A simple fix for measurement error bias due to a latent covariate in propensity score weighting analysis: Factor scores from models that capture the latent covariate's joint distribution with exposure and other covariates

Trang Quynh Nguyen (tnguye28@jhu.edu)
Department of Mental Health, Johns Hopkins Bloomberg School of Public Health
with Cyrus Ebnesajjad, Hwanhee Hong, Elizabeth Stuart

Joint Statistical Meetings
Chicago, IL, August 1, 2016

Propensity score analysis & measurement error

Propensity score methods (e.g., matching, weighting) help balance covariates between exposed and unexposed groups in an observational study, thus remove confounding in estimating the causal effect of exposure on outcome.

Only covariates used in estimating the propensity score are balanced, hence the assumption of no unmeasured confounding:

- No unobserved confounder
- No measurement error: If a confounder is measured with error, there is still some imbalance in the true covariate, and that residual imbalance confounds the estimated treatment effect.

Latent variable and measurement error

Some covariates are latent variables, e.g., depression, anxiety, self-esteem, readiness-to-learn, substance dependence, stigma, etc.

Many are measured using multi-item scales.

> Data example: Does out-of-school suspension in middle/high school lead to more trouble with the law (arrests)? Baseline covariates include the latent constructs delinquency and academic achievement.

A common practice in propensity score analysis is to use a summary score (sum/mean) of the items. This means there is measurement error bias.

With doubly robust propensity score analysis, measurement error bias affects both the propensity score model and the outcome model.

Our investigation

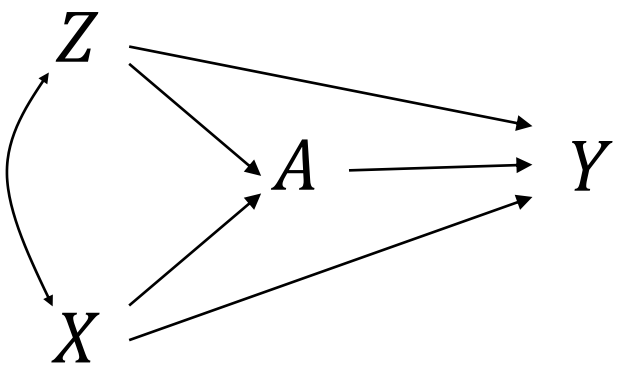
Strategy: use factor scores to represent the latent covariate

- investigate 3 types of factor scores generated from
  - the measurement model (*simple*) (Raykov 2012; Jakubowski 2015)
  - SEM linking the latent factor to the exposure variable (*partially inclusive*)
  - SEM that includes the full exposure assignment model and correlations among covariates (*fully inclusive*)
- comparison methods
  - using the multiple measurements directly
  - using a summary score (mean/sum) of the measurements

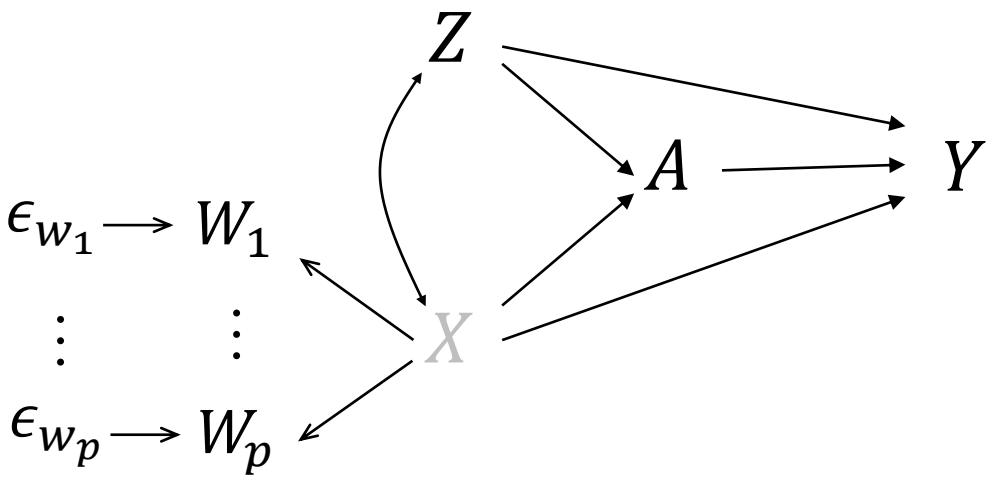
Analysis context: doubly robust propensity score weighting analysis

- propensity score weighting to balance covariates
- regression of outcome on exposure and covariates using weighted sample

Data generating mechanism



Data generating mechanism



Data generating mechanism

Structural model

- X and Z multivariate normal
- Binary A dependent on X and Z via a probit or logit model
- Y normal, dependent on X and Z but not A

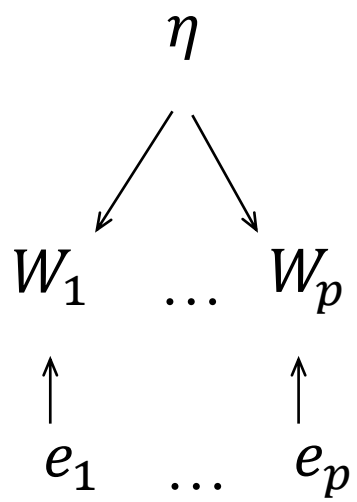
Measurement model

- Ws continuous/ordinal
- Measurement errors normal, non-differential w.r.t. X, Z, A, Y, and independent across the Ws
- Ws have uniform or varying correlations with X (range .4 to .8)

2352 scenarios, each with 1000 simulated n=1000 datasets

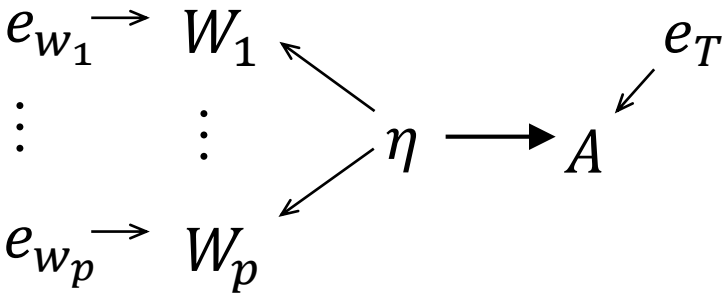
each with number of Ws running from 2 to 10

Factor score types

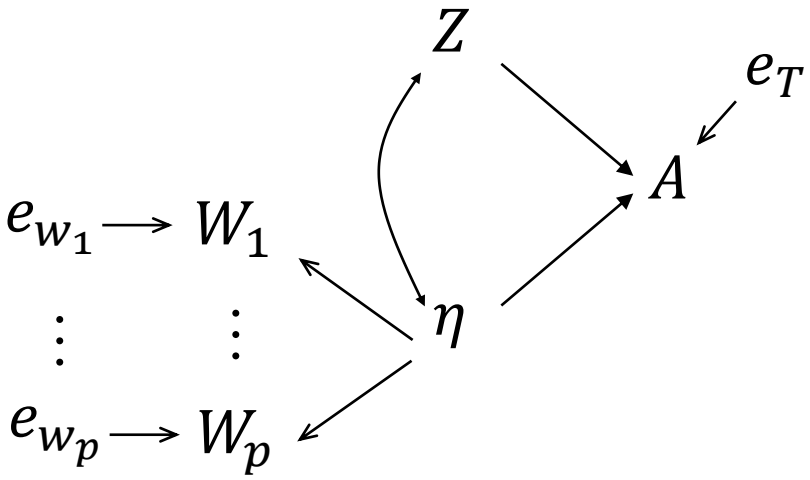


simple

factor score = predicted value of latent η given the observed variables and estimated model parameters



partially inclusive



fully inclusive

Factor score versions

Simple factor score method:

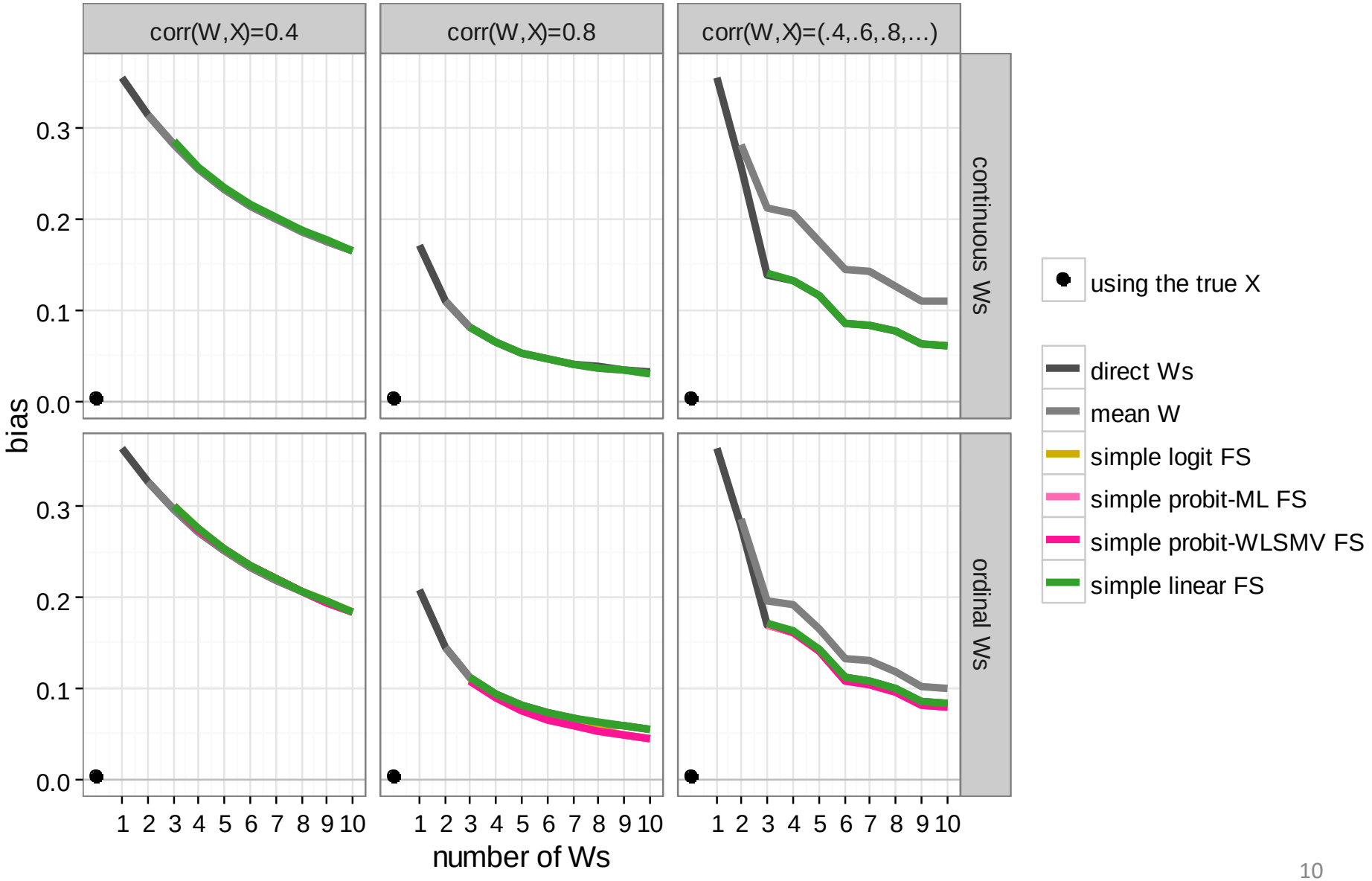
- with continuous Ws: linear measurement model
- with ordinal Ws: logit/probit/linear measurement models

Partially and *fully inclusive* factor scores with continuous and ordinal Ws:

- logit/probit/linear SEMs

Implemented in Mplus using ML and WLSMV estimators

Simple factor score, direct Ws, mean W



alpha=.27-.67; .79-.95; .40-.82

Why *simple* FS not better than direct Ws?

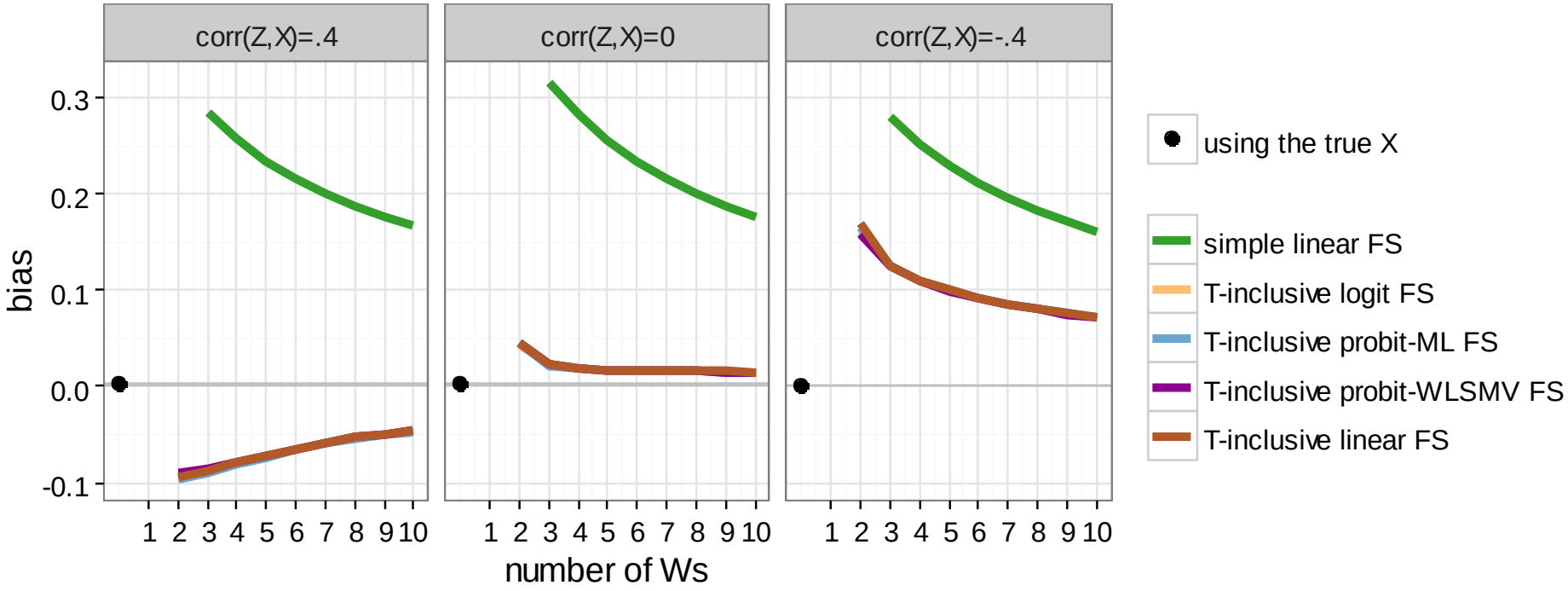
With continuous Ws, the two methods capture the same information about X based on the correlations among the Ws,

correlation of *simple* FS with X = multi-correlation of Ws with X.

The simple factor score does not use any additional information, such as information about the X-A association, or information about the X-Z-A joint distribution,

correlations of *simple* FS with A and Z **<** correlations of X with A and Z.

Partially inclusive vs. *simple* factor scores



Better, but still biased, especially when X and Z are correlated.

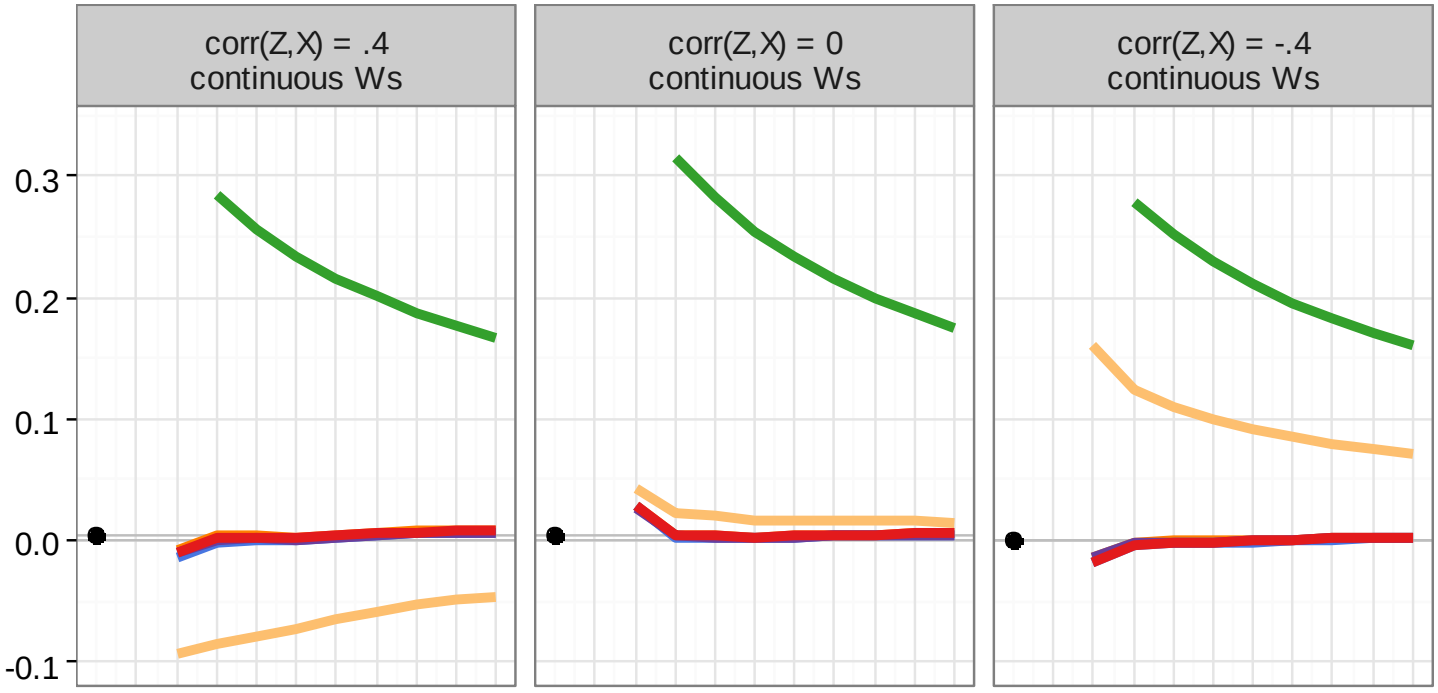
The sign of bias is a function of the sign of the X-Z correlation.

Why doesn't the *partially inclusive* FS method work?

Model imcompatibility:

- the factor score model (imputation model) does not include Z
- the propensity score model (analysis model) includes Z

Fully inclusive vs. *partially inclusive* & *simple* FSs



Fully inclusive factor scores effectively remove bias!

Why does the *fully inclusive* FS method work?

Moving from the *simple* FS to the *fully inclusive* FS

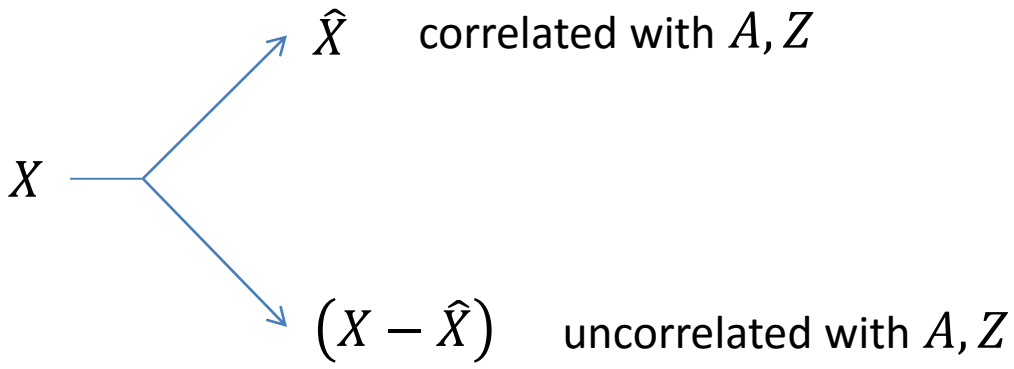
takes us only a tiny bit closer to the true X (the correlation of the FS with X increases very slightly),

but brings us substantially closer to A and Z (the *fully inclusive* FS is more correlated with A and Z than X is).

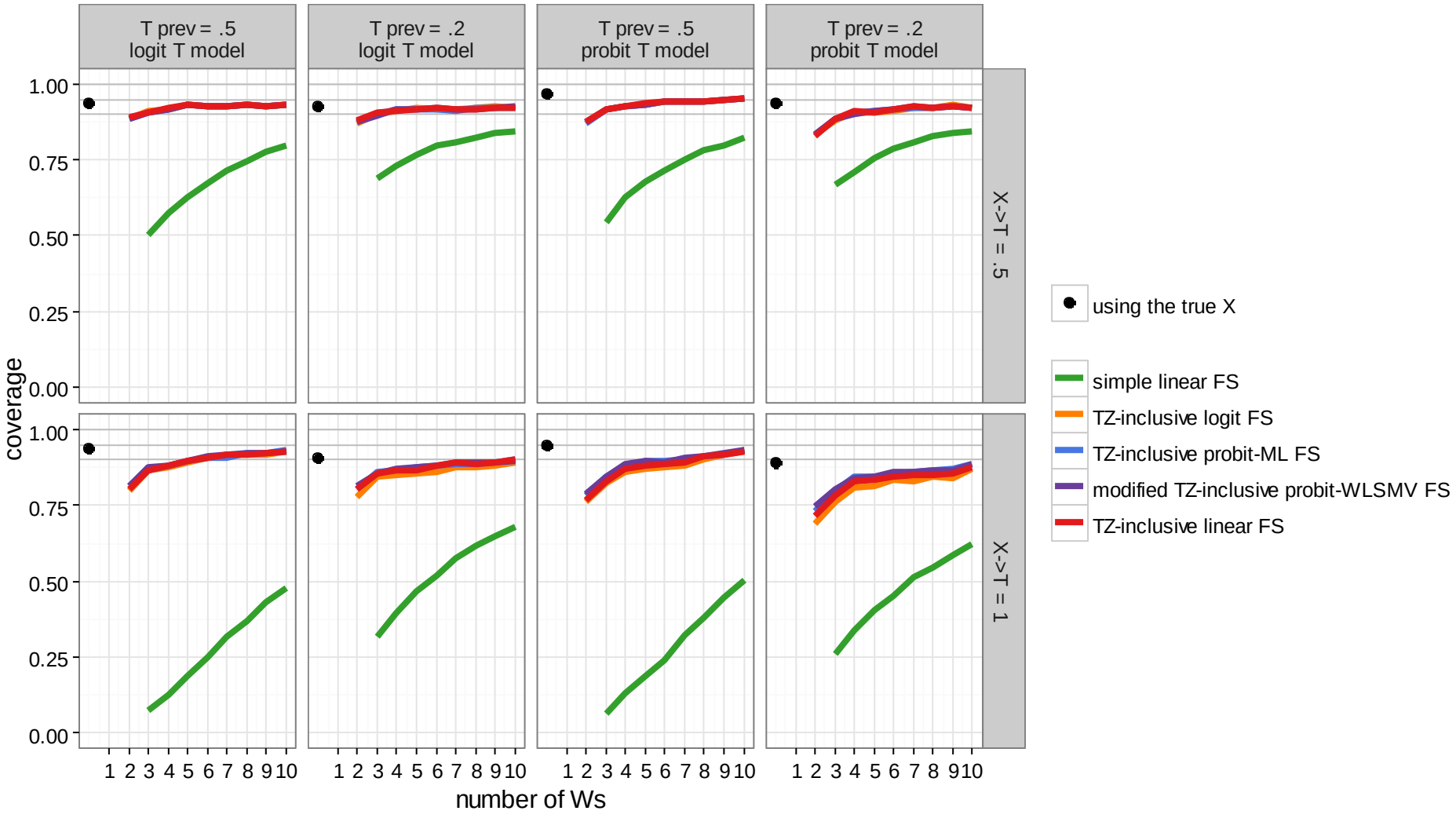
The *fully inclusive* FS approximates the predicted value of X from a regression model using the true X

$$\hat{\eta}\left|\boldsymbol{W}, A, Z \longrightarrow \hat{X}\right|\boldsymbol{W}, A, Z$$

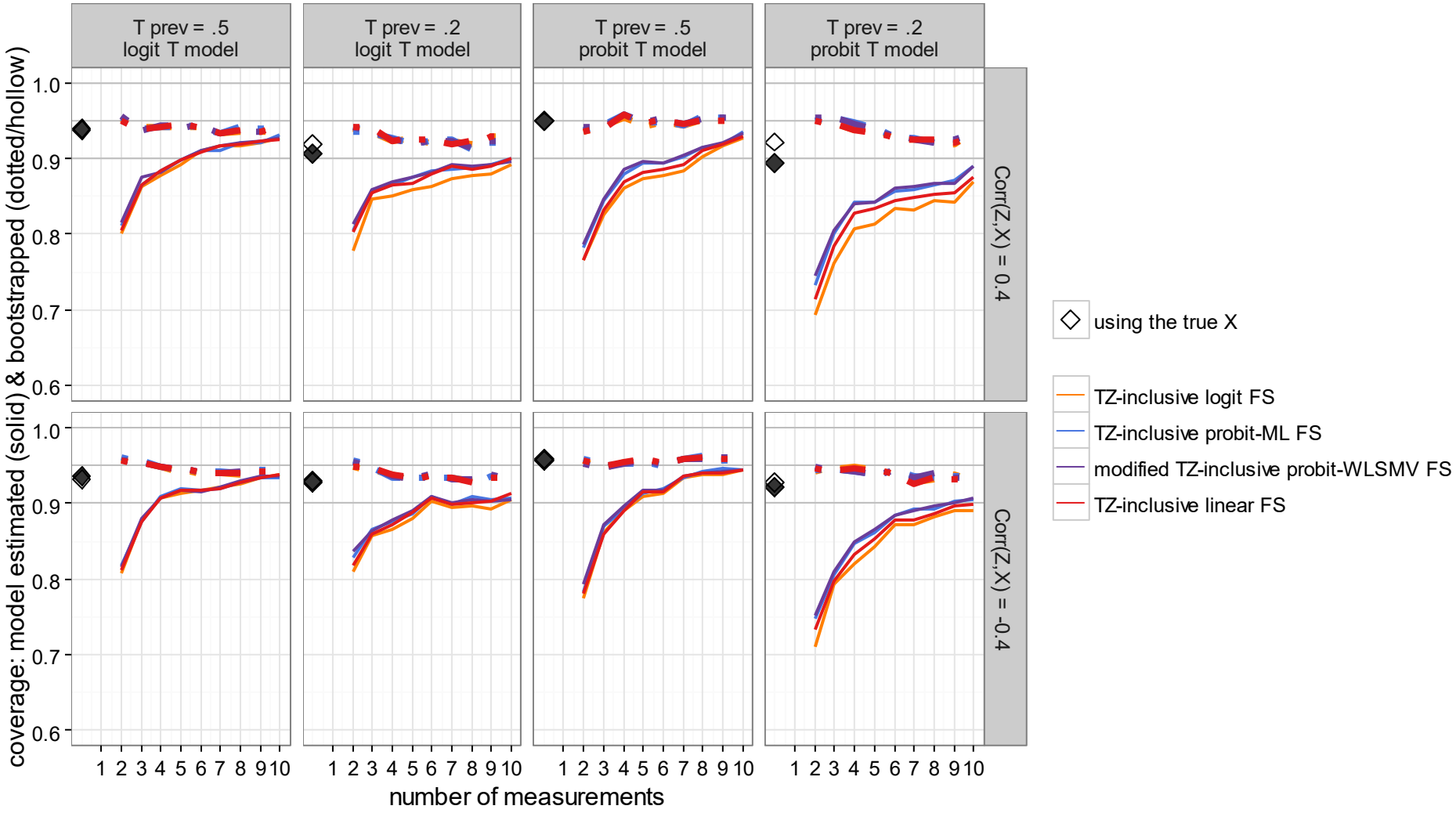
and this predicted value of X is sufficient to remove confounding.



Coverage by outcome-model-estimated CI: *fully inclusive* vs. *simple* factor scores



Coverage by bootstrap interval (dotted/hollow) and by outcome-model-estimated CIs (solid)



Conclusions

A simple fix for bias due to a latent/mismeasured covariate in propensity score weighting analysis is the *fully inclusive* factor score method.

Low coverage may be an issue (if large effect of latent variable on exposure assignment, exposure prevalence far from .5, or few measurements)

Bootstrapping improves coverage.

Note: The regression coefficient of the factor score in the outcome model is NOT an estimate of the effect of the latent variable on the outcome.

Illustrative example

Question of interest: Does out-of-school suspension increase the risk of problems with the law in young adulthood?

Data: The National Longitudinal Study of Adolescent to Adult Health (Add Health)

- representative sample of US adolescents
- 1994-95 (w1, grades 7-12), 1996 (w2), 2001-02 (w3), 2008-09 (w4, ages 24-32)
- public use data, males only

2 sub-samples: using complete data

- A: those who had been suspended prior to wave 1: n=468
- B: those who had never been suspended by wave 1: n=961

Exposure: suspended between waves 1 and 2: 32.9% in A (-> ATE); 6.0% in B (-> ATT)

Outcome: arrested by wave 4: 60.7% in A; 30.9% in B (49.3% in exposed)

Covariates measured at wave 1:

- Manifest: age, race/ethnicity, parent education, parent marital status
- Latent: delinquency, academic achievement

Factor analysis:

- 2 delinquency factors: general (GD, 8 items) and violent (VD, 4 items)
- 1 academic achievement factor (AA, 4 items)

Sample A (suspended before): ATE (for n=468)

Varying adjustment (ignoring measurement error)

		OR (95% CI)	RR	RD	p1	p0
GD, VD, AA	unadjusted	2.01 (1.25,3.23)	1.26	.148	.725	.577
AA	adjusted	1.68 (1.05,2.70)	1.18	.109	.704	.595
GD & VD	adjusted	1.94 (1.21,3.11)	1.24	.138	.724	.586
GD, VD & AA	adjusted	1.63 (1.02,2.60)	1.17	.100	.699	.600

Varying measurement error correction (all three adjusted)

		OR (95% CI)	RR	RD	p1	p0
none	corrected	1.63 (1.02,2.60)	1.17	.100	.699	.600
AA	corrected	1.42 (0.90,2.26)	1.12	.071	.682	.611
GD & VD	corrected	1.53 (0.97,2.43)	1.14	.086	.693	.607
GD, VD & AA	corrected	1.36 (0.87,2.14)	1.10	.062	.678	.616

Sample B (never suspended before): ATT (for n_1=58)

Varying adjustment (ignoring measurement error)

		OR (95% CI)	RR	RD	p1	p0
D1, D2, AA	unadjusted	2.55 (1.40,4.65)	1.68	.200	.493	.293
AA	adjusted	2.34 (1.23,4.46)	1.56	.176	.493	.317
D1 & D2	adjusted	2.02 (1.03,3.95)	1.42	.147	.493	.346
D1, D2 & AA	adjusted	1.86 (0.92,3.79)	1.34	.125	.493	.368

Varying measurement error correction (all three adjusted)

		OR (95% CI)	RR	RD	p1	p0
none	corrected	1.86 (0.92,3.79)	1.34	.125	.493	.368
AA	corrected	1.76 (0.88,3.54)	1.30	.115	.493	.378
D1 & D2	corrected	1.60 (0.77,3.35)	1.24	.095	.493	.399
D1, D2 & AA	corrected	1.55 (0.76,3.18)	1.22	.088	.493	.405

Application and extensions

Application: This kind of correction is useful when

- confounding by the latent covariate(s) is substantial AND measurement error is large
- the estimated treatment effect is borderline significant

Extensions

- Propensity score matching
- More complicated error structures: correlated, differential, non-normal
- More complicated exposure assignment models

Acknowledgements

The Drug Dependence Epidemiology Training Program

NIH funding:

NIDA grant T32DA007292 (Renee M. Johnson)

NIMH grant R01MH099010 (Elizabeth A. Stuart)

Thank you!

Trang Q. Nguyen (tnguye28@jhu.edu)

Supplementary slides

Summary

Methodological context: Propensity score methods are increasingly used to remove confounding to estimate the effect of a treatment (exposure). Key assumption is no unobserved confounding.

Problem: Bias due to a mismeasured/latent covariate

Our work: With multiple measurements of such variable, investigate bias/bias reduction when using a factor score to represent the variable, with several types of factor scores

Main finding: A simple fix is using the factor score from a SEM that includes the measurements and the full treatment assignment model

Illustrative example: Analysis of Add Health data to examine out-of-school suspension as a risk factor for trouble with the law in young adulthood.

Model links and estimators, FS estimation methods

- Linear models: posterior mode (regression-based)
- Logit and probit models fit using ML: posterior mean
- Probit models also fit using WLSMV*: posterior mode

  * *fully inclusive* probit model using WLSMV requires a modification

We used Mplus 7.2.

But most methods can be implemented in other popular statistical packages.

Simulation studies

	Data generating model	Parameter values
Z and X	$\begin{pmatrix} Z \\ X \end{pmatrix} \sim \text{Normal}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$	$\rho = 0, .4, -.4$
T	either logit: $T \sim \text{Binomial}\left(\frac{e^{\beta_0 + \beta_z Z + \beta_x X}}{1 + e^{\beta_0 + \beta_z Z + \beta_x X}} \right)$ or probit: $T^* \sim \text{Normal}\left(\beta_z Z + \beta_x X, 1.7^2\right), T = 1 \cdot (T^* > \tau_T),$ where β_0 and τ_T are set so that $\text{P}(T = 1) = p_T$.	$\beta_z = .5,1$ $\beta_x = .5,1$ $p_T = .5, .4, .3, .2$
Y	$Y \sim \text{Normal}\left((Z + \gamma X + 0T), 4\right).$	$\gamma = 1,2$

Simulation studies

	Data generating model	Parameter values
continuous $\boldsymbol{W}$	$\boldsymbol{W} \sim \text{Normal}(X, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ is diagonal matrix with variance elements based on $\boldsymbol{\phi} = \text{cor}(\boldsymbol{W}, X)$.	$p = 2, \ldots, 10$ 3 cases of uniform W-X correlations: $\boldsymbol{\phi} = .\boldsymbol{4}$ (*low*), $.\boldsymbol{6}$ (*medium*), $.\boldsymbol{8}$ (*high*)
ordinal $\boldsymbol{W}$	$\boldsymbol{W}^* \sim \text{Normal}(X, \boldsymbol{\Sigma})$ as above then categorized into $\boldsymbol{W}$ with four equal mass categories	4 cases of mixed W-X correlations: $\boldsymbol{\phi} = (.4, .6, .4, .6, \ldots)$ (*lome*), $(.4, .8, .4, .8, \ldots)$ (*lohi*), $(.6, .8, .6, .8, \ldots)$ (*mehi*), $(.4, .6, .8, .4, .6, .8, \ldots)$ (*lomehi*)

Bias

The *simple* FS method is similarly biased as the direct Ws method.

The mean W method is more biased than the simple FS and direct Ws methods when W-X correlations are not uniform.

All *simple* FSs perform similarly, except when Ws are ordinal & highly correlated with X – then logit/probit FSs are slightly less biased.

The more measurements and higher W-X correlations, the less bias.

The larger the effects of X on T and Y, the more bias.

Variance

All methods have underestimation of variance when treatment prevalence is far from .5. This problem gets worse for the direct Ws method with increasing number of measurements.

(relevant to high alpha situation when direct Ws looks appealing from a bias perspective)

The *fully inclusive* FS method: other findings

Bias – differentiating the *fully inclusive* factor scores

When treatment prevalence is far from .5, the FS consistent with the true treatment assignment model performs slightly better. This is more pronounced when the effect of X on T increases. However, the differences are small compared to the magnitude of bias of the *simple* FS method.

When Ws are ordinal and some Ws are highly correlated with X, the linear FS is slightly more biased than the non-linear ones.

Variance and coverage

Variance is underestimated, more so when the number of Ws is small, and more so when treatment prevalence is far from .5.

Coverage worsens when the effect of X on T increases.

Coverage is worse when treatment prevalence is far from .5 (also when using X).

In cases where coverage is poor, bootstrapping substantially improves coverage.