice to estimate a latent 'quality score' from a set of observed quality metrics. We propose a flexible, unifying method that can:

- Accommodate non-normal continuous measurements
- Extend to data with mixed measurement scales
- Handle missing or suppressed manifest variable data
- Maintain interpretability of the factor scores (for ranking)

Rank-Based Bayesian Factor Analysis

- For Y, an $n \times p$ matrix of observed measures, define a Gaussian factor model with $k < p$ latent features as:

$$\mathbf{Y} = \eta\mathbf{\Lambda} + \varepsilon$$

where η is an $n \times k$ matrix of factor scores, $\mathbf{\Lambda}$ is a $k \times p$ matrix of loadings, and ε is an $n \times p$ matrix of idiosyncratic noise

- Considering the Gaussian copula, $F(y_1, \ldots, y_p) = \Phi_p[\Phi^{-1}\{F_i(y_1)\}, \ldots, \Phi^{-1}\{F_p(y_p)\}|\mathbf{C}]$ with correlation matrix $\mathbf{C}$, define:

$$\eta_i \sim N(0, \mathbf{I}), \qquad \mathbf{z}_i|\eta_i \sim N(\mathbf{\Lambda}\eta_i, \mathbf{I}), \qquad y_{ij} = F_j^{-1}\left(\Phi\left(z_{ij}\right)\right)$$

where $\mathbf{z}$ are latent Gaussian variables, F_j^{-1} is the pseudo-inverse of F_j, and $P(\mathbf{Y}|\mathbf{C},\ F_1,\ \ldots,\ F_p) \approx P(\mathbf{Z} \in D(\mathbf{Y})|\mathbf{C})$[1]

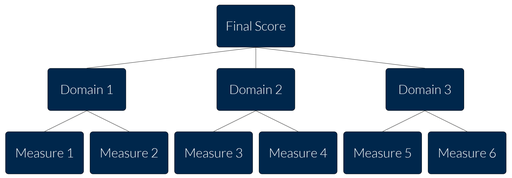
- Thus, only the order statistics of $y_j;\ j = 1, \ldots, p$ are needed to jointly estimate the factor structure underlying mixed data
- Conjugacy allows for closed-form updates to Λ, η, and $\text{diag}(\sigma_{Y_j})$ without the need for Metropolis-Hastings type approaches
- Missing values in Y are initialized in Z and imputed with repeated Gibbs samples from a truncated Normal distribution
- To maintain interpretability of the factor scores, we employ intermediate factor rotations at each Gibbs update which eliminates the need for constraint on the loadings[2]

Simulation Study

A schematic of the data generated is given in Figure 2 below. A shown, we assume a three-level Normal hierarchy:

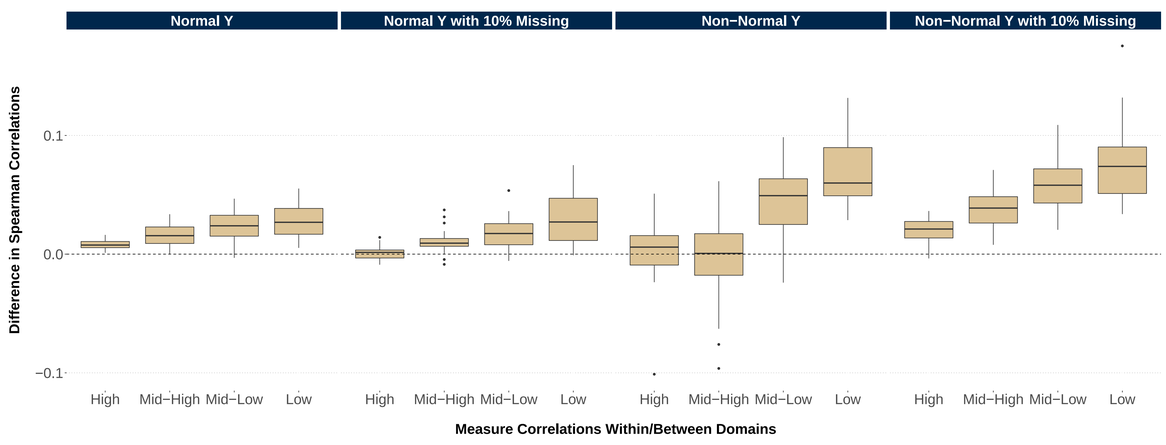
- Data are simulated under assumption that the true rank comes from several underlying correlated measures
- Missingness was set in 10% of the observed data
- $\log(\cdot)$, $\exp(\cdot)$, and discrete transformations were then applied
- Correlations between measures 1-6 were varied systematically by changing the Level 3 variances

Figure 2: Data Generation for Simulation Study



The proposed method performed best in the non-Normal setting with missing data as compared to traditional factor analysis. Results were more comparable in the Normal setting with no missingness. These results are presented in Figure 3:

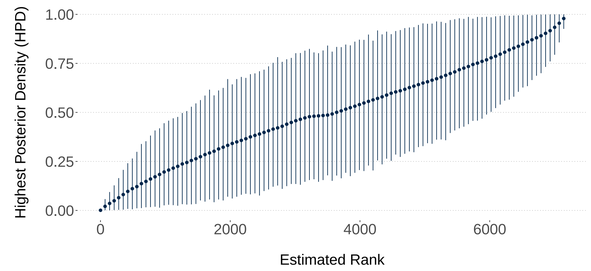
Figure 3: Simulation Results Varying the Measure Correlations



Data Results

- Factor loading estimates were consistent with traditional Frequentist approaches currently employed
- Estimated facility ranks using the proposed method had a 94% correlation with the current ranking methodology
- Estimation of facility rankings from a factor model allow for measures of uncertainty through the posterior distribution

Figure 4: Estimated Ranks and 95% HPD Intervals



Conclusions

- In simulation, the proposed method shows higher correlation with the ground truth ranking when the manifest variables are non-normal or have missingness
- The proposed method allows for quantification of the variability in estimating both the factor loadings and scores and preservation of the underlying rankings
- Consistent results in the data example favor consideration of the proposed method to rank healthcare providers with mixed-type quality data

References

1. Hoff, P.D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1), 265-283.
2. Roková, V. & George, E.I. (2016). Fast Bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516), 1608-1622.

BIOSTATISTICS