Overview of the ACIC Data Challenge, 2019

Susan Gruber, Putnam Data Sciences, LLC
Geneviève Lefebvre, Université du Québec à Montréal
Tibor Schuster, McGill University
Alexandre Piché, Mila, Université de Montréal , Element AI

Atlantic Causal Inference Conference, Montréal, Canada, May 24, 2019

ACIC Data Challenge

- Initiated in 2016 by Jennifer Hill, Vince Dorie, Uri Shalit, Marc Scott, Dan Cervone
  - ATT parameter
  - Data from a single publicly available dataset (55,000 x 6500)
  - 7700 datasets (77 DGPs)
  - Code submitted to organizers
- This year
  - ATE parameter
  - Low and high dimensional tracks
  - Covariates simulated or drawn from 7 source datasets
    - healthcare, business, social
  - 6400 datasets (32 DGPs per track)
  - Teams analyze data and submit results files

Covariate Data Sources
Low-Dimensional Track

- UCI Machine Learning Repository *https://archive.ics.uci.edu/ml/index.php*
  1. cervical cancer
  2. spam email
  3. credit card defaults
  4. student performance
- Vanderbilt University *http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets*
  5. right heart catherization

1. Kelwin Fernandes, Jaime S. Cardoso, and Jessica Fernandes. 'Transfer Learning with Partial Observability Applied to Cervical Cancer Screening.' Iberian Conference on Pattern Recognition and Image Analysis. Springer International Publishing, 2017.
2. Creators: Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt. Hewlett-Packard Labs, 1501 Page Mill Rd., Palo Alto, CA 94304. Donor: George Forman. Generated: June-July 1999
3. NI-Cheng Yeh. icyeh@chu.edu.tw, 140910@mail.tku.edu.tw
Department of Information Management, Chung Hua University, Taiwan, and Department of Civil Engineering, Tamkang University, Taiwan.
4. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.
5. Connors *et al.* (1996): The effectiveness of RHC in the initial care of critically ill patients. *J American Medical Association* 276:889-897.

Covariate Data Sources
High-Dimensional Track

- UCI Machine Learning Repository *https://archive.ics.uci.edu/ml/index.php*
  1. epilepsy
- Columbia University *http://www.stat.columbia.edu/~gelman/arm/examples/*
  2. speed dating
- Simulation
  - Block-dependent covariates simulated from copula models
  - Gaussian, Student, Gumbel, Frank, Joe, Clayton
  - Small to large values for Kendall's tau
  - Allowed for separate and flexible modeling of dependency and marginal distributions

1. Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE (2001. IIndications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E, 64, 061907
2. *Andrew Gelman, http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating*

Data Generating Processes

- Real-world and simulated sources of covariates
- Simulated treatments and outcomes
    - Easy main terms models
    - Poor overlap
    - Complex functional form
    - Treatment effect heterogeneity
    - IVs

Submissions

19 Teams submitted results for 29 different methods

- Industry and academia
- US, Canada, Korea, Germany

Both tracks (20 methods)

1. ac-tmle3, cvtmle3bbd
2. BART, BART_TMLE, BARTcv, BARTpscore, XBARTtmle
3. BCF
4. eb, ensemble, median, psc, rfdripw
5. FisherBART, Fisheric
6. GRF-NET
7. l-learner
8. PCAPS
9. PCATS
10. Std

High-Dim Only (2)

11. DR-CFR
12. SBBSP

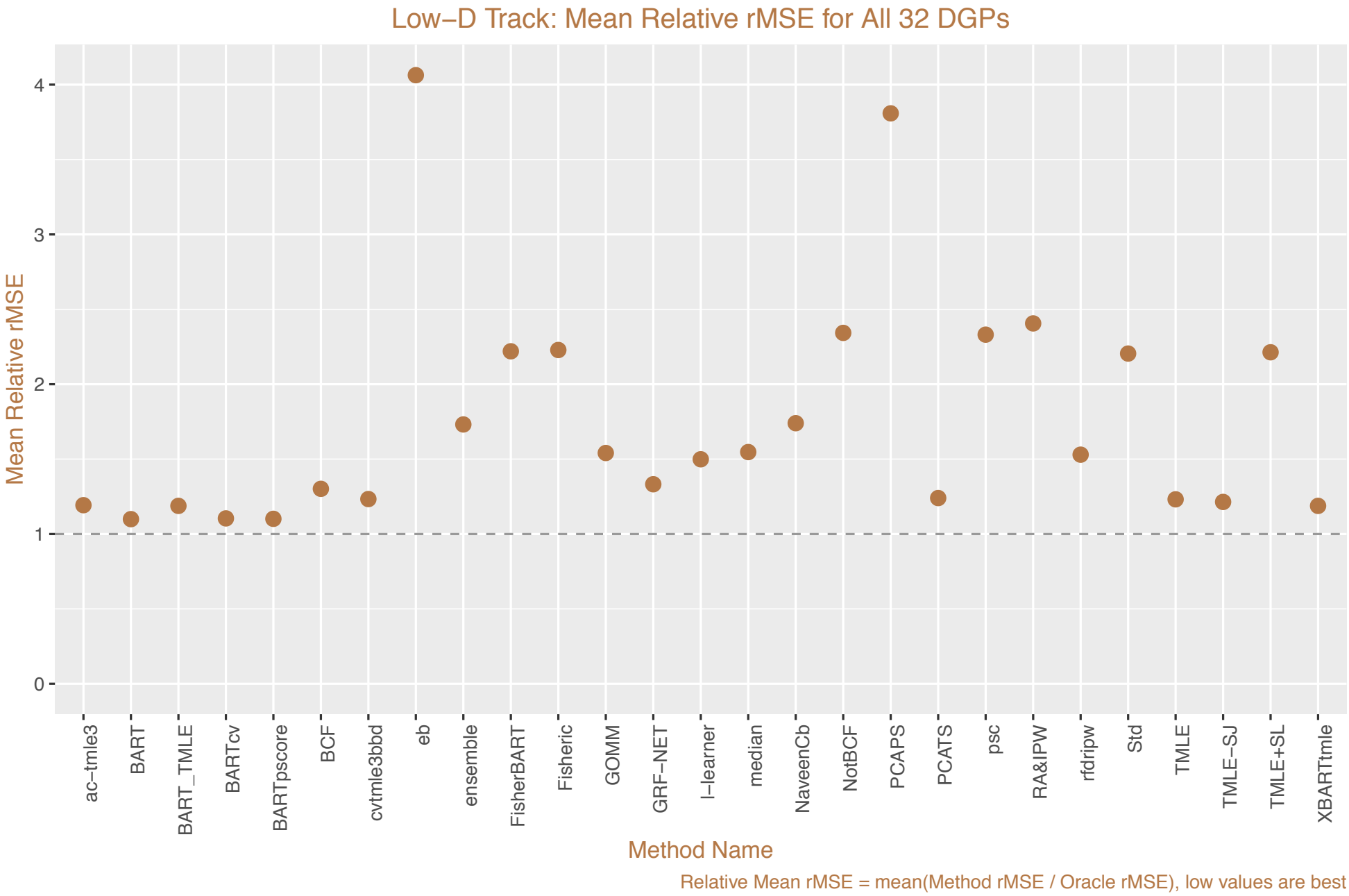
Low-Dim Only (7)

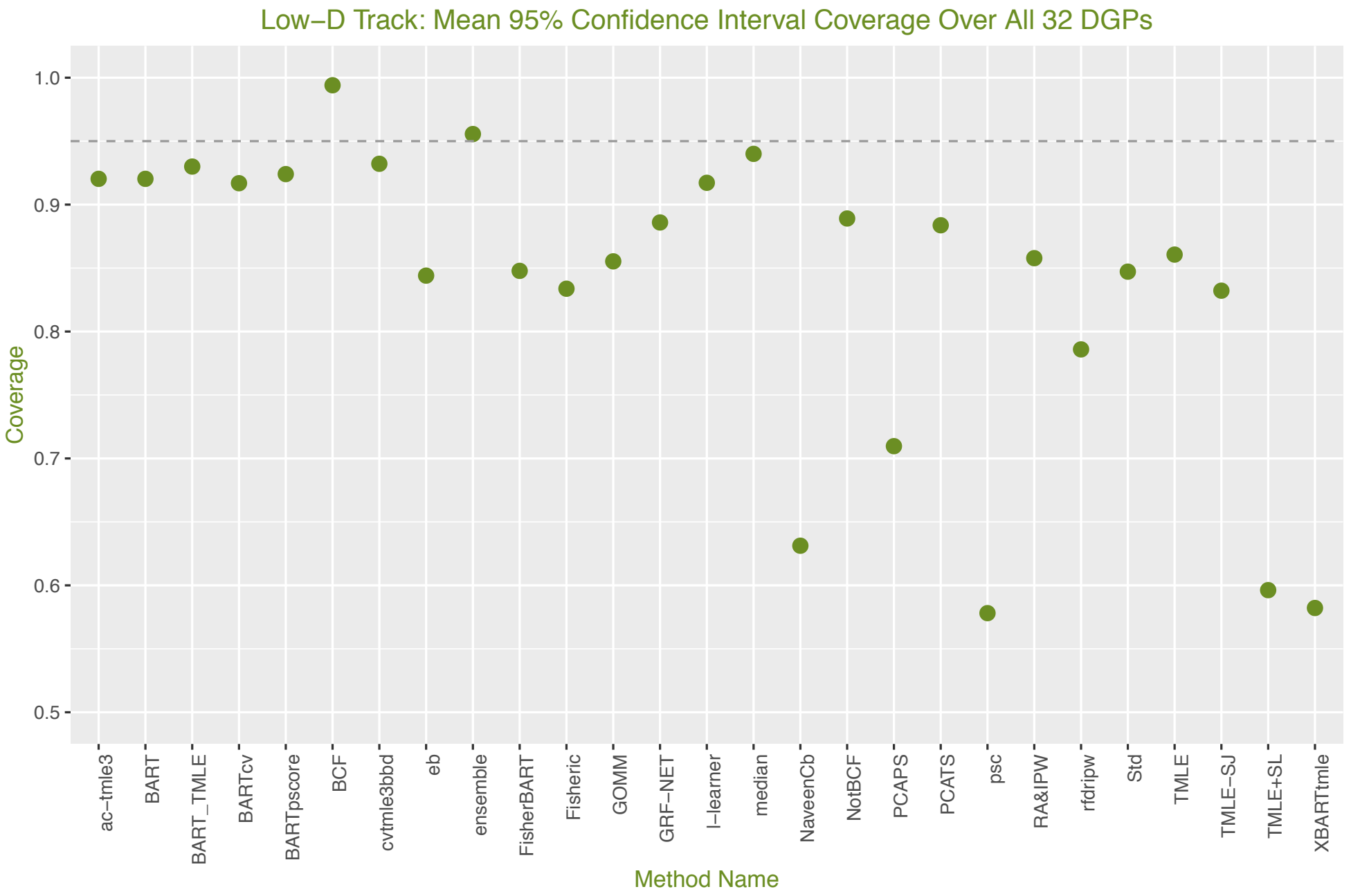
13. GOMM
14. NaveenCb
15. NotBCF
16. RA&IPW
17. TMLE
18. TMLE-SJ
19. TMLE+SL

Overview of Results

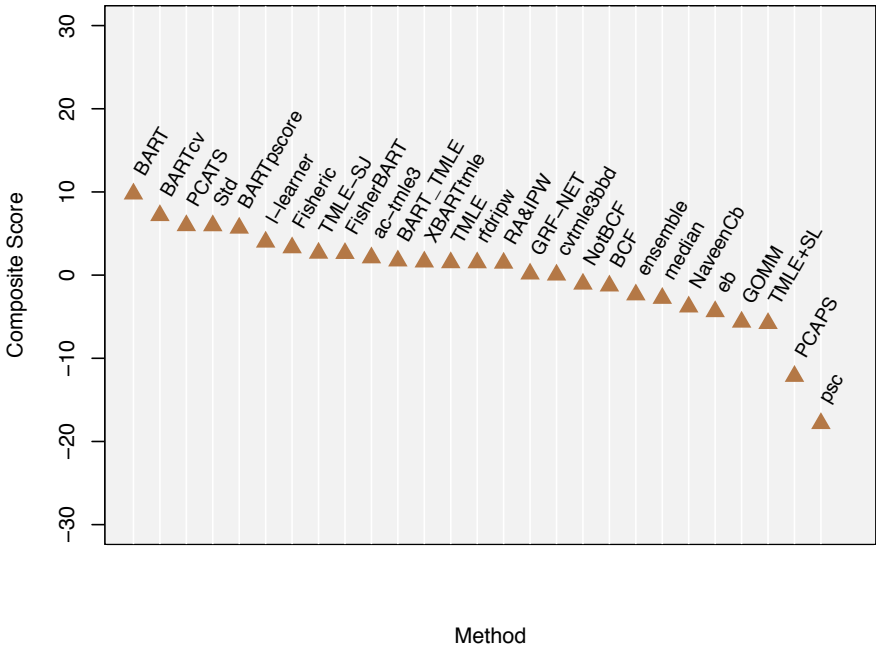
- Today: Summary plots of rMSE, bias, SD, relative to Oracle coverage, and composite rank scores in each track
  - example: relative rMSE = $\frac{1}{32}\sum_{i=1}^{32}\frac{rMSE_{m,i}}{rMSE_{oracle,i}}$
  - Composite rank score
    - Range = -32 to 32
    - 1 point for 1st place, ½ point for 2nd place, ...,
      -1 point for last place, -1/2 for next-to-last place, ...

- Many additional plots and files *available today* on the ACIC Data Challenge Website

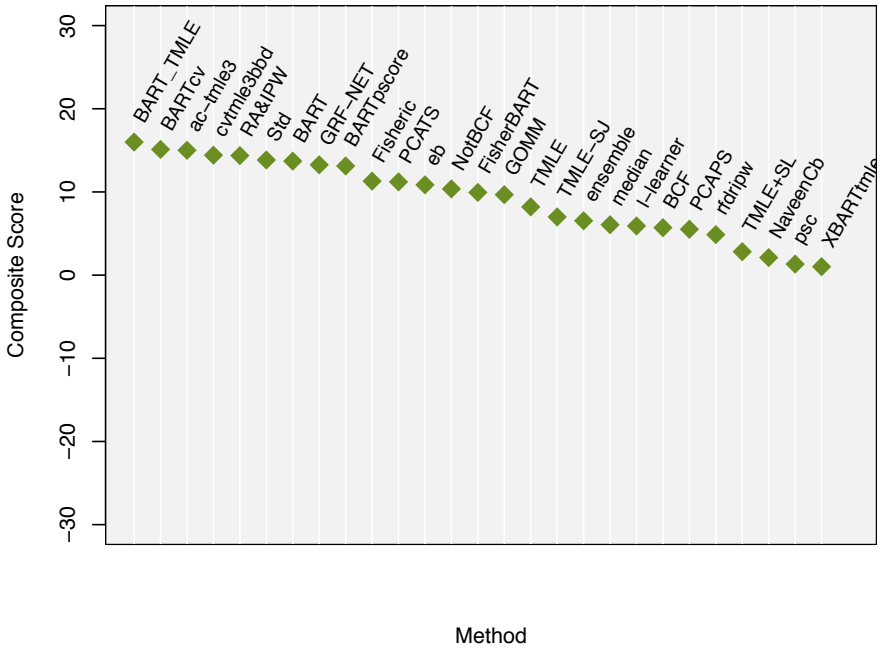

Low–D Track: Mean Relative rMSE for All 32 DGPs
Mean Relative rMSE
0
1
2
3
4
ac–tmle3
BART
BART_TMLE
BARTcv
BARTpscore
BCF
cvtmle3bbd
eb
ensemble
FisherBART
Fisheric
GOMM
GRF–NET
I–learner
median
NaveenCb
NotBCF
PCAPS
PCATS
psc
RA&IPW
rfdripw
Std
TMLE
TMLE–SJ
TMLE+SL
XBARTtmle
Method Name
Relative Mean rMSE = mean(Method rMSE / Oracle rMSE), low values are best


Low−D Track: Mean 95% Confidence Interval Coverage Over All 32 DGPs
Coverage
1.0
0.9
0.8
0.7
0.6
0.5
ac−tmle3
BART
BART_TMLE
BARTcv
BARTpscore
BCF
cvtmle3bbd
eb
ensemble
FisherBART
Fisheric
GOMM
GRF−NET
l−learner
median
NaveenCb
NotBCF
PCAPS
PCATS
psc
RA&IPW
rfdripw
Std
TMLE
TMLE−SJ
TMLE+SL
XBARTtmle
Method Name

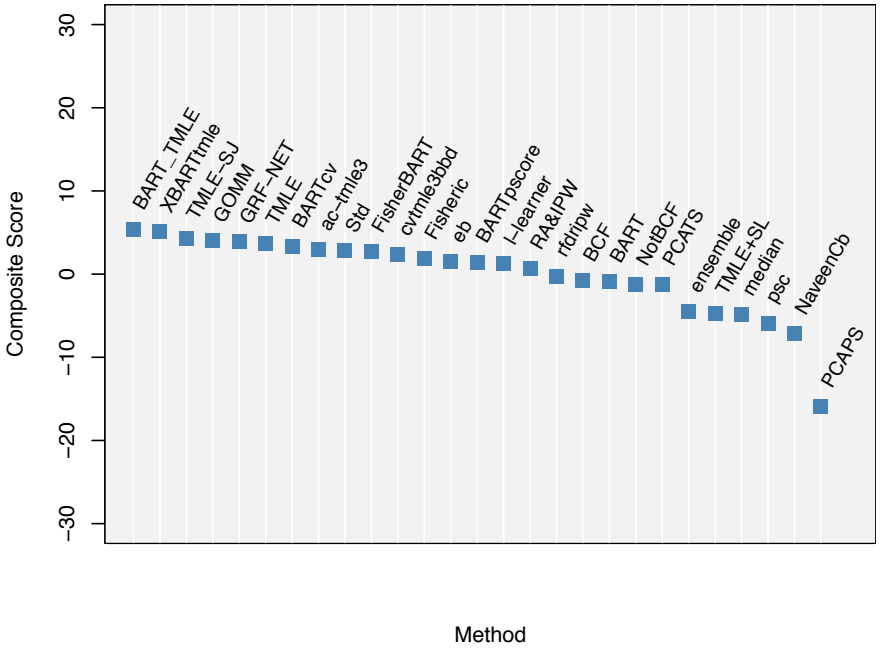
Low–Dim: Composite rMSE Scores



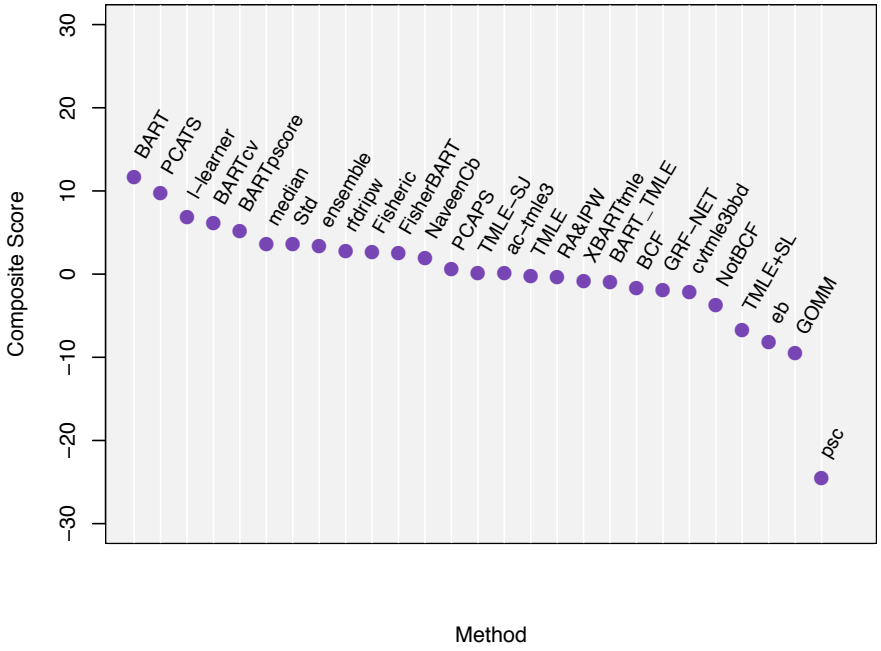
Low–Dim: Composite Coverage Scores

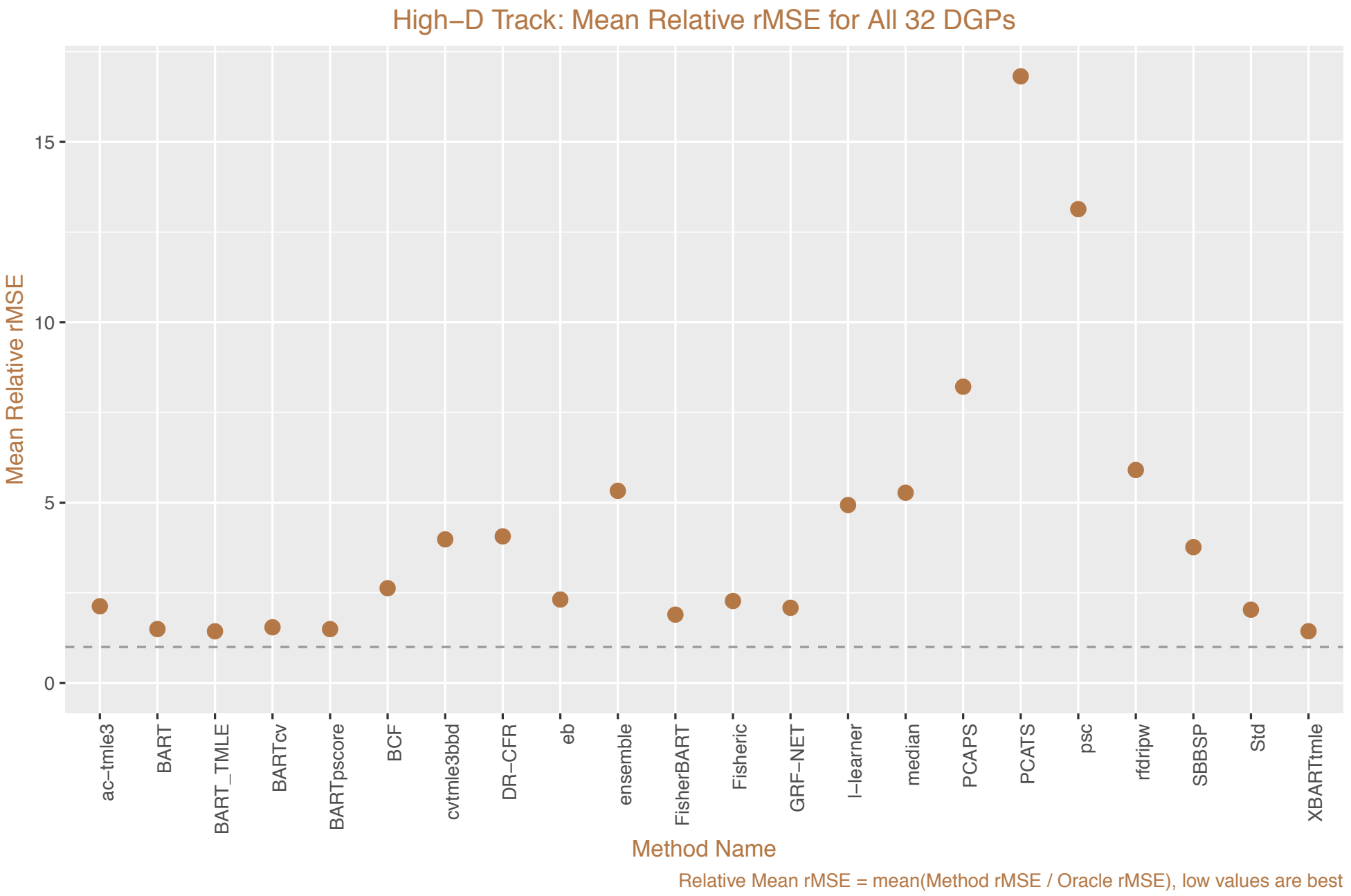


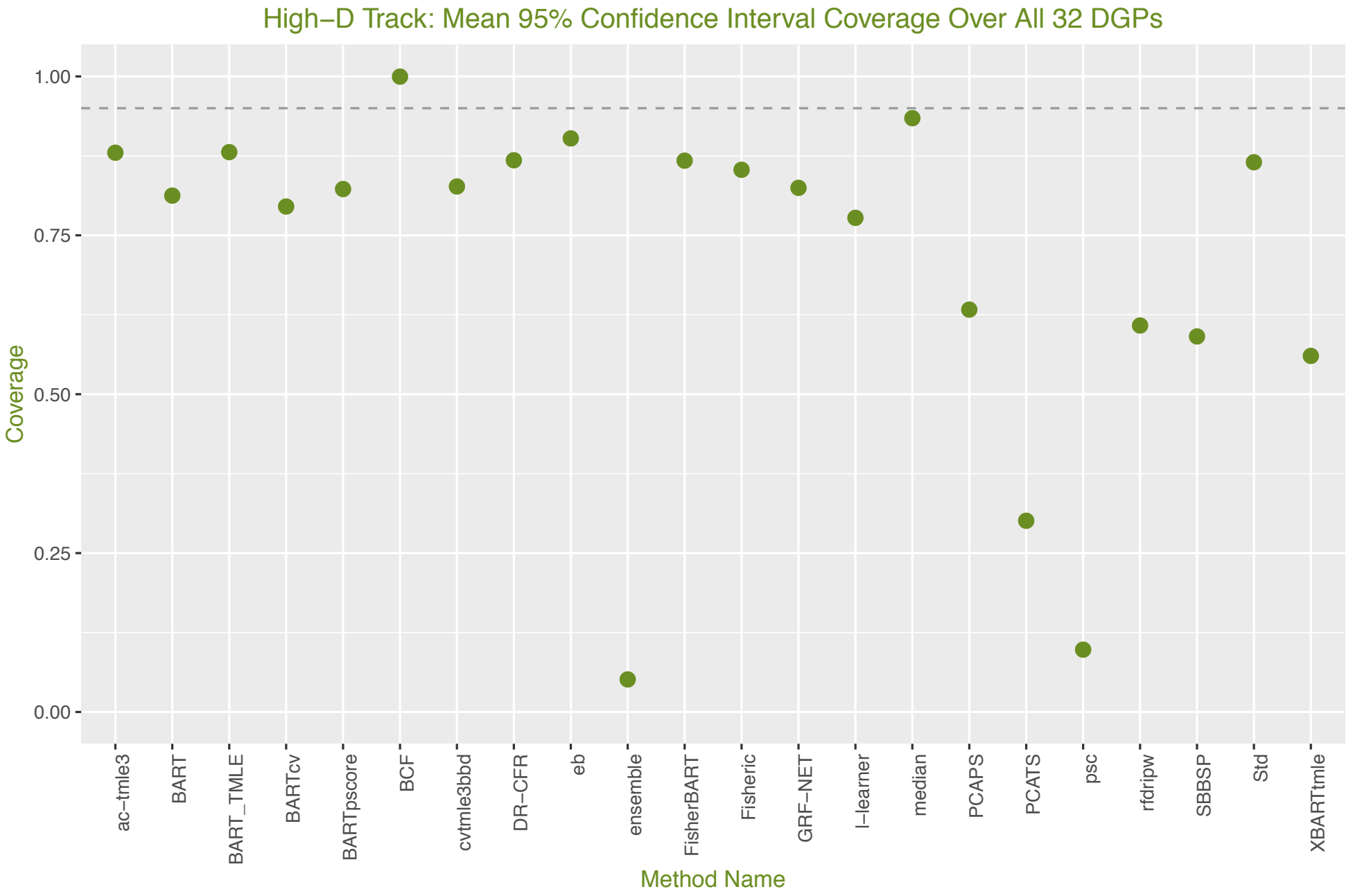
Low–Dim: Composite Bias Scores

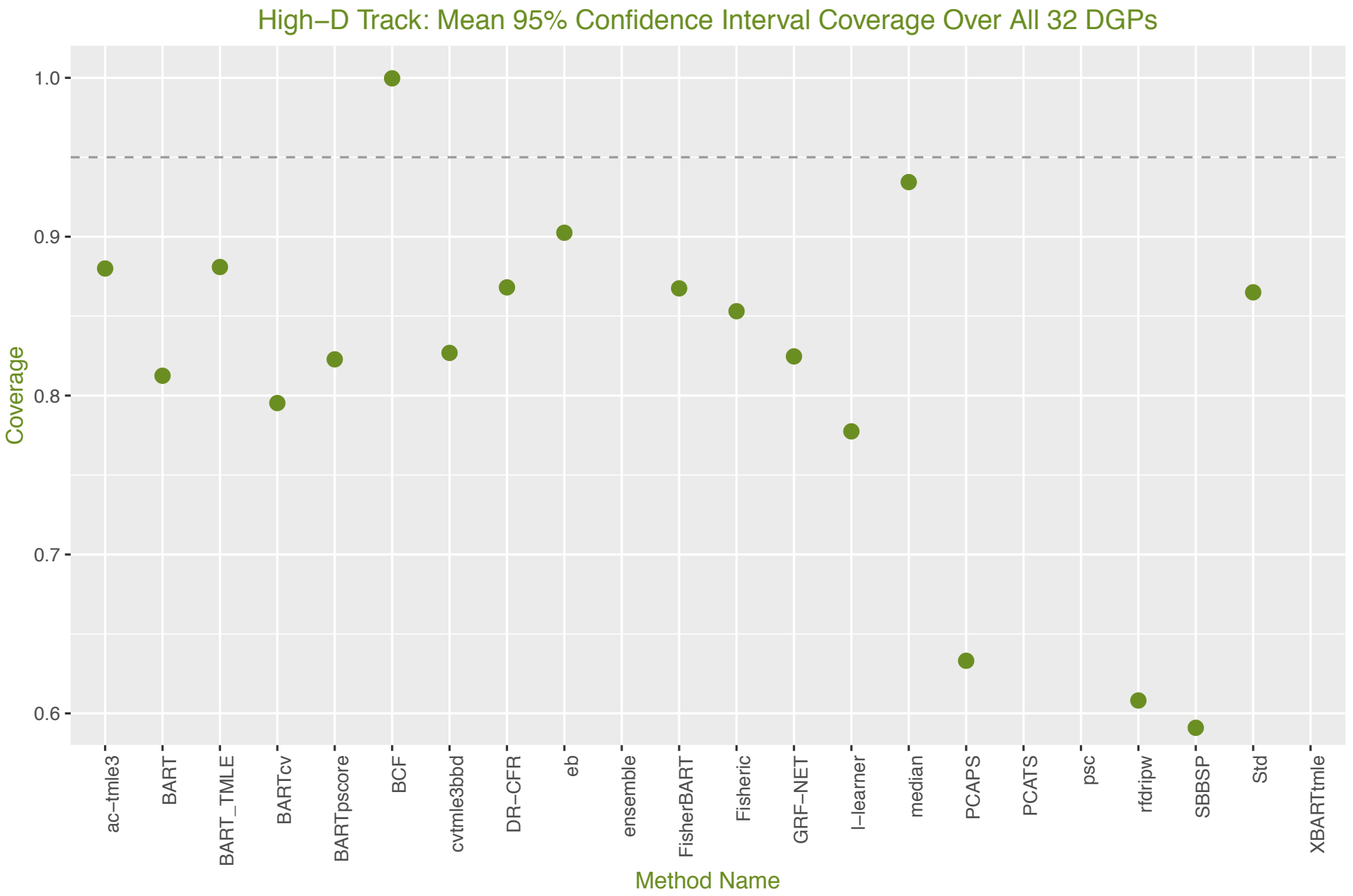


Low–Dim: Composite SD Scores

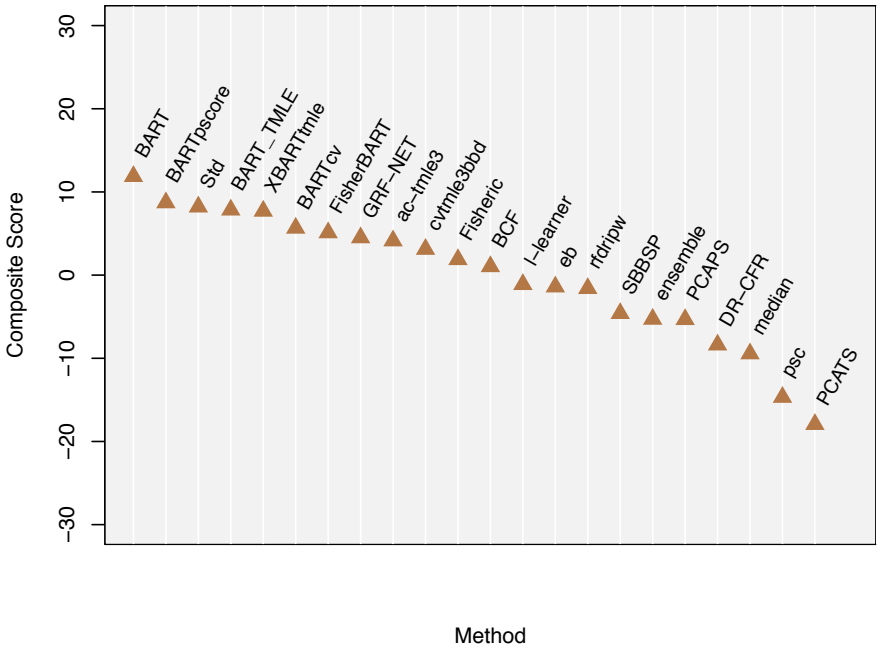



High–D Track: Mean Relative rMSE for All 32 DGPs
Mean Relative rMSE
15
10
5
0
ac–tmle3
BART
BART_TMLE
BARTcv
BARTpscore
BCF
cvtmle3bbd
DR–CFR
eb
ensemble
FisherBART
Fisheric
GRF–NET
l–learner
median
PCAPS
PCATS
psc
rfdripw
SBBSP
Std
XBARTtmle
Method Name
Relative Mean rMSE = mean(Method rMSE / Oracle rMSE), low values are best

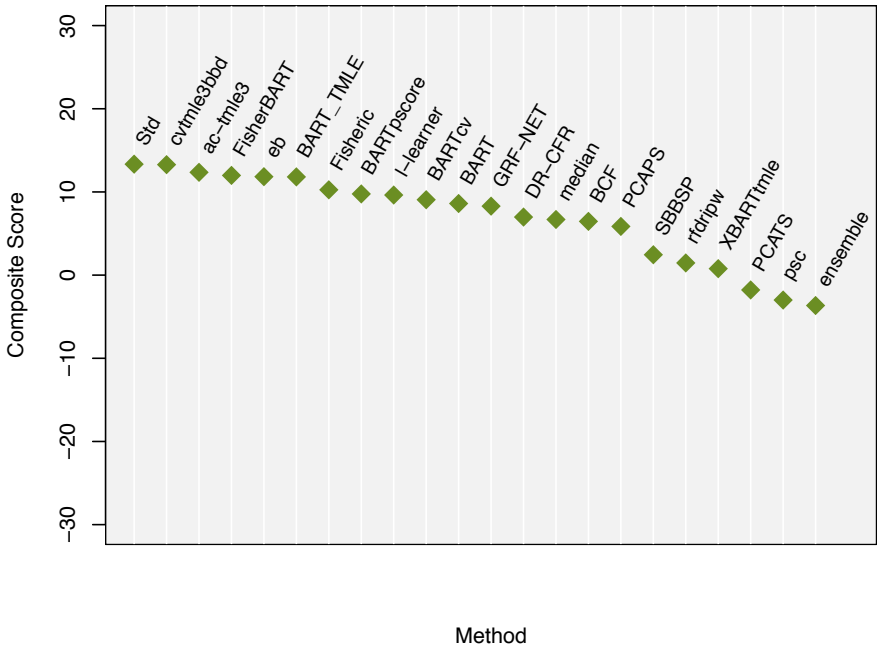

High−D Track: Mean 95% Confidence Interval Coverage Over All 32 DGPs
Coverage
1.00
0.75
0.50
0.25
0.00
ac−tmle3
BART
BART_TMLE
BARTcv
BARTpscore
BCF
cvtmle3bbd
DR−CFR
eb
ensemble
FisherBART
Fisheric
GRF−NET
l−learner
median
PCAPS
PCATS
psc
rfdripw
SBBSP
Std
XBARTtmle
Method Name


High–D Track: Mean 95% Confidence Interval Coverage Over All 32 DGPs
Coverage
1.0
0.9
0.8
0.7
0.6
ac–tmle3
BART
BART_TMLE
BARTcv
BARTpscore
BCF
cvtmle3bbd
DR–CFR
eb
ensemble
FisherBART
Fisheric
GRF–NET
l–learner
median
PCAPS
PCATS
psc
rfdripw
SBBSP
Std
XBARTtmle
Method Name

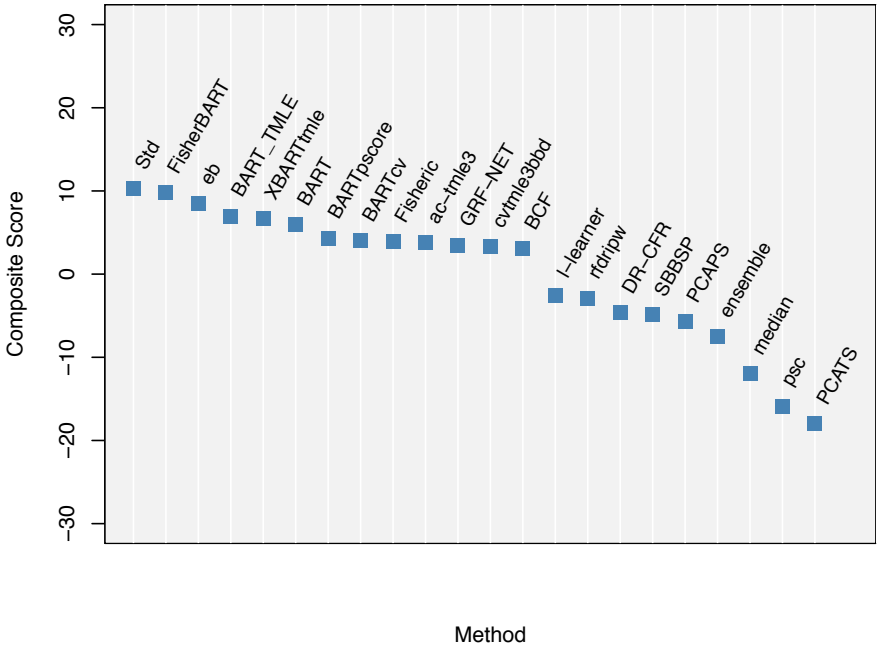
High–Dim: Composite rMSE Scores



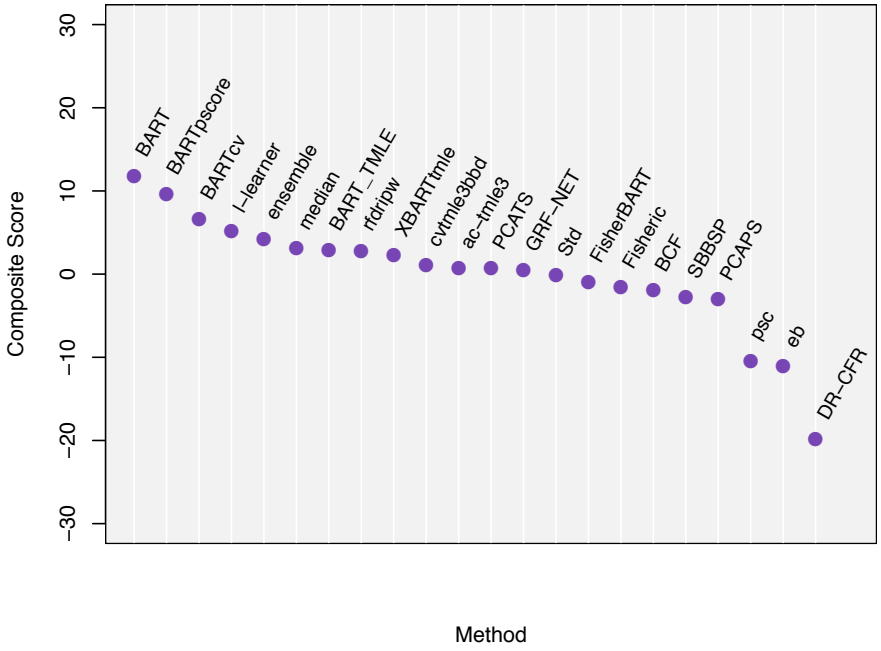
High–Dim: Composite Coverage Scores



High–Dim: Composite Bias Scores



High–Dim: Composite SD Scores



Some Questions for Discussion

1. What are other meaningful performance metrics?
2. What direction for future challenges?

   longitudinal data, right censoring, unmeasured confounding, ...
3. What else can we learn from the results?
4. Are observable characteristics of the data and methodologies a reliable guide for analytic choices?

Thank you!

- All Participants
- Google for site development tools and hosting
- UC Irvine, Vanderbilt University, Columbia University data repositories