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

Debangan Dey
ddey1@jhu.edu

@debangan07
Dept. of Biostatistics,
Johns Hopkins Bloomberg School of Public Health

Dr. Vadim Zipunnikov
Dept. of Biostatistics,
Johns Hopkins Bloomberg School of Public Health

Contents

- Motivation
- AUC and Rank Statistics
- Gaussian Copula
- Latent R-square
- Simulation
- Data example

Large scale health surveys (NHANES)

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), serum albumin, systolic blood pressure etc.

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

What does AUC mean? How large is large?

Table 6 Ranking of individual mortality predictors importance based on AUC criteria in from univariate logistic regressions

Rank	Adjusted weights		Unadjusted weights		Unweighted	
	Variable	AUC	Variable	AUC	Variable	AUC
1	TAC	0.783	TAC	0.784	TAC	0.753
2	MVPA	0.756	MVPA	0.757	Age	0.735
3	Age	0.747	Age	0.746	MVPA	0.729
4	$ASTP_{sl/nw}$	0.745	Sedentary, sleep, or non-wear	0.745	$ASTP_{sl/nw}$	0.727
5	Sedentary, sleep, or non-wear	0.744	$ASTP_{sl/nw}$	0.745	Sedentary, sleep, or non-wear	0.724

*The Table is from - *Organizing and Analyzing the Activity Data in NHANES, Andrew Leroux et al, Statistics in Biosciences, Feb 2019.*

Motivation

- Prediction of binary outcomes is an important problem: 5-year mortality in NHANES
- 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

Our contributions

- 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.

Setup

- (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.

Population-level AUC (A)

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

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

Population-level Rank Correlations

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 bivariate distribution.

AUC ⟷ Kendall, Wilcoxon, Spearman

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

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

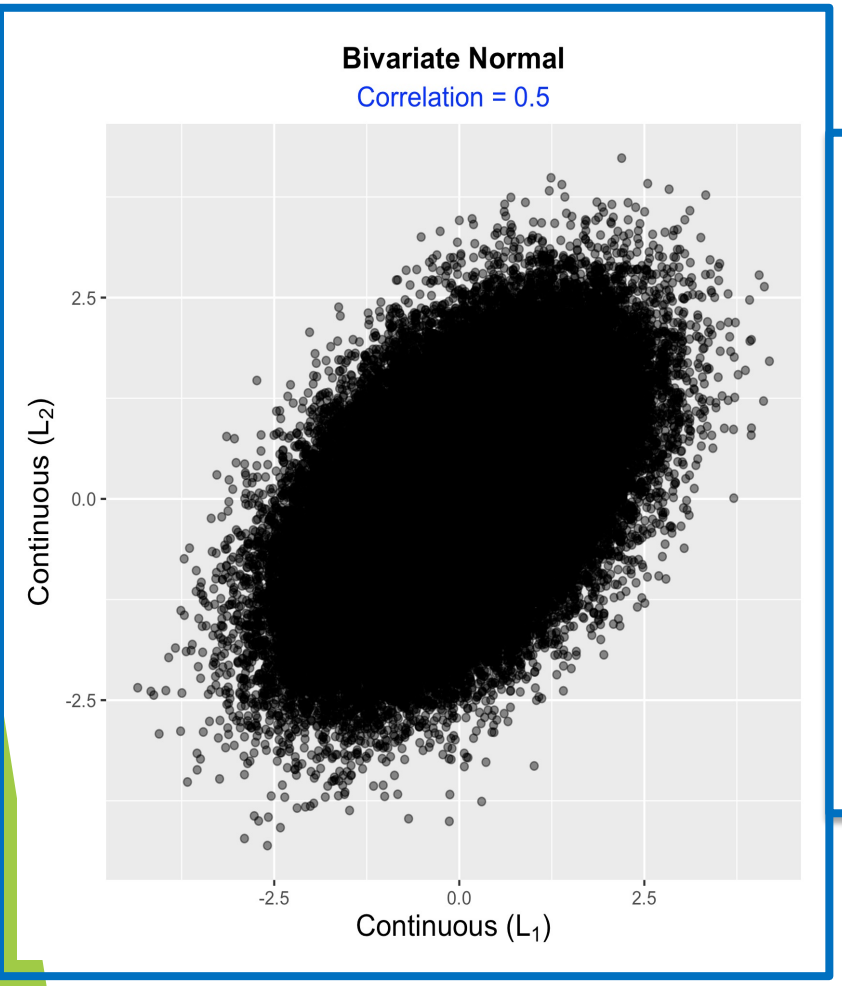
$$A_S = \frac{1}{2} + \left| \frac{r_S - (6p^2 - 6p + 3)}{12p^2(1-p)} \right|$$

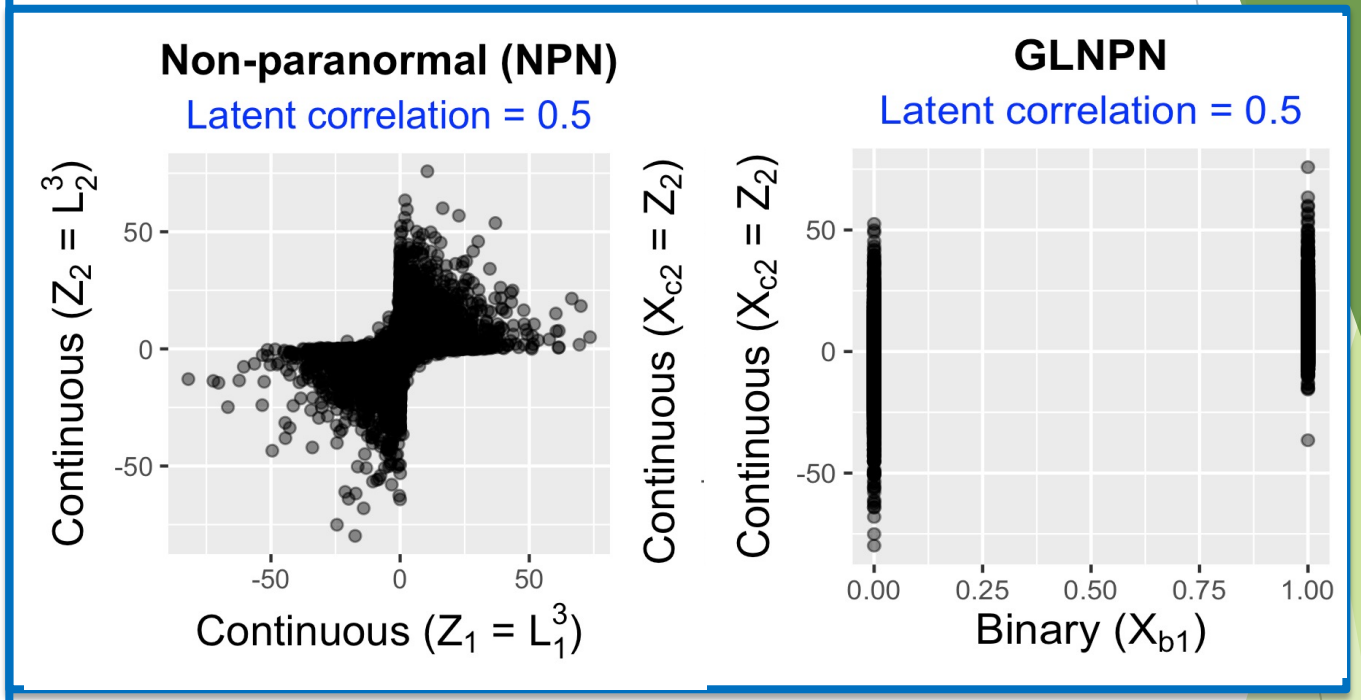
$$A = A_K = A_W = A_S$$

Why we need Gaussian Copula?

- Need to relate Quadrant correlation (robust) to AUC.
- Define an alternative goodness-of-fit measure, latent R-square (R_l^2) to keep in analogy with the continuous case.

Illustration





Latent → **Observed**

Gaussian Semiparametric Copula

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

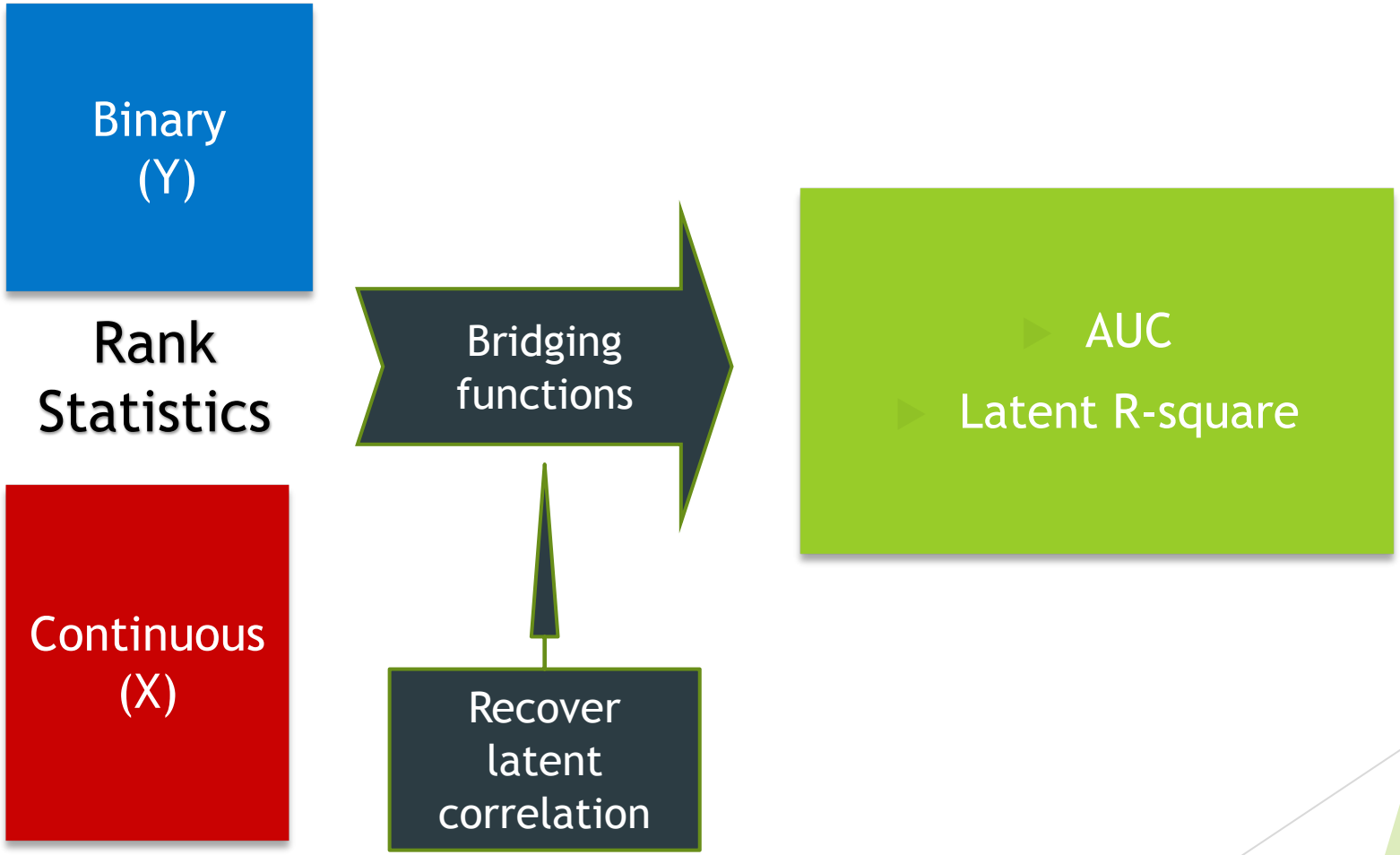
Definition 3.2. *Suppose we have binary variable* Y *and continuous variable* X. *Then if there exists latent variable* Z, *montone functions* f_Z, f_X *such that,* $(Y, X) = (I\{f_Z(Z) > \Delta\}, X)$ *and,* $(U, V) = (f_Z(Z), f_X(X)) \sim N_2(0, 0, 1, 1, r)$, *then we define* (Y, X) *to follow* Latent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

Fan, Jianqing, et al. "High dimensional semiparametric latent graphical model for mixed data." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79.2 (2017): 405-421.

Properties

- Invariant under monotone transformations.
- Generalizes biserial, tetrachoric correlations.
- Latent correlation can be recovered from rank statistics using bridging functions (G_K, G_S, G_Q)

How does it work?



Under Gaussian Copula

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

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

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

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

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

Latent R-square (R_l^2)

- Goodness-of-fit measure in binary-continuous model.
- Proposed in analogous to typical R^2 statistic.

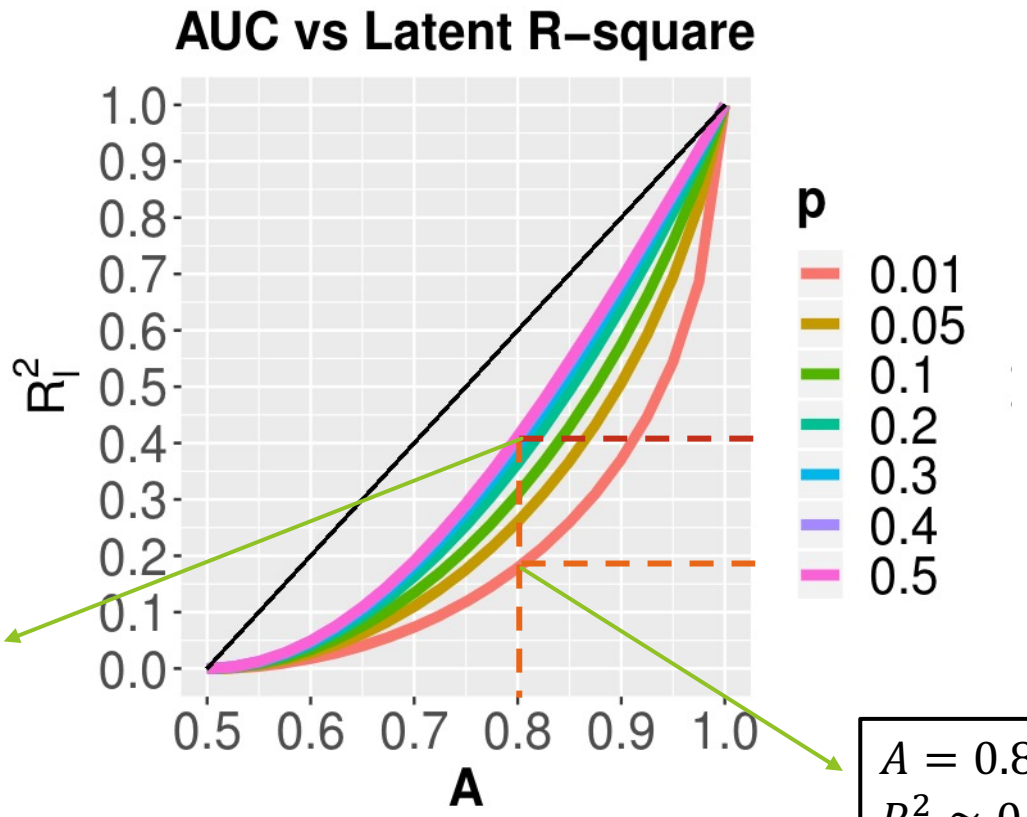
$$\begin{aligned} R_{lK}^2 &= (G_K^{-1}(r_K))^2 \\ R_{lS}^2 &= (G_S^{-1}(r_S))^2 \\ R_{lQ}^2 &= (G_Q^{-1}(r_Q))^2 \\ R_l^2 = R_{lK}^2 &= R_{lS}^2 = R_{lQ}^2 \end{aligned} \tag{15}$$

- Quantifies the amount of variance explained by the predictor variable in the latent space.

Latent R-square and why is it important?

- Non-linear function of AUC.
- Relates traditional variance explained (R-square) intuition into binary-continuous case.
- Takes into account loss of information from dichotomization.
- AUC can be high, with low proportion of cases but latent R-square will be lower.

Illustration (Same AUC, different variance explained)



$A = 0.8, p = 0.5,$
$R_l^2 \approx 0.4$

$A = 0.8, p = 0.01,$
$R_l^2 \approx 0.2$

Complex Surveys

$$\hat{r}_K = \frac{1}{\sum_{i<j} \frac{1}{\hat{w}(i,j)}} \sum_{i<j} \frac{1}{\hat{w}(i,j)} [(Y_i - Y_j) sgn(X_i - X_j)]$$

$$\hat{W} = \frac{1}{\sum_{i:Y_i=1} \frac{1}{w(i)}} \sum_{i:Y_i=1} \frac{1}{w(i)} \hat{F}_X(X_i) - \frac{1}{\sum_{i:Y_i=0} \frac{1}{w(i)}} \sum_{i:Y_i=0} \frac{1}{w(i)} \hat{F}_X(X_i) \tag{19}$$

$$\hat{r}_S = 12 \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} [\hat{F}_Y(Y_i) \hat{F}_X(X_i)] - 3$$

$$\hat{r}_Q = \frac{1}{\sum_{i=1}^{n} \frac{1}{w(i)}} \sum_{i=1}^{n} \frac{1}{w(i)} sgn((Y_i - \hat{M}_Y)(X_i - \hat{M}_X)) \tag{20}$$

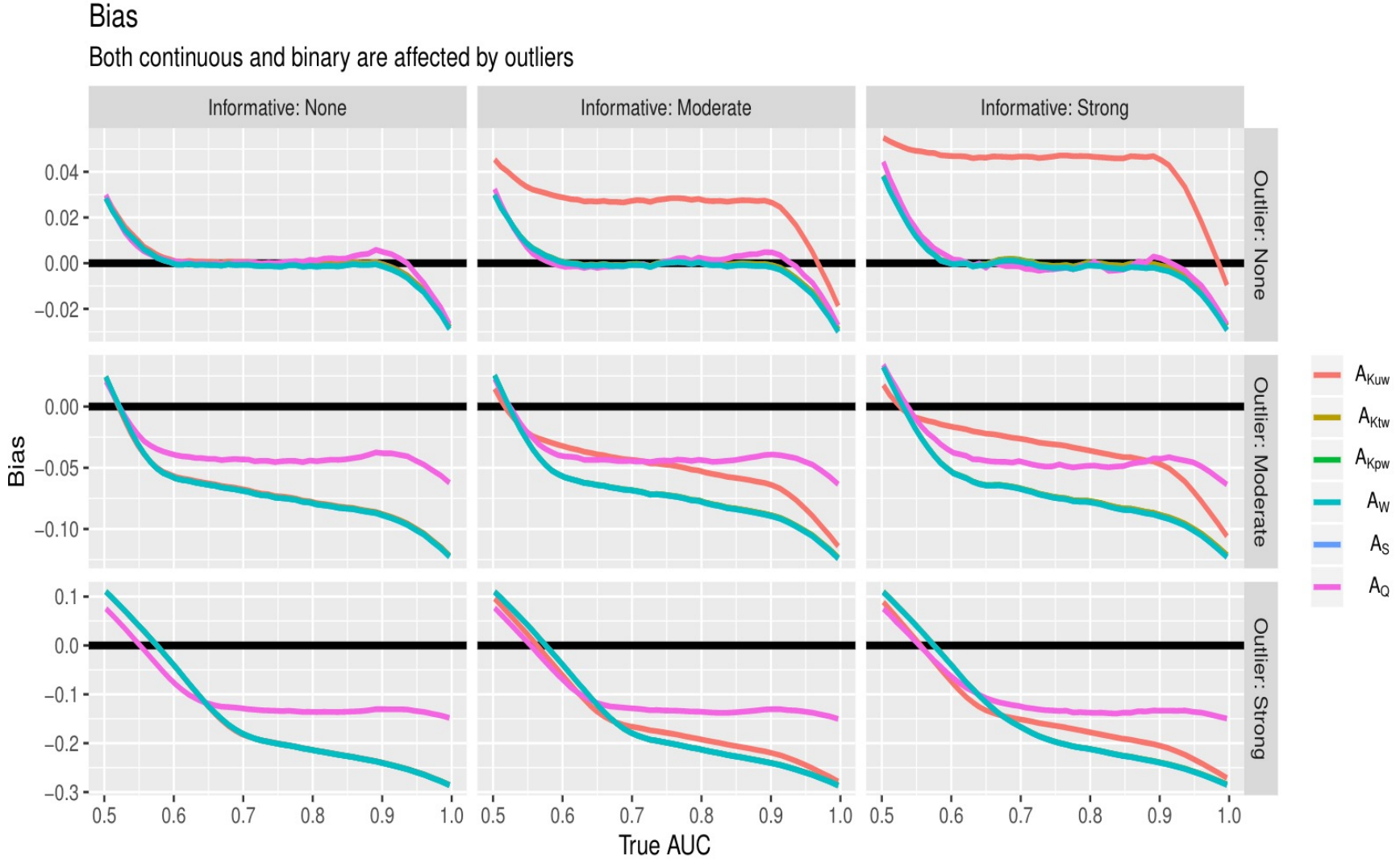
Depending on how we specify $\hat{w}(i,j)$, we define different estimates of r_K in the following way - (1) $\hat{r}_{Kuw}$ ($Unweighted : \hat{w}(i,j) = 1$), (2) $\hat{r}_{Ktw}$ ($True\ weighted : \hat{w}(i,j) = w(i,j)$) and (3) $\hat{r}_{Kpw}$ ($Pairwise\ product\ weighted : \hat{w}(i,j) = w(i)w(j)$).

AUC and R_l^2 estimates

- Use bridging functions
- Can get AUC estimates $\widehat{A}_{Kuw}$, $\widehat{A}_{Ktw}$, $\widehat{A}_{Kpw}$, $\widehat{A}_{W}$, $\widehat{A}_{S}$ and $\widehat{A}_{Q}$
- Can get Latent R-square estimates $\hat{R}^2_{lKuw}$, $\hat{R}^2_{lKtw}$, $\hat{R}^2_{lKpw}$, $\hat{R}^2_{lW}$, $\hat{R}^2_{lS}$ and $\hat{R}^2_{lQ}$. $\hat{R}^2_{lKuw}$.

Simulation setup

- Two stage stratified cluster-sampling (10 strata, 10 PSU)
- Population size = 60000, sample size = 600
- Population latent correlation varied between (-0.995,0.995) with 200 equally spaced points.
- **9 scenarios considered based on**
  - **(a) Strata informativeness (0: none, 1: moderate, 2: strong)**
  - **(b) Outlyingness (0% : none, 5%: moderate, 15%: strong).**
- Each experiment run 100 times to get bias, SE, MSE of estimates


Bias
Both continuous and binary are affected by outliers
Informative: None
Informative: Moderate
Informative: Strong
Outlier: None
Outlier: Moderate
Outlier: Strong
Bias
True AUC
0.04
0.02
0.00
−0.02
0.00
−0.05
−0.10
0.1
0.0
−0.1
−0.2
−0.3
0.5
0.6
0.7
0.8
0.9
1.0
A_{Kuw}
A_{Ktw}
A_{Kpw}
A_W
A_S
A_Q

Data example (NHANES 2003-2006)

- 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.*

	Variables	R^2_{lKuw}	Rank	R^2_{lKpw}	Rank	R^2_{lS}	Rank	R^2_{lQ}	Rank
1	TAC	0.33 (0.28, 0.39)	1	0.29 (0.2, 0.37)	1	0.29 (0.2, 0.37)	1	0.23 (0.17, 0.29)	2
2	MVPA	0.27 (0.23, 0.35)	3	0.25 (0.18, 0.32)	2	0.25 (0.18, 0.32)	2	0.26 (0.19, 0.35)	1
3	Age	0.29 (0.24, 0.36)	2	0.23 (0.15, 0.3)	3	0.23 (0.15, 0.3)	3	0.19 (0.12, 0.24)	4
4	ASTP	0.26 (0.22, 0.3)	4	0.22 (0.16, 0.31)	4	0.22 (0.16, 0.31)	4	0.19 (0.13, 0.25)	3
5	Albumin	0.11 (0.1, 0.14)	5	0.13 (0.08, 0.17)	5	0.13 (0.08, 0.17)	5	0.1 (0.06, 0.15)	5
6	Systolic BP	0.01 (0.01, 0.01)	6	0 (0, 0.02)	6	0 (0, 0.02)	6	0 (0, 0.01)	6

Table 2: *R^2_l estimates and 95% confidence intervals for continuous predictors in NHANES 2003-2006.*

Questions

Thank you!