Semiparametric Gaussian Copula Regression Modelling of Mixed Data Types (SGCRM)

Debangan Dey, Vadim Zipunnikov

Johns Hopkins Bloomberg School of Public Health

Introduction: Large Health Surveys

Self-reported questionnaire data on dietary preferences, smoking status, drinking status, health status, mobility problem status etc.

Physical activity, sleep, cardiometabolic biomarkers and comorbidities. For example, total activity county (TAC), albumin systolic BP etc.

Hundreds or thousands of binary (0/1), ordinal, truncated, continuous, and categorical variables.

Can we build a flexible modelling framework for joint and mutually consistent conditional modeling of mixed data (binary, ordinal, truncated and continuous)?

Generalized Latent Non-paranormal (GLNPN)

A random vector $X = (X_1, \ldots, X_p)' \sim NPN(0, \Sigma, f)$ if there exist monotonically increasing transformation functions $f = (f_1, \ldots, f_p)$ such that - $Z = f(X) = (f_1(X_1), \ldots, f_p(X_p)) \sim N(0, \Sigma)$ where $\Sigma_{jj} = 1$ for all j.

- c, t, o, b subscript denotes **c**ontinuous, **t**runcated, **o**rdinal, **b**inary respectively

$$X_{cj} = f_{cj}^{-1}(Z_{cj}), 1 \leq j \leq p_c$$

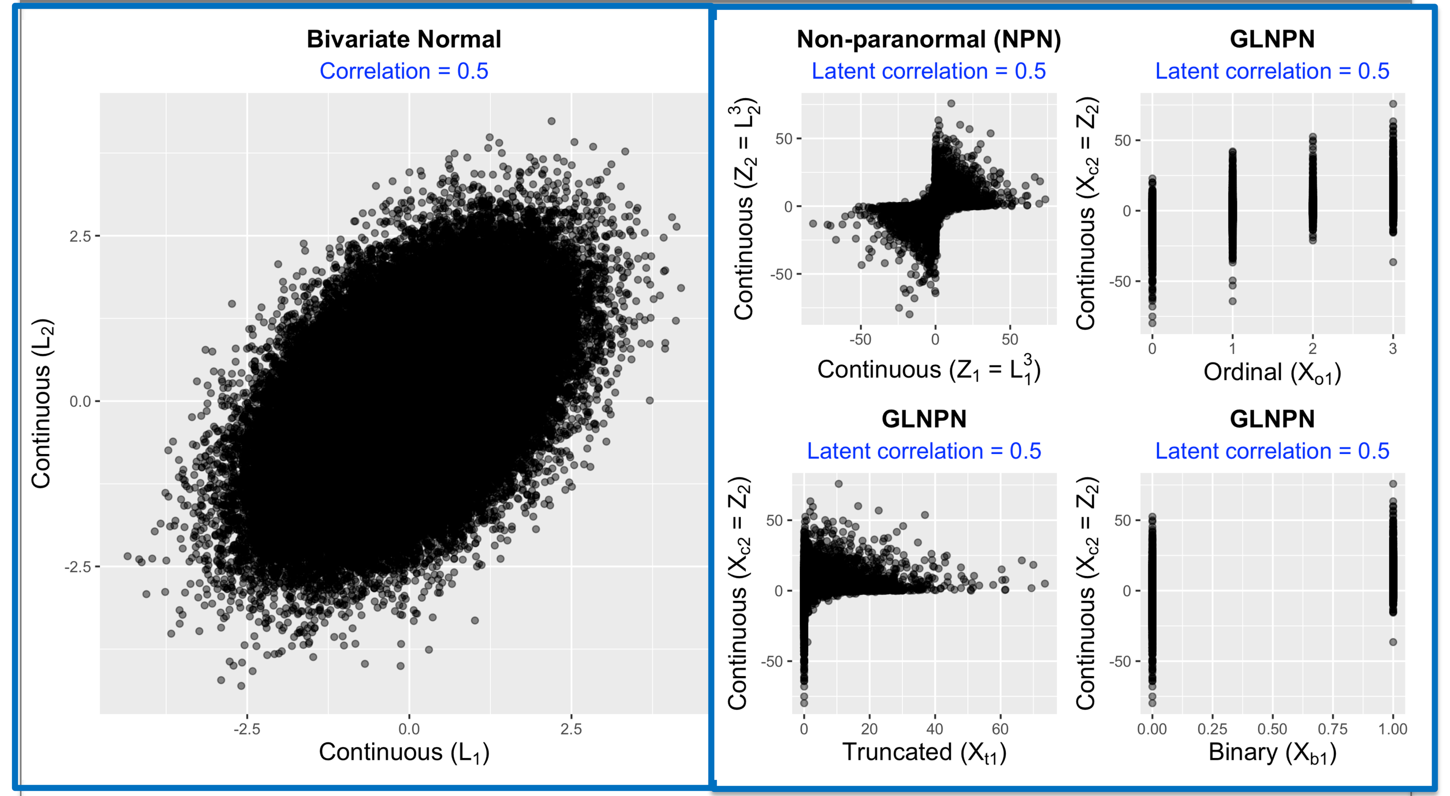
$$X_{tj} = f_{tj}^{-1}(Z_{tj}) I(Z_{tj} > \Delta_{tj}), 1 \leq j \leq p_t$$

$$X_{oj} = \sum_{k=0}^{c_j} k I(\Delta_{ojk} \leq Z_{oj} < \Delta_{oj(k+1)}), 1 \leq j \leq p_o$$

$$X_{bj} = I(Z_{bj} > \Delta_{bj}), 1 \leq j \leq p_b$$

$$Z = (Z_c, Z_t, Z_o, Z_b)' \sim NPN(0, \Sigma, f)$$
$$X = (X_c, X_t, X_o, X_b)' \sim GLNPN(0, \Sigma, f, \Delta)$$

Illustration



Latent → **Observed**

Estimation

- Δ (the set of cutoffs) are estimated through method of moments.
- Kendall's Tau (τ) measures concordance and is calculated as follows -

$$\hat{\tau}_{jk} = \frac{2}{n(n-1)} \sum_{1 \leq i < i' < n} sgn\{(X_{ij} - X_{i'j})(X_{ik} - X_{i'k})\}$$

- $(a - b)$ has the same sign as $(f(a) - f(b))$ for any increasing transformation f. *Makes Kendall's Tau invariant under monotone increasing transformation.*
- Observed Kendall's Tau can be bridged (using known one-to-one transformations) to the corresponding elements of the latent correlation matrix Σ.

Traditional model

For a generalized linear model for mixed data, the assumption looks like –

$$g(E(Y_i|\mathbf{X}_i)) = \sum_{k \in \{c,t,o,b\}} \sum_{j=1}^{p_k} X_{kji}\beta_{kj}$$

where, g() is a pre-specified link function.

SGCRM

$$f_Y(Z_i^Y) = \sum_{k \in \{c,t,o,b\}} \sum_{j=1}^{p_k} f_{kj}(Z_{kji}^X)\beta_{kj} + \epsilon_i, i = 1, \ldots, n$$

- $Z_Y, Z_X \sim NPN(0, \Sigma, f)$ are the latent variables corresponding to Y (outcome) and X (covariates).
- Σ can be partitioned as follows

$$\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX} & \Sigma_{XX} \end{bmatrix}$$

- β estimated as $\Sigma_{XX}^{-1}\Sigma_{XY}$ from the estimated latent correlation matrix.

Proved results

- We present bridging functions for all pairs of types of variables.
- We prove asymptotic normality of our estimators.
- We provide multiple imputation approaches embedded within the framework

Data Analysis (NHANES 2003-06)

$Mortality \sim MobilityProblem + HealthStatus + Education + Age + TAC$

	Probit regression		SGCRM	
	Covariate	Coefficients	Covariate	Coefficients
1	MobilityProblem1	0.281 (0.129, 0.432)	Mobility Problem	0.157(0.053,0.262)
2	Health Status (2)	0.084 (-0.233, 0.421)	Health Status	0.073(0.006,0.201)
3	Health Status (3)	0.291 (-0.01, 0.613)		
4	Health Status (4)	0.299 (-0.022, 0.64)		
5	Health Status (5)	0.711 (0.322, 1.113)		
6	Education (2)	0.237 (0.019, 0.455)	Education	-0.017(-0.028,0.139)
7	Education (3)	0.085 (-0.116, 0.286)		
8	Education (4)	0.086 (-0.128, 0.301)		
9	Education (5)	0.025 (-0.219, 0.266)		
10	Age	0.042 (0.034, 0.05)	Age	0.307(0.252,0.397)
11	scaled TAC	-0.474 (-0.679, -0.276)	TAC	-0.204(-0.28,-0.11)

Conclusions

- Traditional regression modelling:
  - Requires different model formulation (probit, ordinal probit, Gaussian truncated, etc.) for different type of outcome and those are, not, mutually consistent.
  - Requires likelihood, time-consuming and difficult to optimize under certain scenarios. Sensitive to outliers.
  - Have to manually adjust for scales before fitting the model.
- SGC Regression Modelling
  - One joint model and derive mutually consistent conditional model estimates for specific choice of outcome.
  - Estimation procedure (method of moments and rank correlation) makes it robust and fast.
  - Takes care of different scales of variables naturally by construction.

References & Acknowledgements

1. Fan, Jianqing, Han Liu, Yang Ning, and Hui Zou. "High dimensional semiparametric latent graphical model for mixed data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79, no. 2 (2017): 405-421.

2. Yoon, G., Carroll, R. J., and Gaynanova, I. (2018). Sparse semiparametric canonical correlation analysisfor data of mixed types.arXiv preprint arXiv:1807.05274.

3. Dey, D. and Zipunnikov, V. (2019). Connecting population-level auc and latent scale-invariantr2viasemiparametric gaussian copula and rank correlations.arXiv preprint arXiv:1910.14233