Connecting population-level AUC and latent scale-invariant R^2 via Semiparametric Gaussian Copula and rank correlations

Debangan Dey[1], Vadim Zipunnikov[1]
Johns Hopkins Bloomberg School of Public Health[1]

Motivation

- Prediction of binary outcomes is an important problem, for example: 5-year mortality in National Health and Nutrition Examination Survey
- Many pseudo-R^2 proposals to quantify Goodness-of-fit in binary-outcome and continuous-predictor(s) models.
- AUC is the most widely used non-parametric summary. But it has many shortcomings and limitations.
- What is AUC? Do we have intuition about the (0.5, 1) scale? Is 0.8 large (enough)?
- Under complex survey designs (NHANES), AUC requires knowledge of pairwise survey-weights

Contribution

- AUC and three rank statistics (Kendall's Tau, Spearman's rho, Wilcoxon rank-sum) are linearly related.
- AUC and Quadrant correlation are linked under semi-parametric Gaussian Copula assumptions.
- Relating AUC and rank correlation creates more robust estimates.
- We introduce more intuitive latent R-square (R_l^2) scale in analogy to well-understood continuous case.
- How AUC can be calculated using single participant weights.

Notations

- (Y,X) with Y denoting binary and X being continuous.
- M_Y, M_X - the population medians of Y and X. F_Y, F_X are the cdfs of Y and X.
- P(Y=1)=p
- X_1 and X_0 denotes random variables (X|Y=1) and (X|Y=0), respectively.
- The suffix uw and pw means unweighted and pairwise-weighted (product of individual weights)

Definition of Rank Correlations and AUC

$$A = max(P(X_1 > X_0), P(X_1 < X_0)).$$

It's trivial to see that, $P(X_1 > X_0) = 1 - P(X_1 < X_0)$, hence, $A \geq \frac{1}{2}$.

1. **Kendall's Tau**: $r_K = E((Y_i - Y_i')sgn(X_i - X_i'))$,
2. **Wilcoxon's rank-sum statistic**: $W = P(X \leq X_1) - P(X \leq X_0)$
3. **Spearman correlation**. $r_S = 12E[F_Y(Y)F_X(X)] - 3$,
4. **Quadrant correlation**. $r_Q = E[sgn((Y - M_Y)(X - M_X))]$,.

where (Y_i, X_i) and $(Y_i^{'}, X_i^{'})$ are two independent copies following the same bi-variate distribution.

Relation between AUC and Rank Correlations

$$A_K = \frac{1}{2} + \left|\frac{r_K}{4p(1-p)}\right|$$

$$A_W = \frac{1}{2} + |W|$$

$$A_S = \frac{1}{2} + \left|\frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)}\right|$$

$$A = A_K = A_W = A_S$$

Semi-parametric Gaussian Copula (SGC) : Defining latent R-square

We need to relate Quadrant correlation (robust) to AUC and also define an alternative goodness-of-fit measure, **latent R-square (R_l^2) to keep in analogy with the continuous case**.

Definition 3.1. *We say that (Y, X) follows a* **Nonparanormal** *distribution if there exists monotone functions f_Y, f_X such that $(U, V) = (f_Y(Y), f_X(X)) \sim N_2(0, 0, 1, 1, r)$.*

Definition 3.2. *Suppose we have binary variable Y and continuous variable X. Then if there exists latent variable Z, montone functions f_Z, f_X such that, $(Y, X) = (I\{f_Z(Z) > \Delta\}, X)$ and, $(U, V) = (f_Z(Z), f_X(X)) \sim N_2(0, 0, 1, 1, r)$, then we define (Y, X) to follow* Latent non-paranormal distribution.

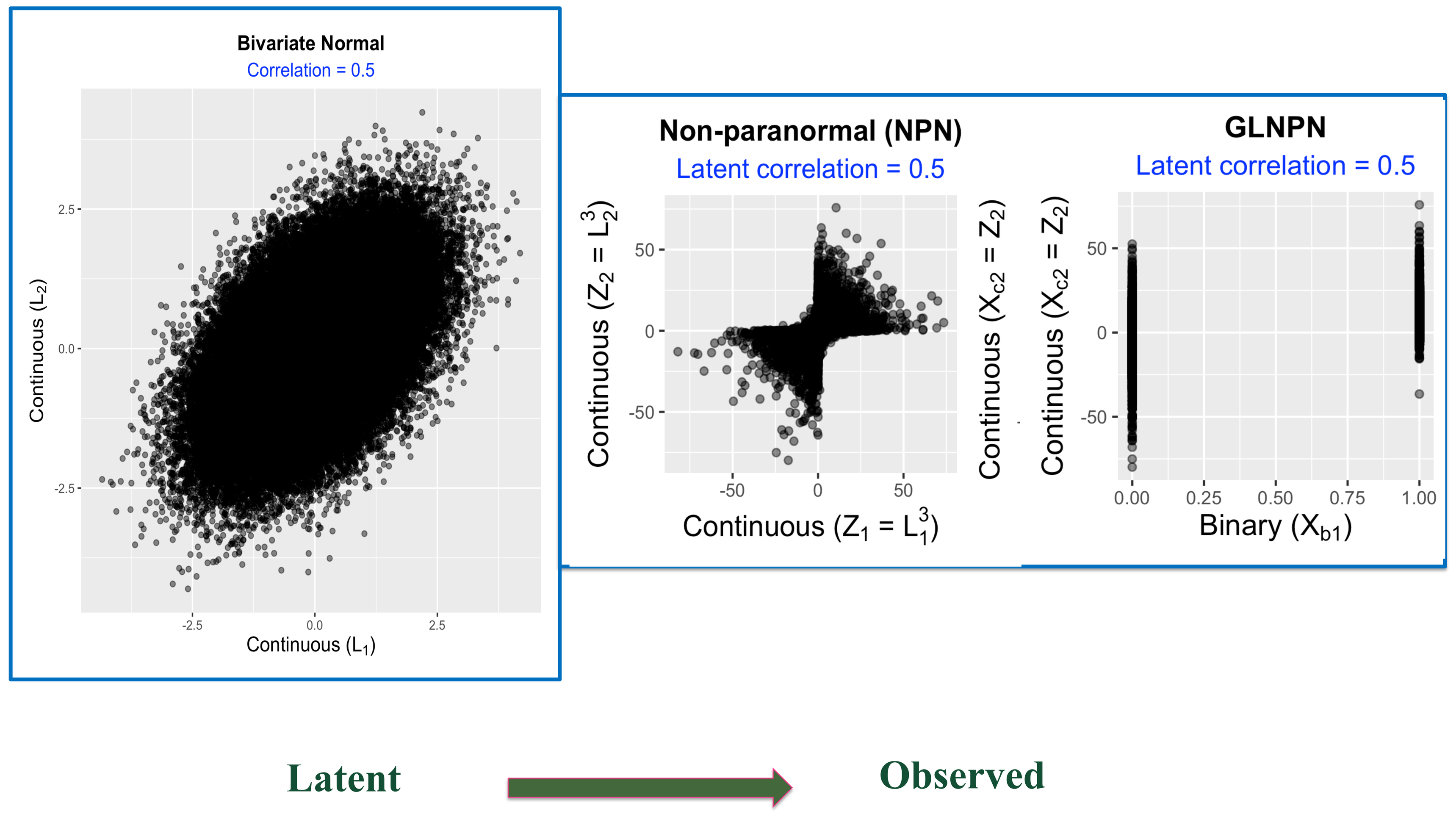


Figure 1: Illustration of Semi-parametric Gaussian Copula

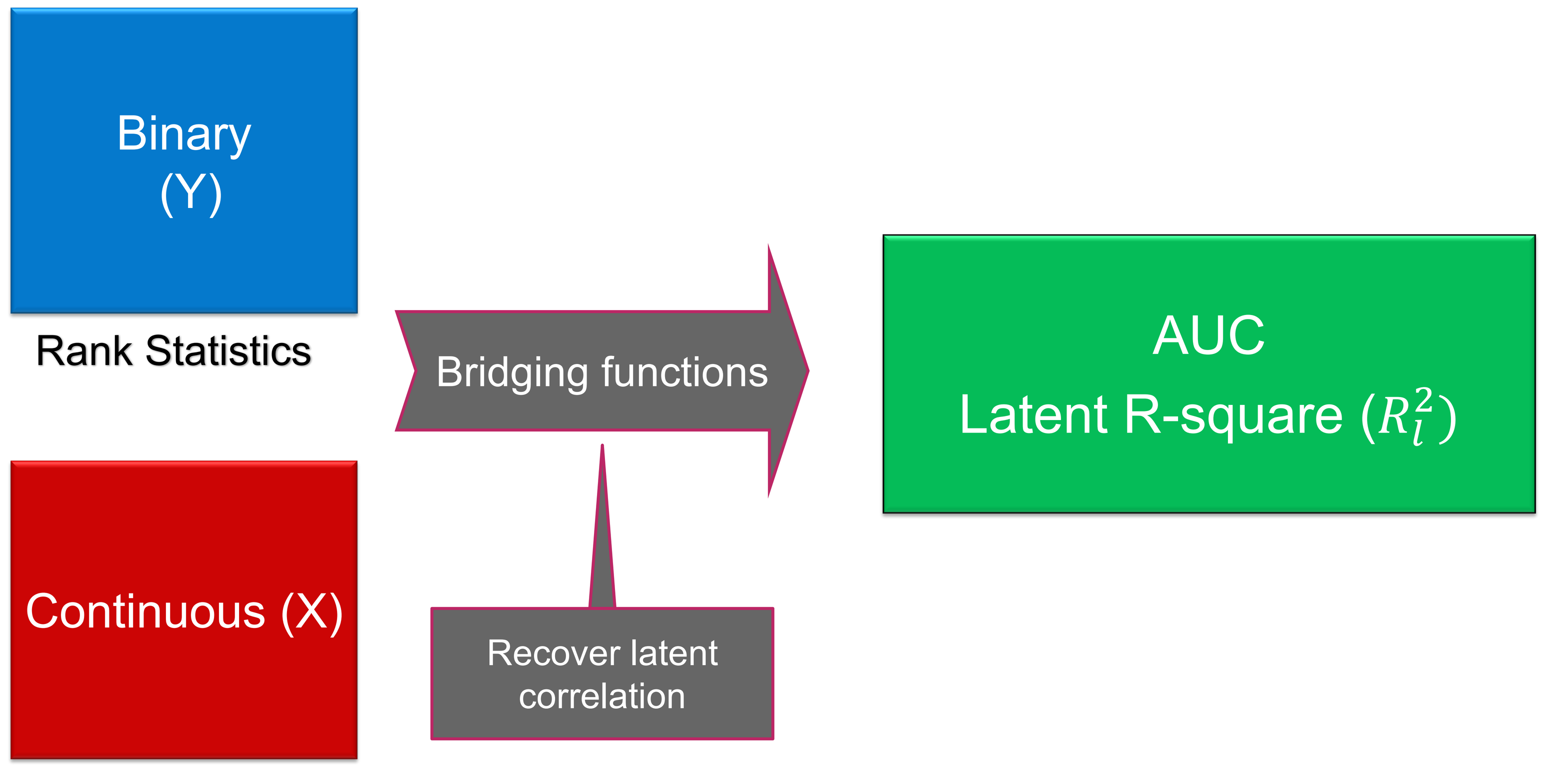


Figure 2: Flowchart of calculating R-square

Relation between AUC and Rank Correlations (under SGC)

$$A_K = \frac{1}{2} + \left|\frac{r_K}{4p(1-p)}\right|$$

$$A_W = \frac{1}{2} + |W|$$

$$A_S = \frac{1}{2} + \left|\frac{G_K(G_S^{-1}(r_S))}{4p(1-p)}\right| = \frac{1}{2} + \left|\frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)}\right|$$

$$A_Q = \frac{1}{2} + \left|\frac{G_K(G_Q^{-1}(r_Q))}{4p(1-p)}\right|$$

$$A = A_K = A_W = A_S = A_Q$$

Latent R^2 more fundamental: Same AUC, but different R^2

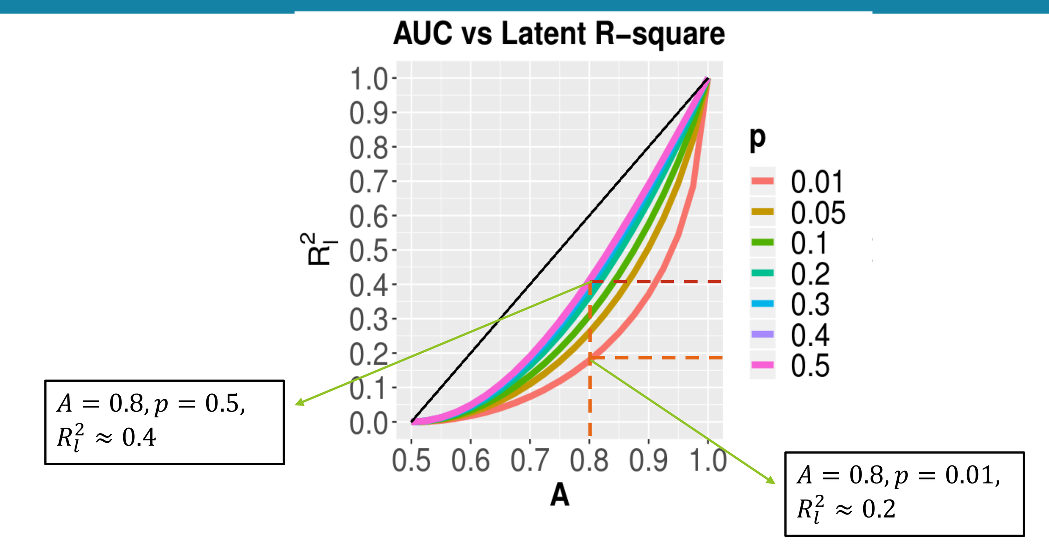
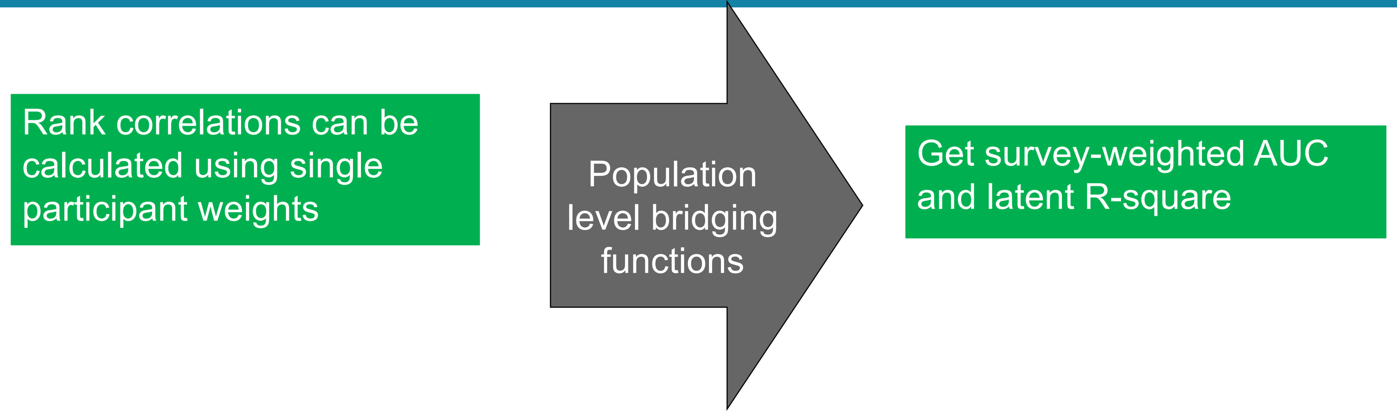


Figure 4: AUC vs Latent R-square (with varying p)

Complex surveys (NHANES): AUC from single participant weights



- Age range (50-84).
- Y: 5 year mortality.
- X: Age/Albumin/Systolic BP/TAC/MVPA/ASTP.
- 3069 subjects with 507 deaths, so p = 0.17
- 100 replicate survey bootstrap confidence intervals are reported in brackets.

	Variables	A_{Kuw}	Rank	A_{Kpw}	Rank	A_W	Rank	A_S	Rank	A_Q	Rank
1	TAC	0.75 (0.75, 0.75)	1	0.8 (0.75, 0.83)	1	0.8 (0.75, 0.83)	1	0.8 (0.75, 0.83)	1	0.77 (0.73, 0.8)	2
2	MVPA	0.73 (0.73, 0.73)	3	0.78 (0.74, 0.81)	2	0.78 (0.73, 0.81)	2	0.78 (0.74, 0.81)	2	0.78 (0.75, 0.82)	1
3	Age	0.74 (0.74, 0.74)	2	0.77 (0.72, 0.8)	3	0.76 (0.72, 0.8)	4	0.77 (0.72, 0.8)	3	0.74 (0.7, 0.77)	4
4	ASTP	0.73 (0.73, 0.73)	4	0.76 (0.73, 0.8)	4	0.76 (0.73, 0.81)	3	0.76 (0.73, 0.8)	4	0.74 (0.7, 0.78)	3
5	Albumin	0.65 (0.65, 0.65)	5	0.7 (0.66, 0.73)	5	0.7 (0.66, 0.73)	5	0.7 (0.66, 0.73)	5	0.68 (0.64, 0.71)	5
6	Systolic BP	0.54 (0.54, 0.54)	6	0.53 (0.5, 0.57)	6	0.53 (0.5, 0.57)	6	0.53 (0.5, 0.57)	6	0.5 (0.5, 0.57)	6

Table 1: *AUC estimates and 95% bootstrap confidence intervals for continuous predictors in NHANES 2003-2006.*

Simulation Studies: Robust AUC (Quadrant)

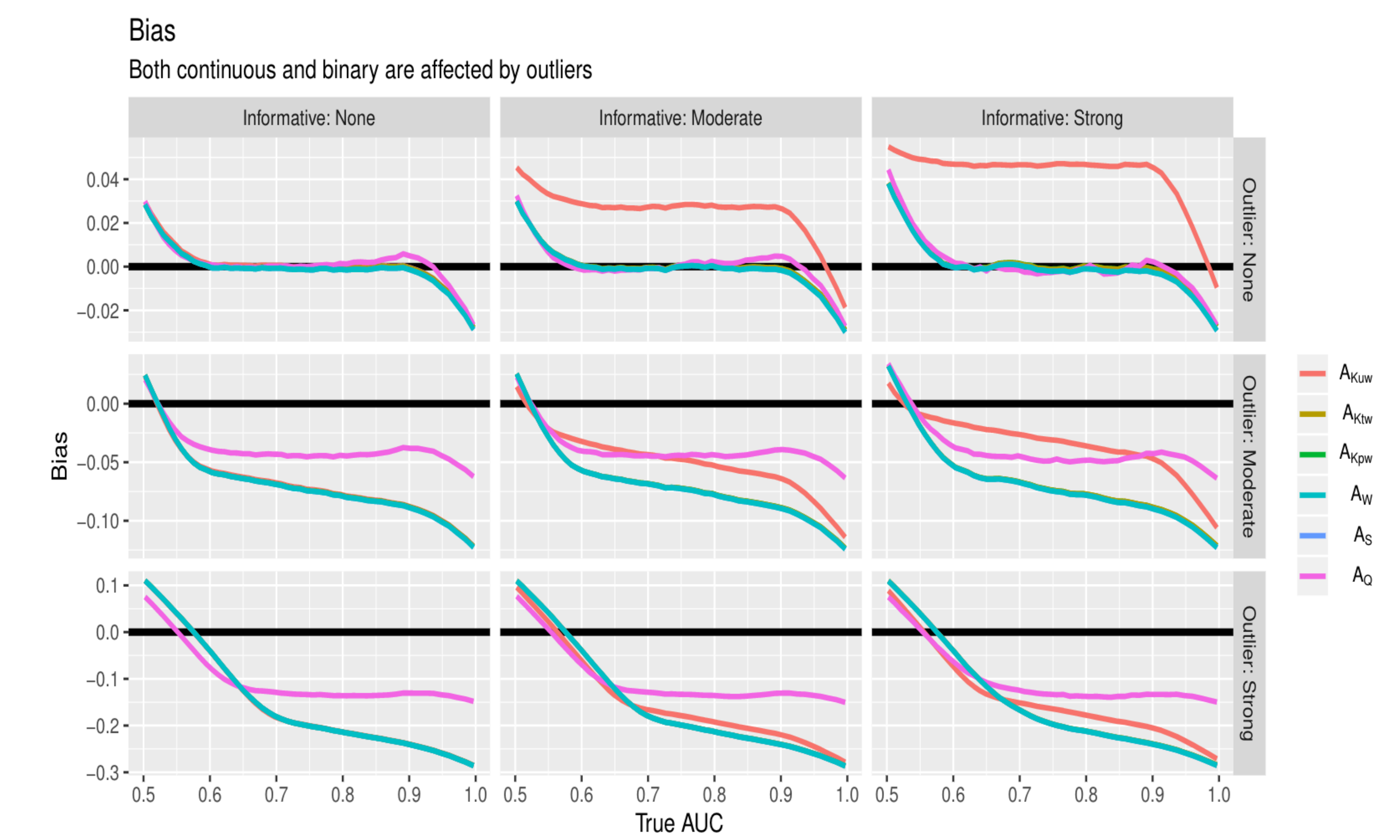


Figure 5: As outlyingness increases, AUC calculated from Quadrant shows less bias

References

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. Lumley, T. and Scott, A. J. (2013). Two-sample rank tests under complex sampling.Biometrika100,831–842.