Efficiency of Classification Algorithms as an Alternative to Logistic Regression in Propensity Score Adjustment for Survey Weighting

Ramón Ferri-García [1] María del Mar Rueda [1]

[1] Department of Statistics and Operational Research, University of Granada

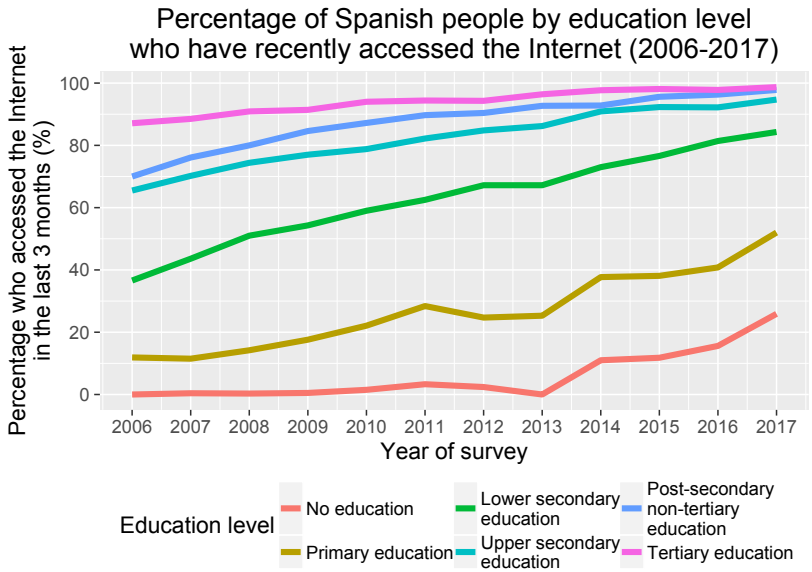
Barcelona, 27th October 2018

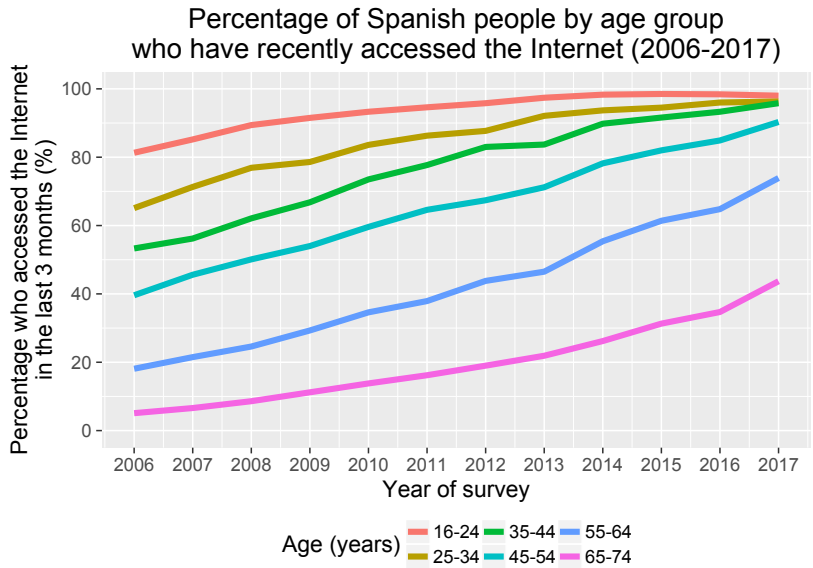
Outline

Selection bias in online surveys is a main concern:

- Participants are often self-selected (non-probabilistic sampling)
- Online sampling frames are available only for narrow populations (Schonlau and Couper, 2017)
- Some calibration procedures (GREG) are ineffective in removing volunteering bias (Dever et al., 2008)

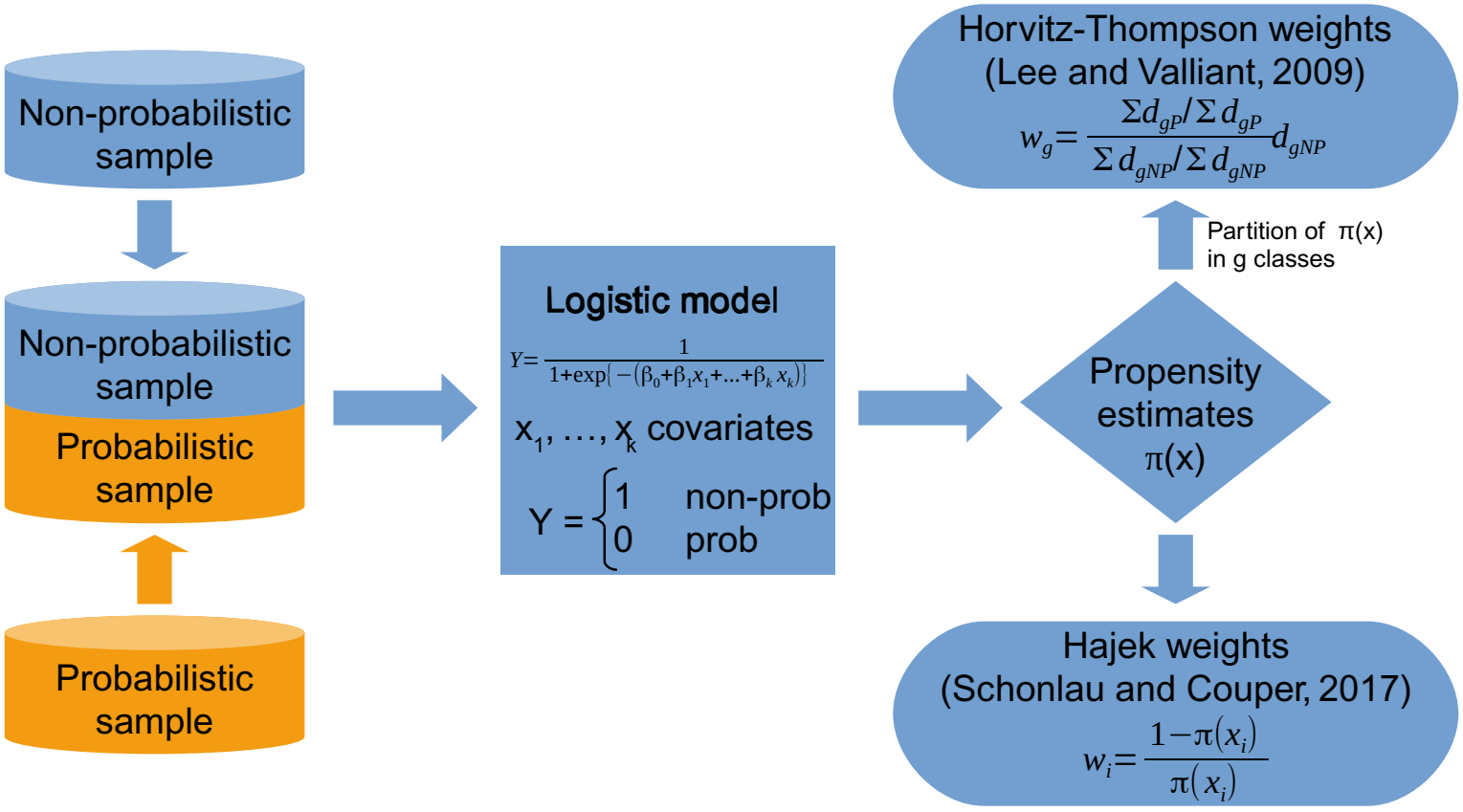
Coverage bias in online surveys is important and often associated to demographic variables





Source: Survey of information technologies in households, National Statistics Institute (INE)

Propensity Score Adjustment (PSA) has been proposed as a method for removing selection bias.



Propensity Score Adjustment (PSA) was originally developed for balancing experimental designs (Rosenbaum and Rubin, 1983), but it has been proposed as a method to remove selection bias in online surveys.

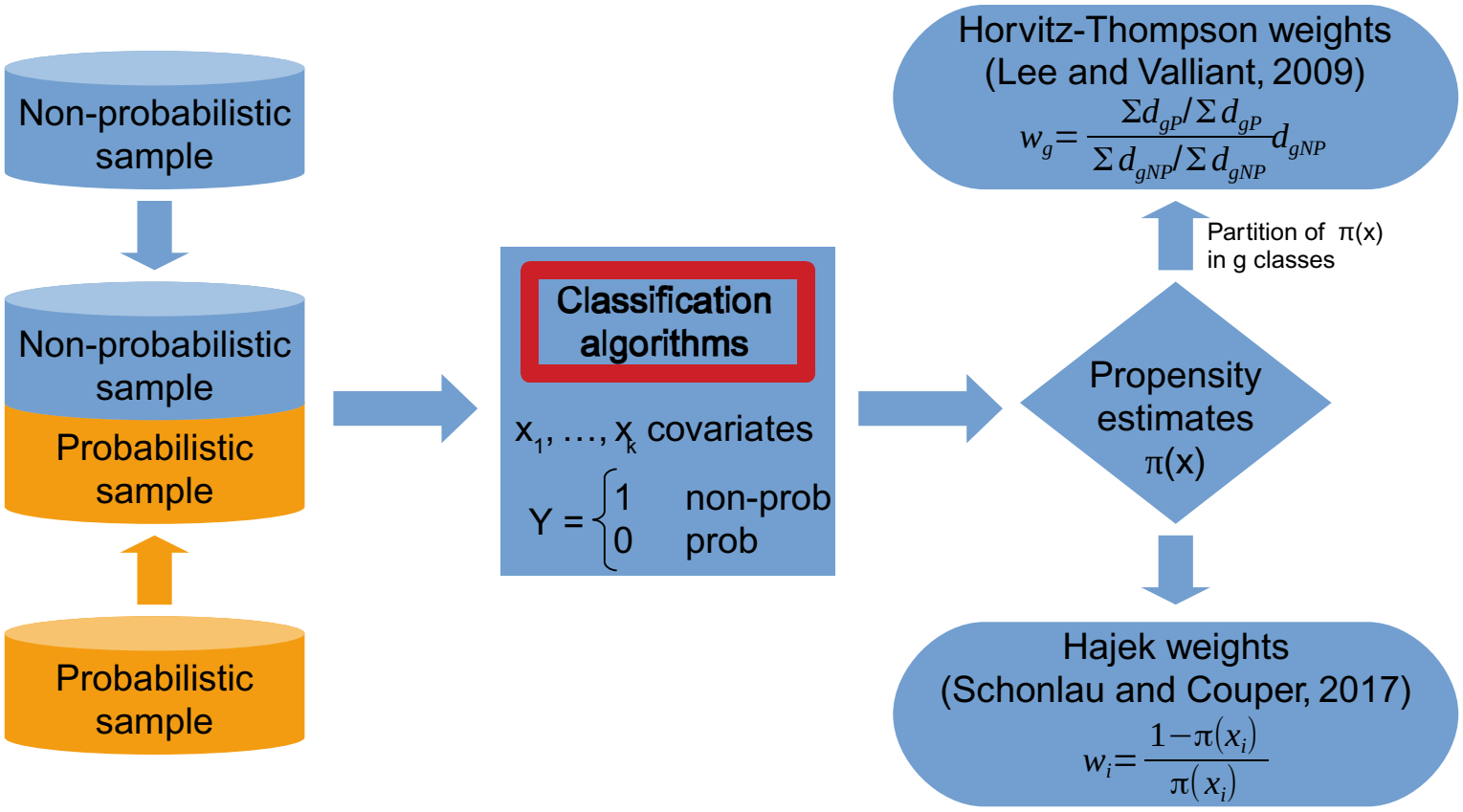
- Its application requires a probabilistic sample on which covariates have been measured.
- Has been proved as effective in online surveys at the cost of increasing variance (Lee, 2006; Lee and Valliant, 2009).
- Efficacy strongly dependent on covariates' choice and the relationship between Internet access and target variables (Valliant and Dever, 2011).

Logistic regression is the standard model in literature for propensity estimation. It provides robust and efficient weights, but it has several drawbacks:

- Requires linearity assumptions on risk
- Strongly dependent of functional form and shape of covariates
- Does not capture interactions on its usual configuration for PSA (D'Agostino, 1998; Westreich et al., 2010)

In addition, alternatives to logistic regression as a classificator, in terms of accuracy, have arisen over time on the experimental design context.

Proposal


Non-probabilistic sample
Non-probabilistic sample
Probabilistic sample
Probabilistic sample
Classification algorithms
$x_1, \ldots, x_k$ covariates
$Y = \begin{cases} 1 & \text{non-prob} \\ 0 & \text{prob} \end{cases}$
Propensity estimates $\pi(x)$
Partition of $\pi(x)$ in g classes
Horvitz-Thompson weights (Lee and Valliant, 2009)
$w_g = \frac{\Sigma d_{gP} / \Sigma d_{gP}}{\Sigma d_{gNP} / \Sigma d_{gNP}} d_{gNP}$
Hajek weights (Schonlau and Couper, 2017)
$w_i = \frac{1-\pi(x_i)}{\pi(x_i)}$

The use of Machine Learning (ML) classification algorithms for propensity score estimation has been studied in experimental design:

- Decision trees (Setoguchi et al., 2008; Lee et al., 2010; Watkins et al., 2013; Wyss et al., 2014; Linden and Yarnold, 2017)
- Neural networks (Cavuto et al., 2006; Glynn et al., 2006)
- Boosting (McCaffrey et al., 2013; Pirracchio et al., 2014; Watkins et al., 2013; Zhao et al., 2016)

However, research on PSA with ML in survey estimation is sparse.

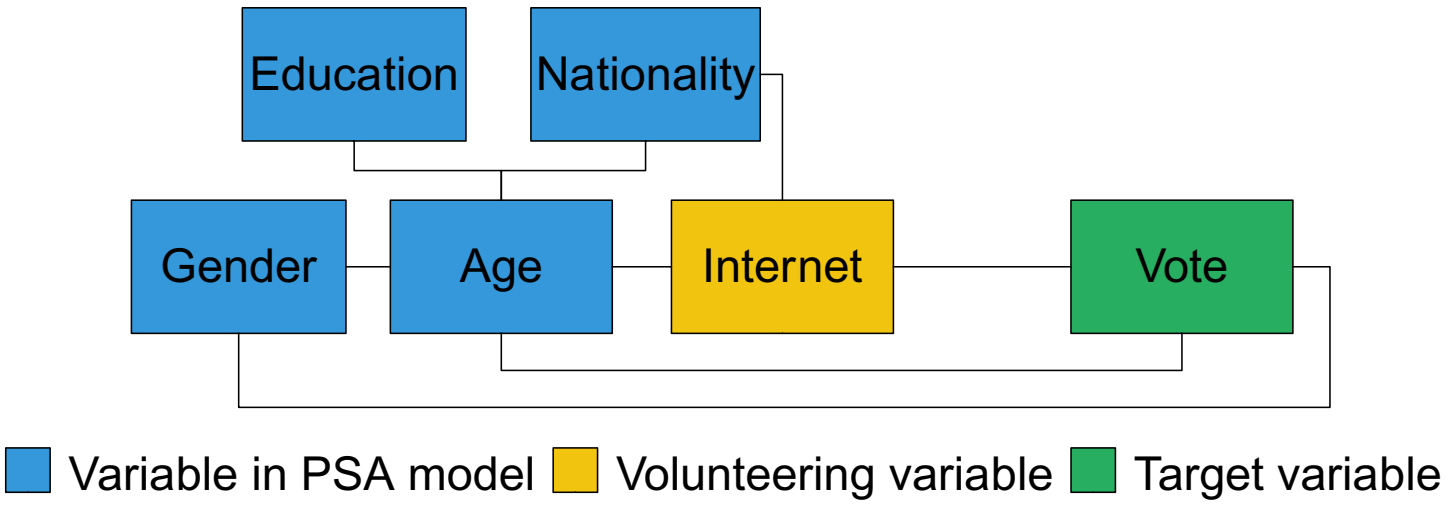
Simulation with artificial data

Simulation structure from Ferri-García and Rueda (2018) which is similar to the Bethlehem (2010) simulation with several modifications

- Population size of $N = 50000$ from which, in each simulation:
  - A probabilistic (reference) sample of $n_{rs} = 500$ is extracted
  - An Internet people (volunteer) sample of $n_{vs} = 500, 750, 1000, 2000, 5000, 7500, 10000$ is extracted
- 4 covariates (age, gender, nationality and education) and a target variable (voting intention) with 3 options:
  - Party 1 (dependent on gender)
  - Party 2 (dependent on age)
  - Party 3 (dependent on Internet access)

Simulation with artificial data

Relationships between variables can be observed in the following diagram:



Simulation with artificial data

Reweighting with PSA (Hajek-type weights) was performed over 500 simulations with several algorithms and parameters:

- Decision trees: J48, C5.0 and CART
  - Min. num. of obs. per node: $0.5\%, 1\%, 5\%$ of dataset
  - Confidence for pruning: $0.1, 0.25, 0.5$
- K-nearest neigbours with $k = 3, 5, 7, 9, 11, 13$
- Naïve Bayes with Laplace smoothing: $k = 1, 2, 5, 10$
- Boosting:
  - Random Forest with num. of trees $= 100, 200, 500$
  - GBM with depth $= 4, 6, 8$ and learning rate $= 0.1, 0.01, 0.001$

Simulation with real data

Spanish Life Conditions Survey (2012 edition) dataset as pseudopopulation:

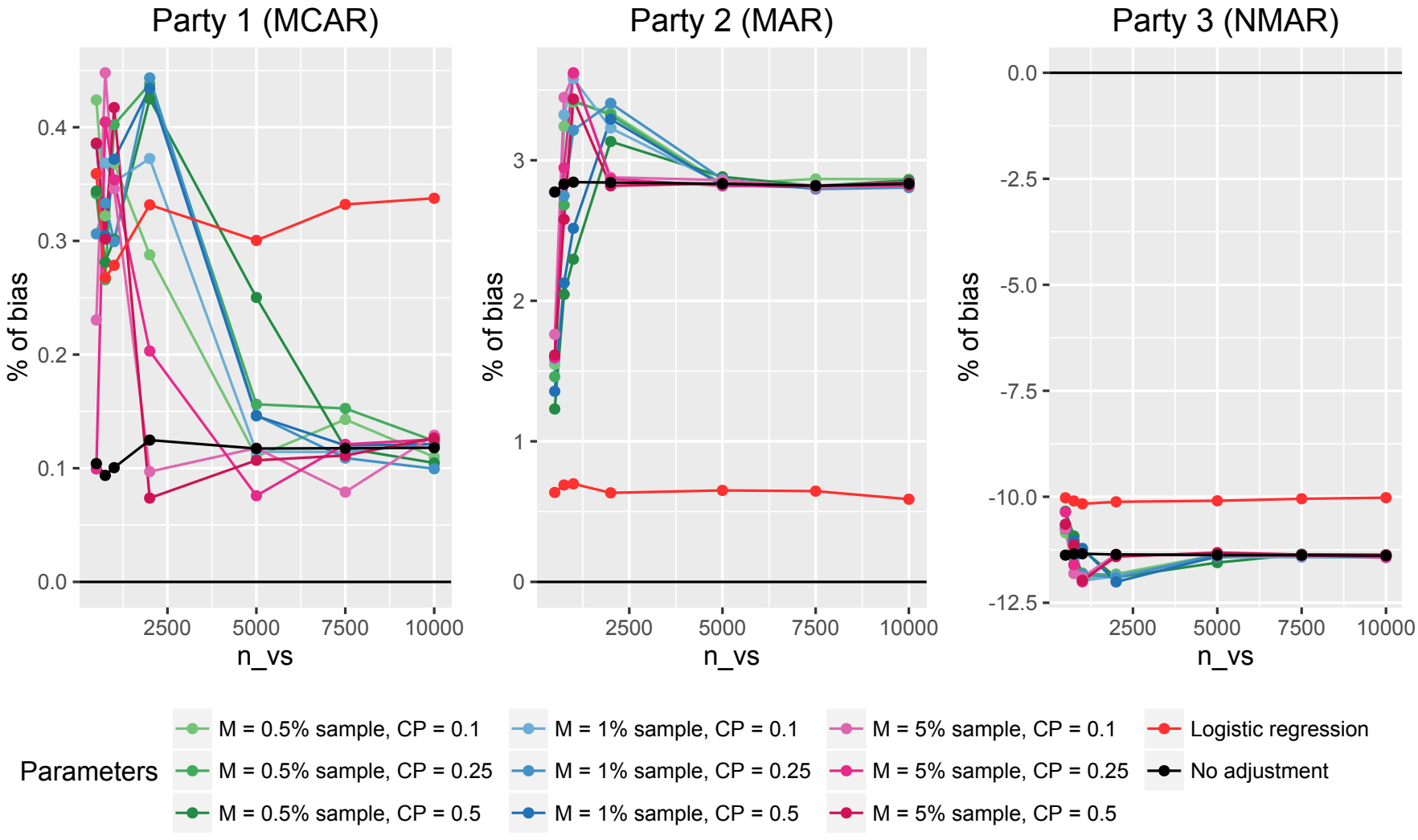
- $N = 28210$ and 61 potential covariates after assessing for missing data.
- Owning of computer at home as volunteering variable
- Two target variables:
  - Self-reported health: bad/not bad (MAR)
  - $>$2 members in household (NMAR)
- Four groups of covariates:
  - G1: demographic variables
  - G2: G1 + health-related variables
  - G3: G1 + poverty-related variables
  - G4: all potential covariates

Simulation with real data

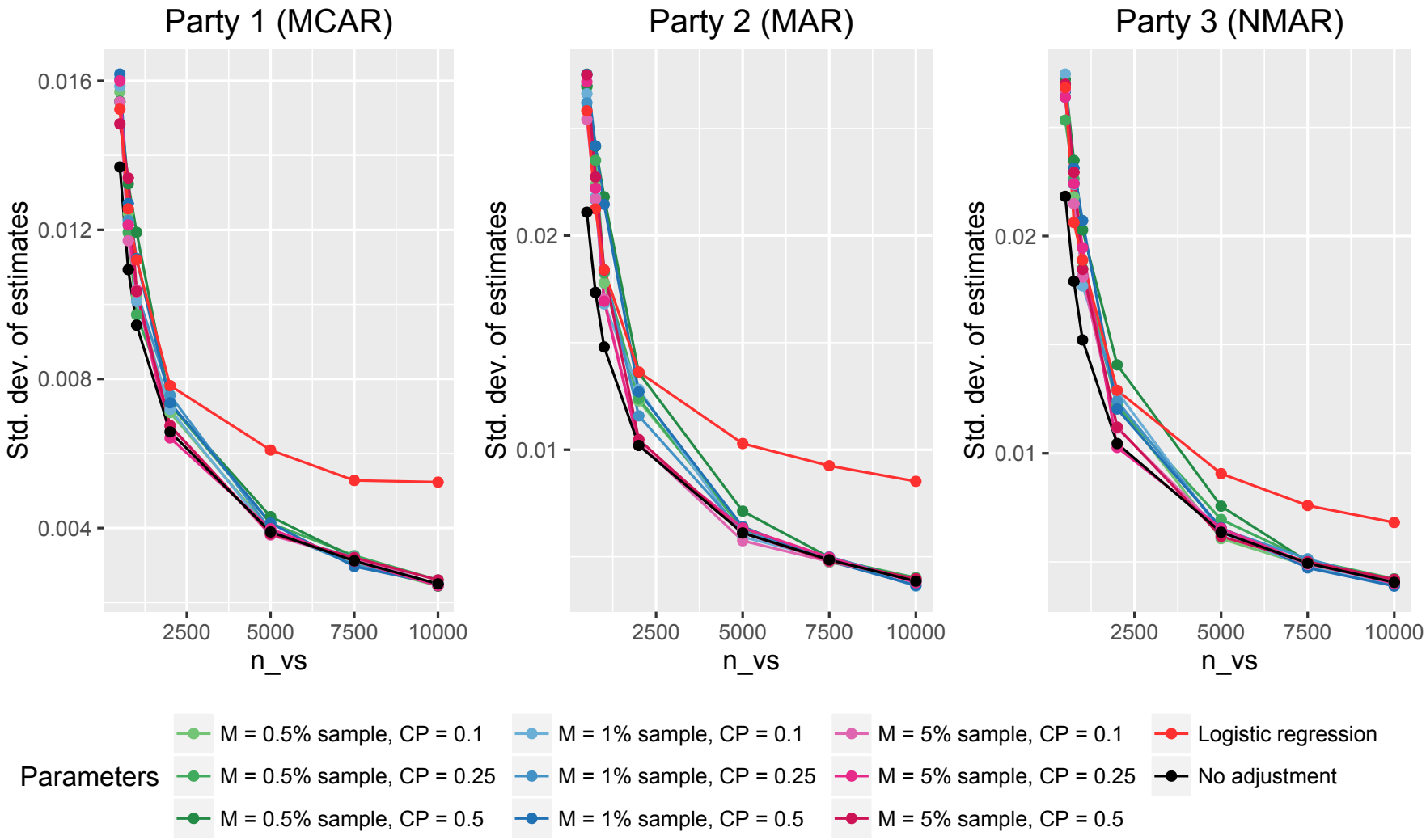
Reweighting with PSA (Hajek-type weights) was performed over 500 simulations, with $n_{vs} = 500, 750, 1000, 2000, 5000$ and the following algorithms and parameters:

- Decision trees: J48, C5.0 and CART with default parameters
- K-nearest neigbours with $k = 3$
- Naïve Bayes without Laplace smoothing
- Boosting:
  - Random Forest with 100 trees
  - GBM with interaction depth (ID) $= 8$ and learning rate (LR) $= 0.001$

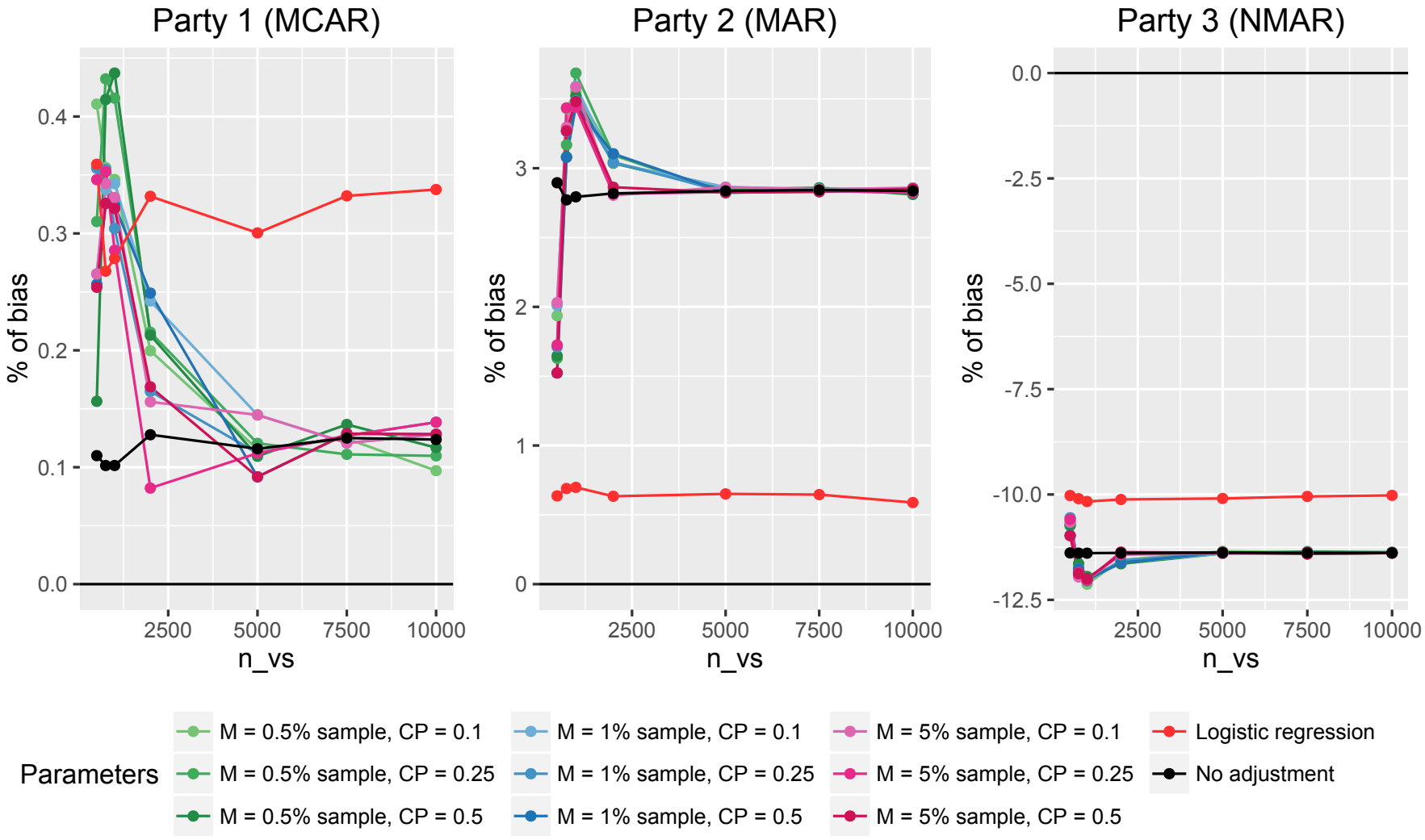
PSA with J48 in artificial data



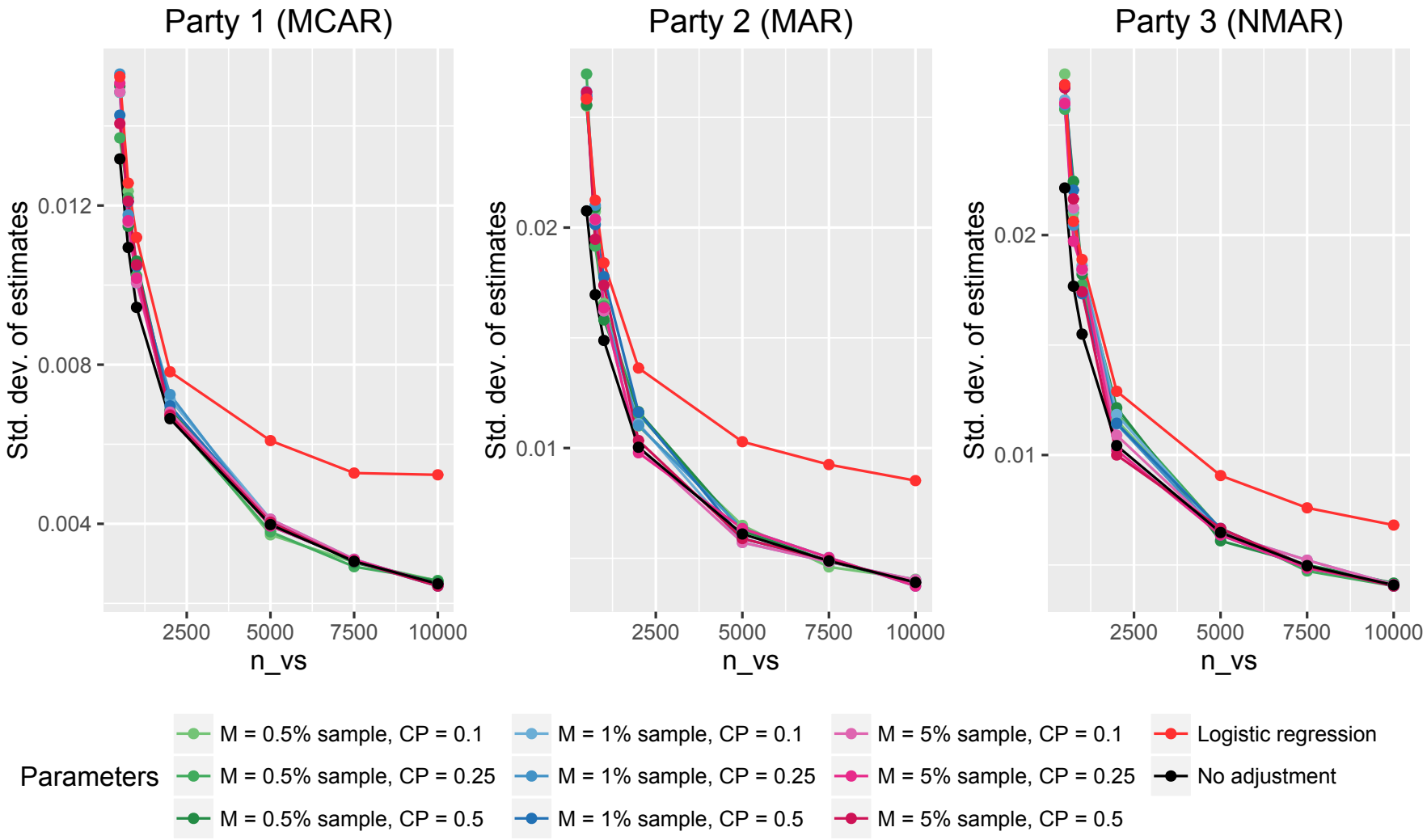
PSA with J48 in artificial data



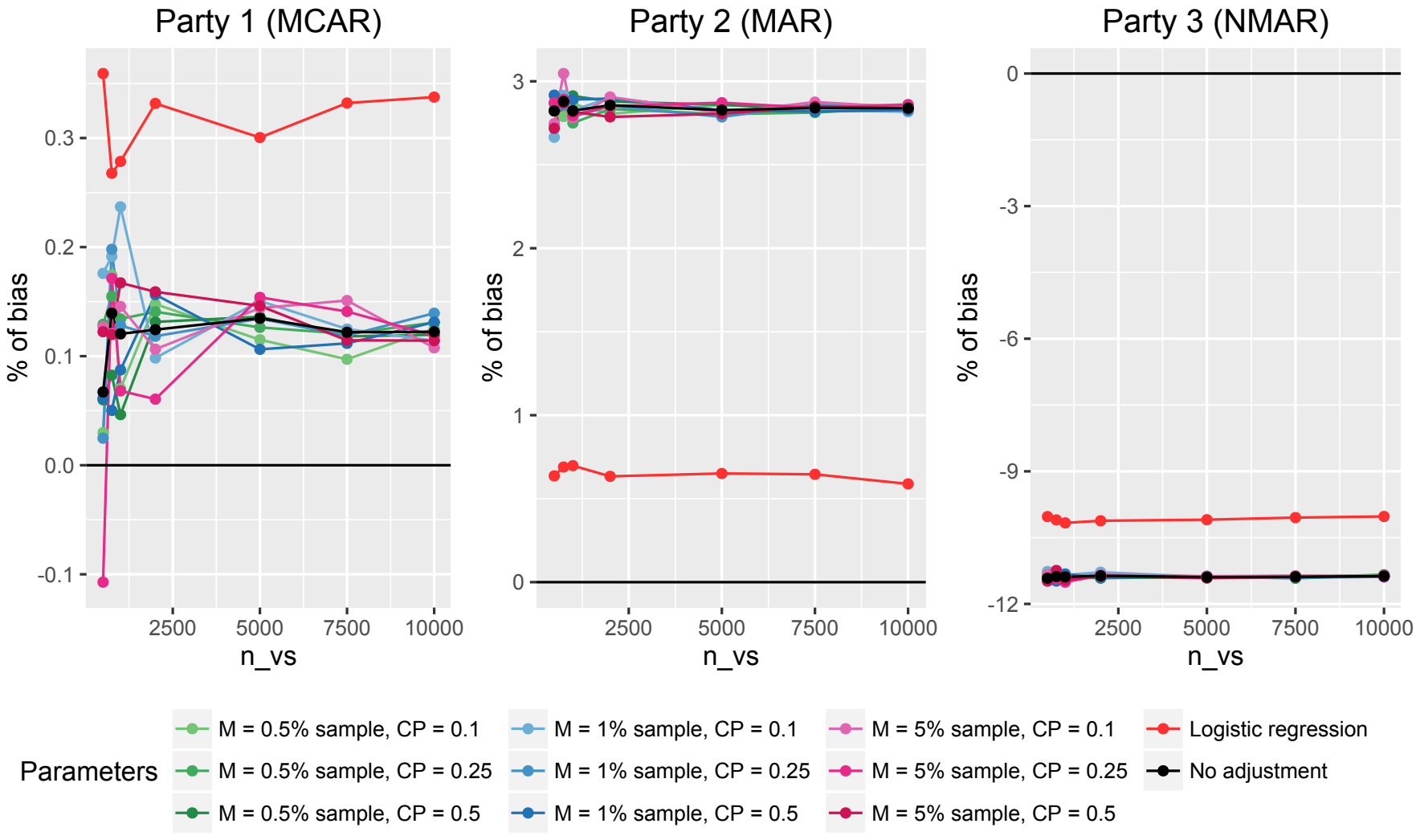
PSA with C5.0 in artificial data



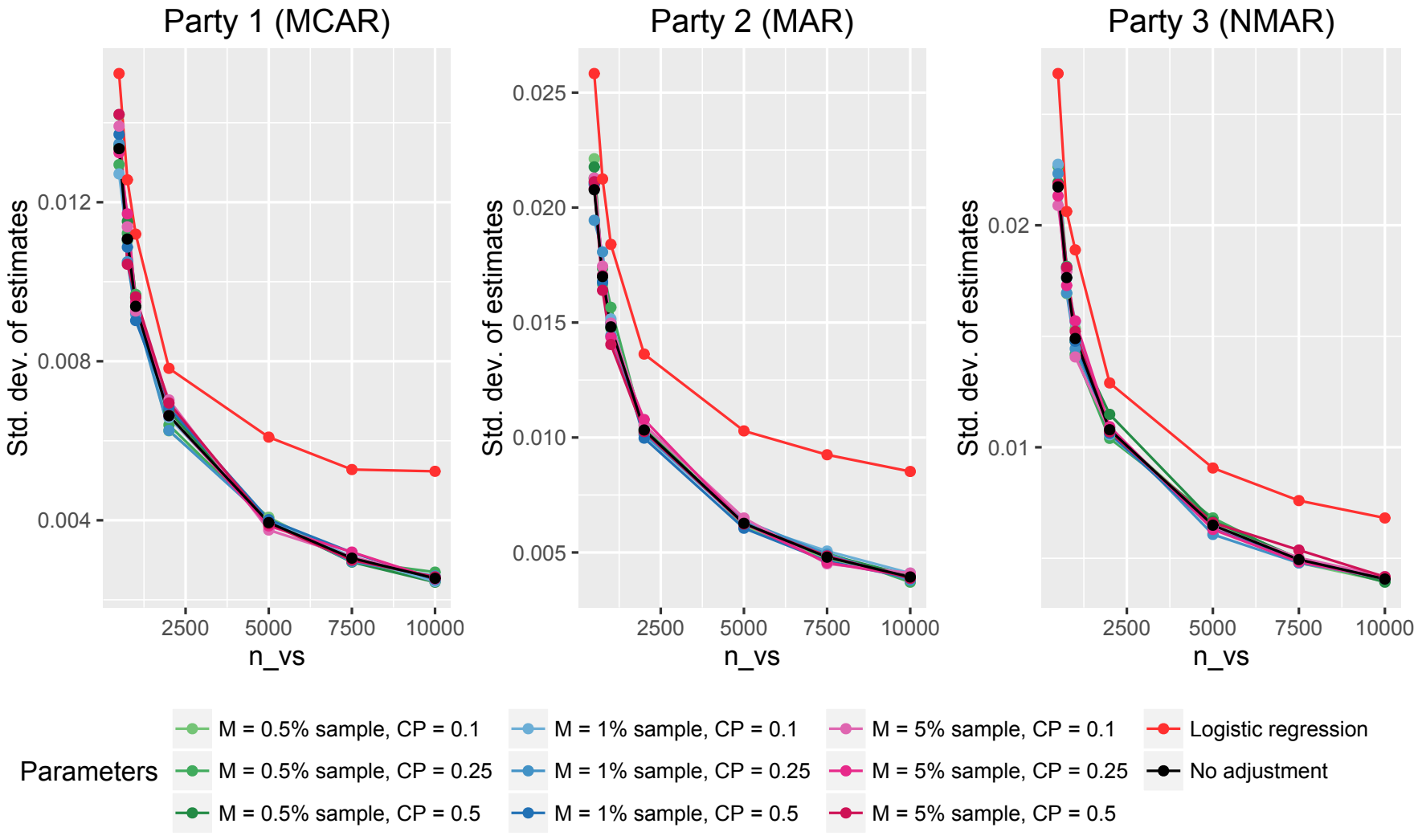
PSA with C5.0 in artificial data



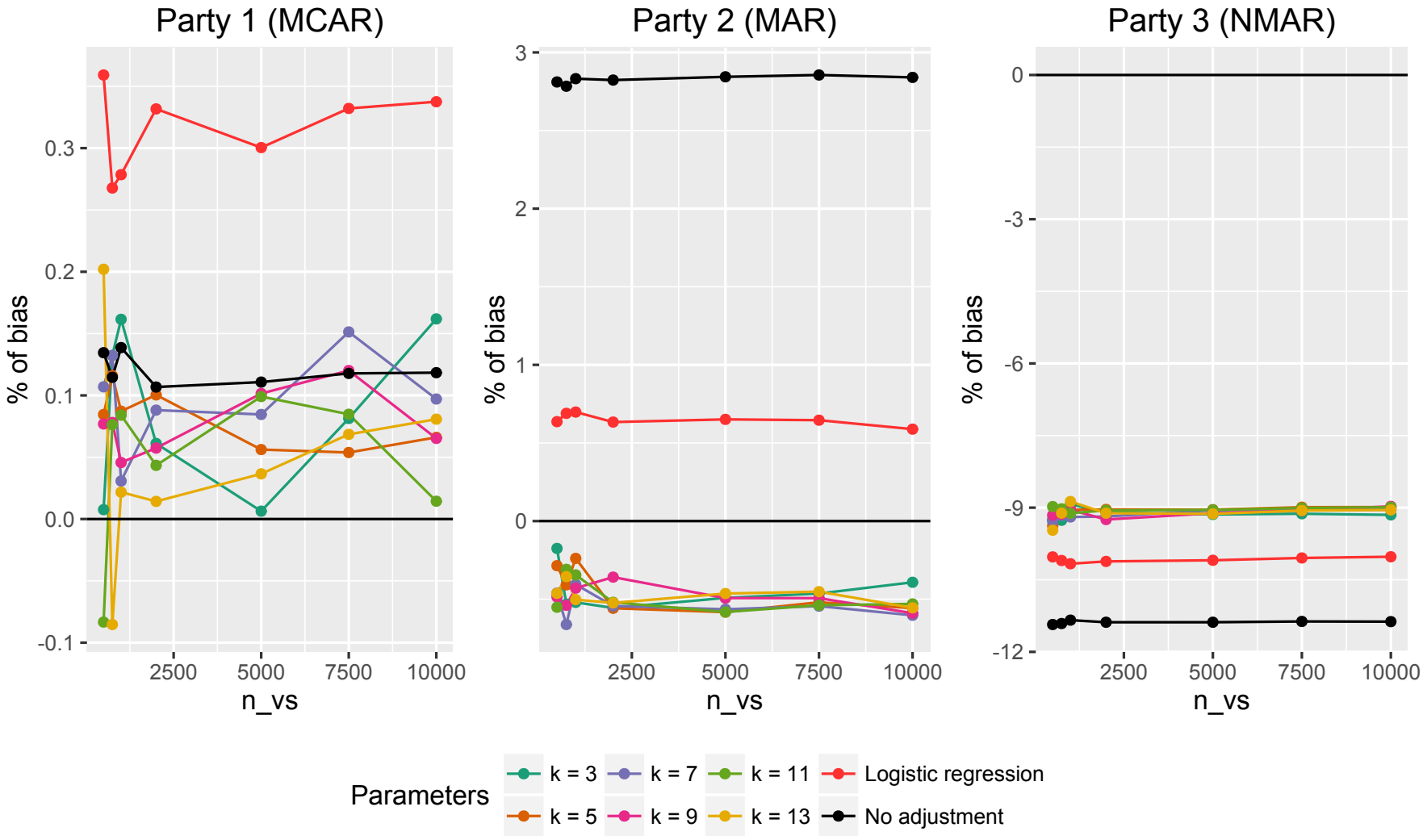
PSA with CART in artificial data



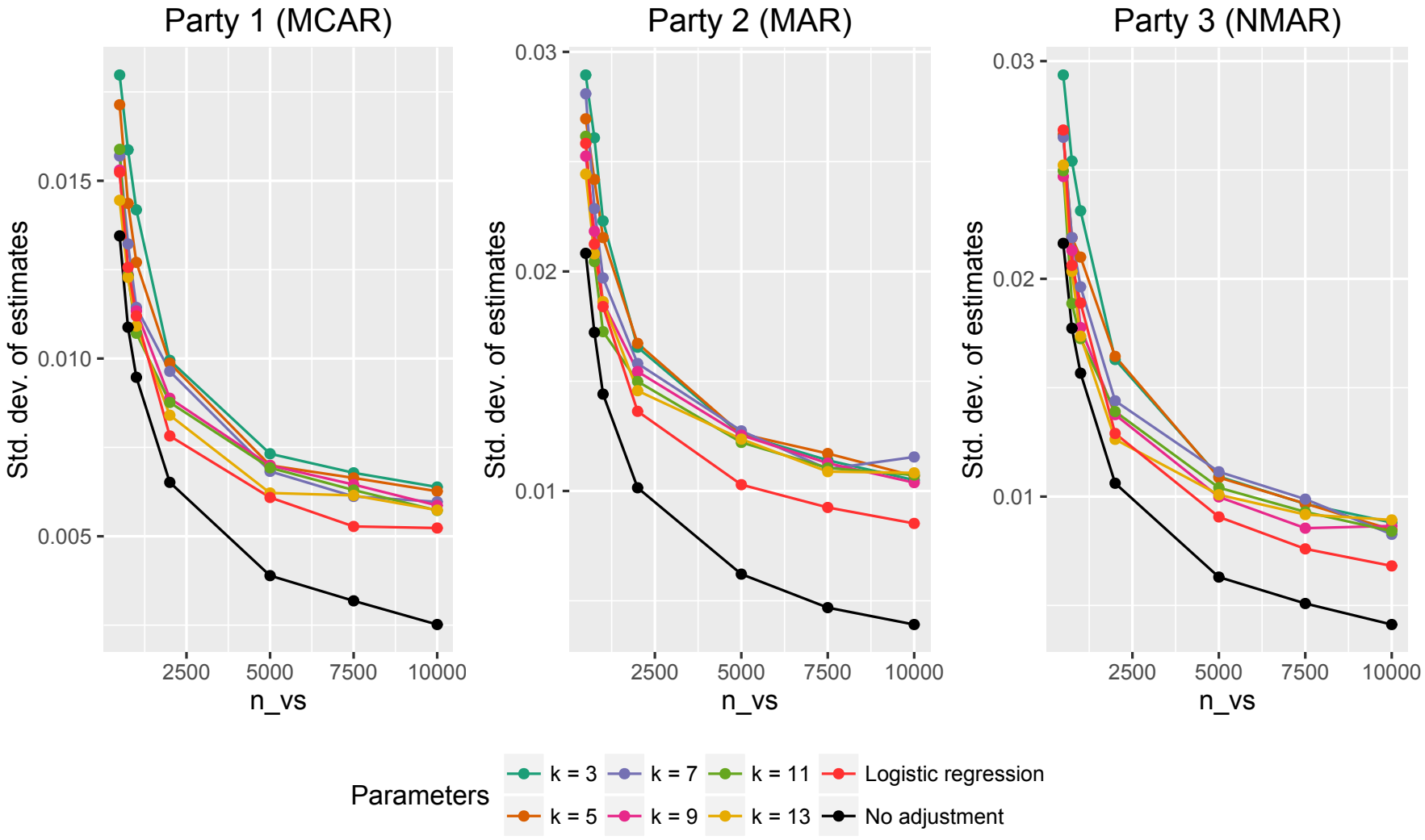
PSA with CART in artificial data



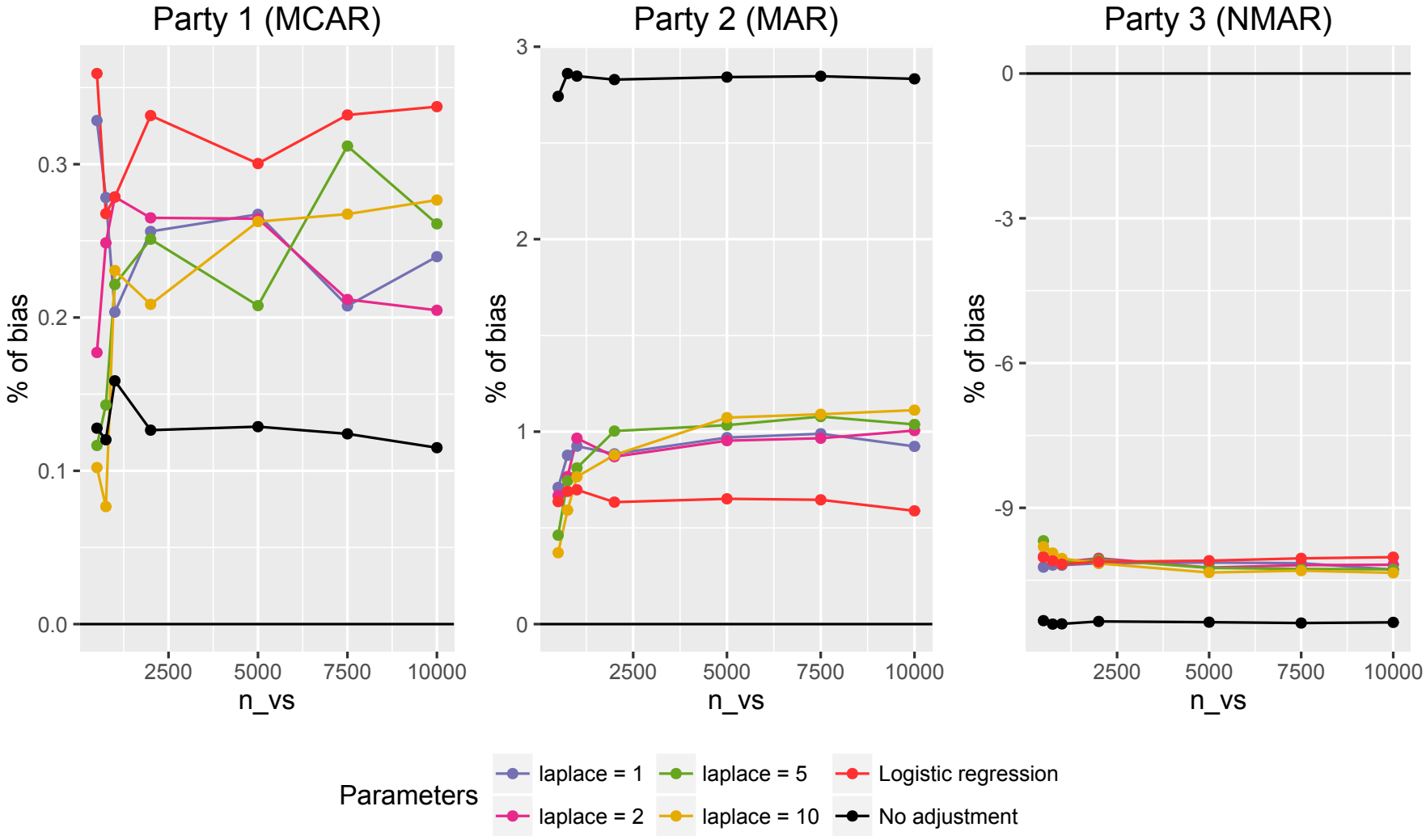
PSA with kNN in artificial data



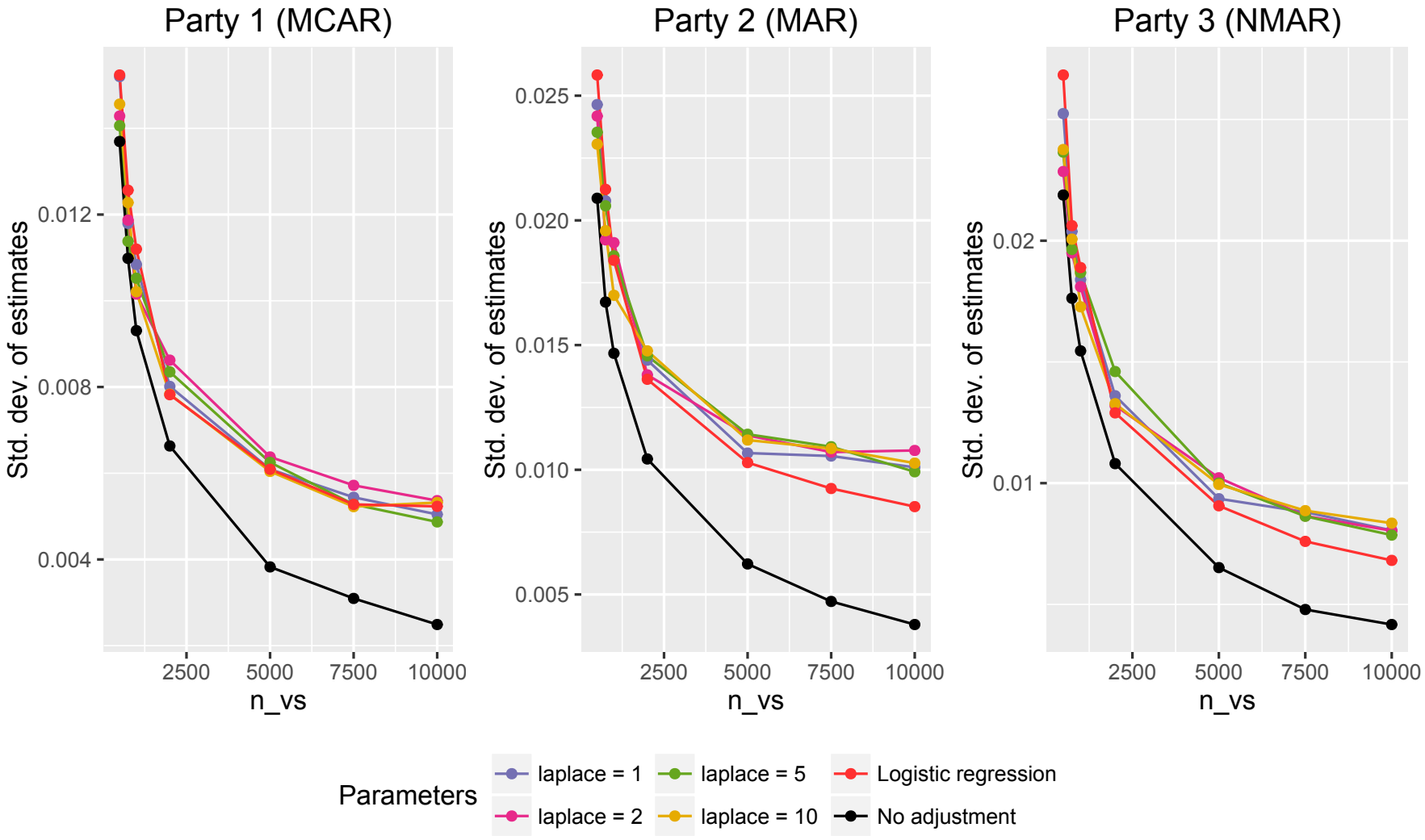
PSA with kNN in artificial data



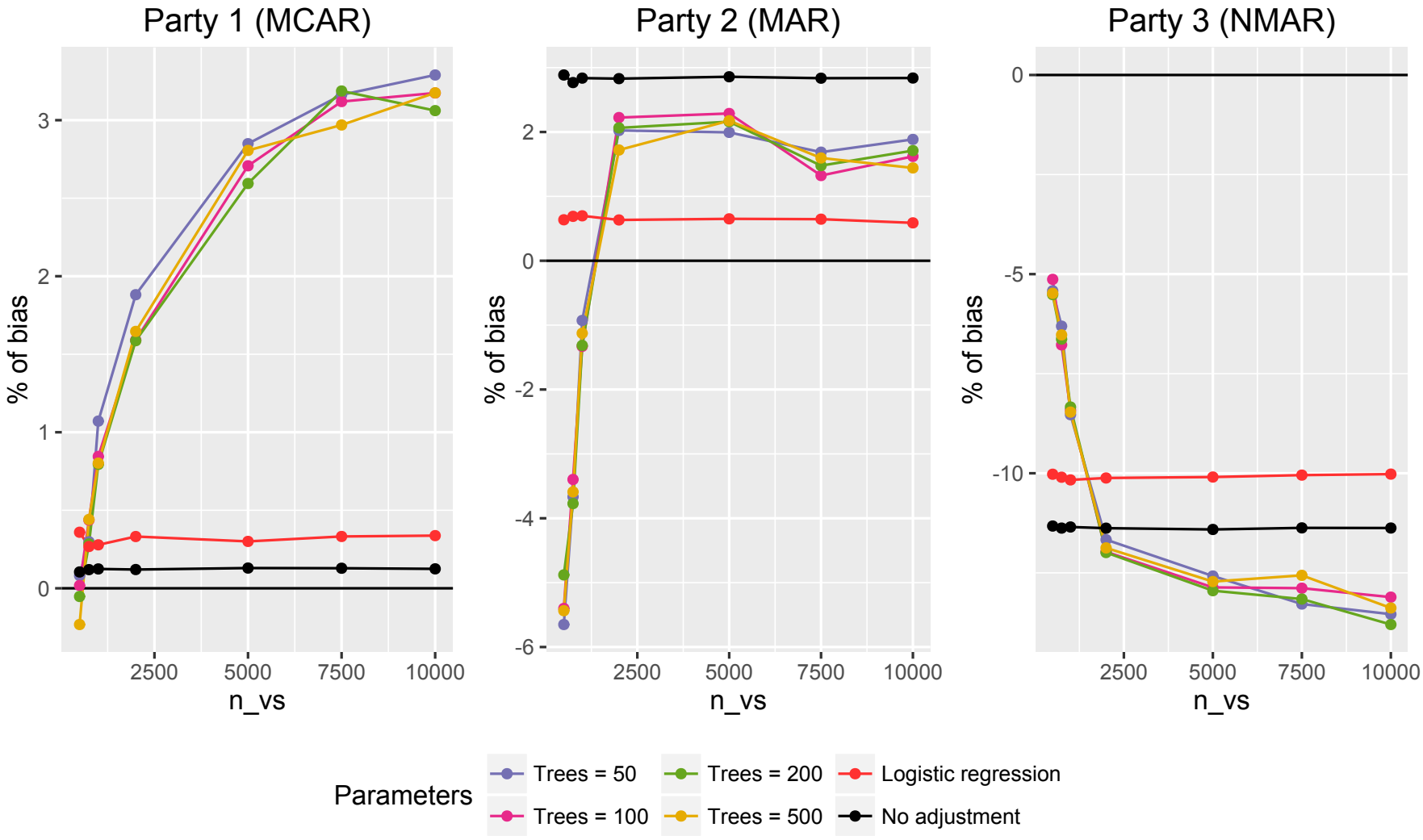
PSA with Naïve Bayes in artificial data



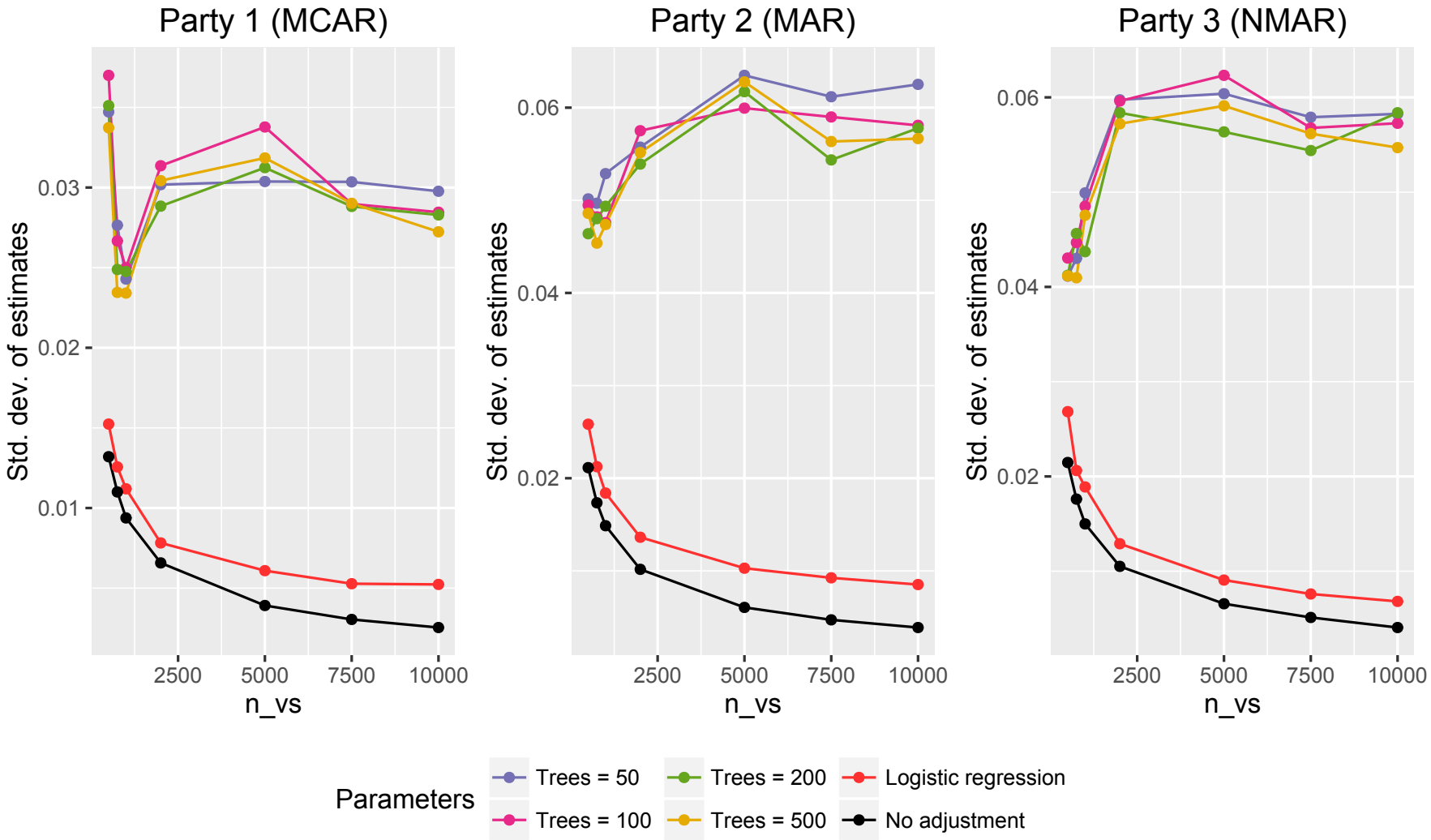
PSA with Naïve Bayes in artificial data



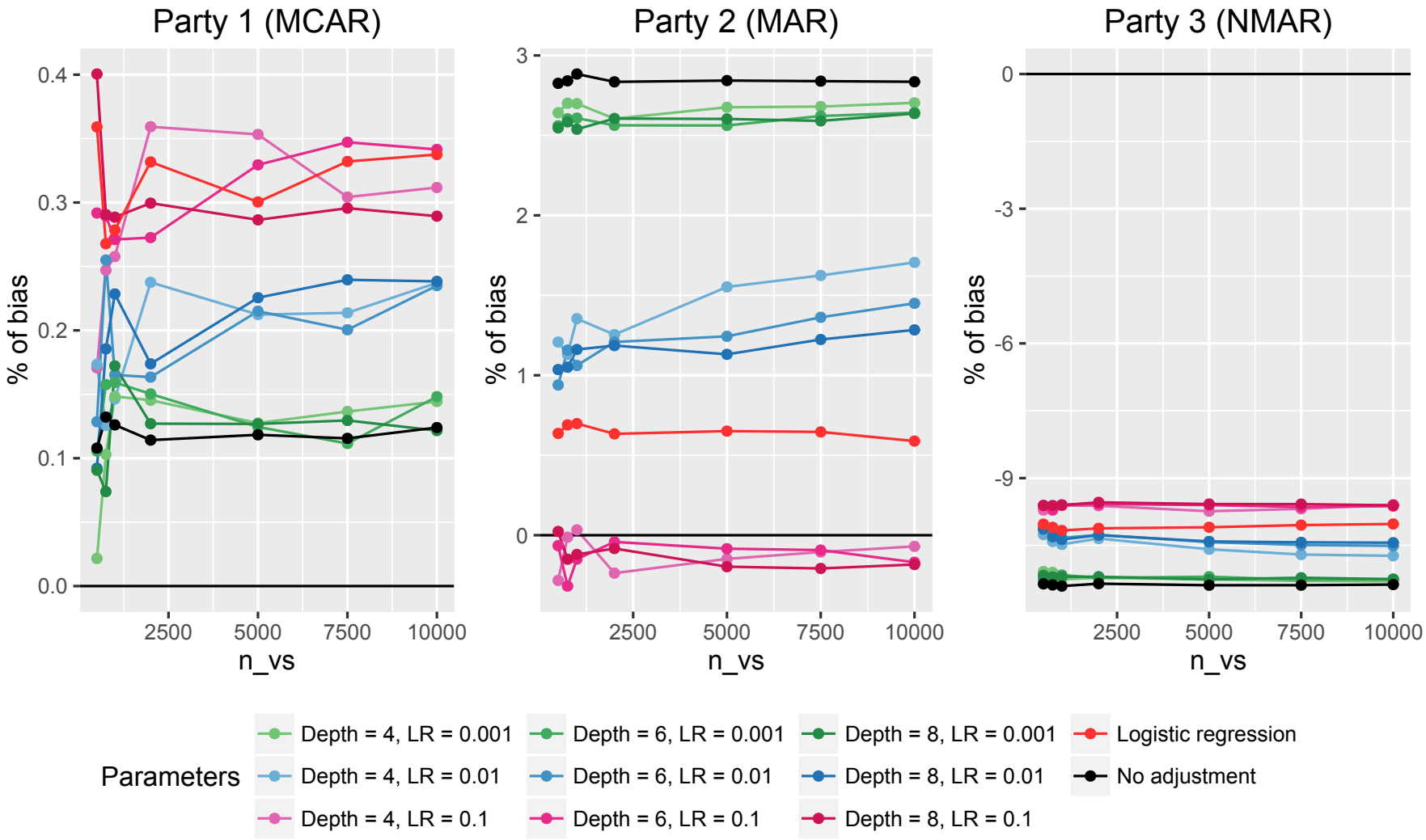
PSA with Random Forest in artificial data



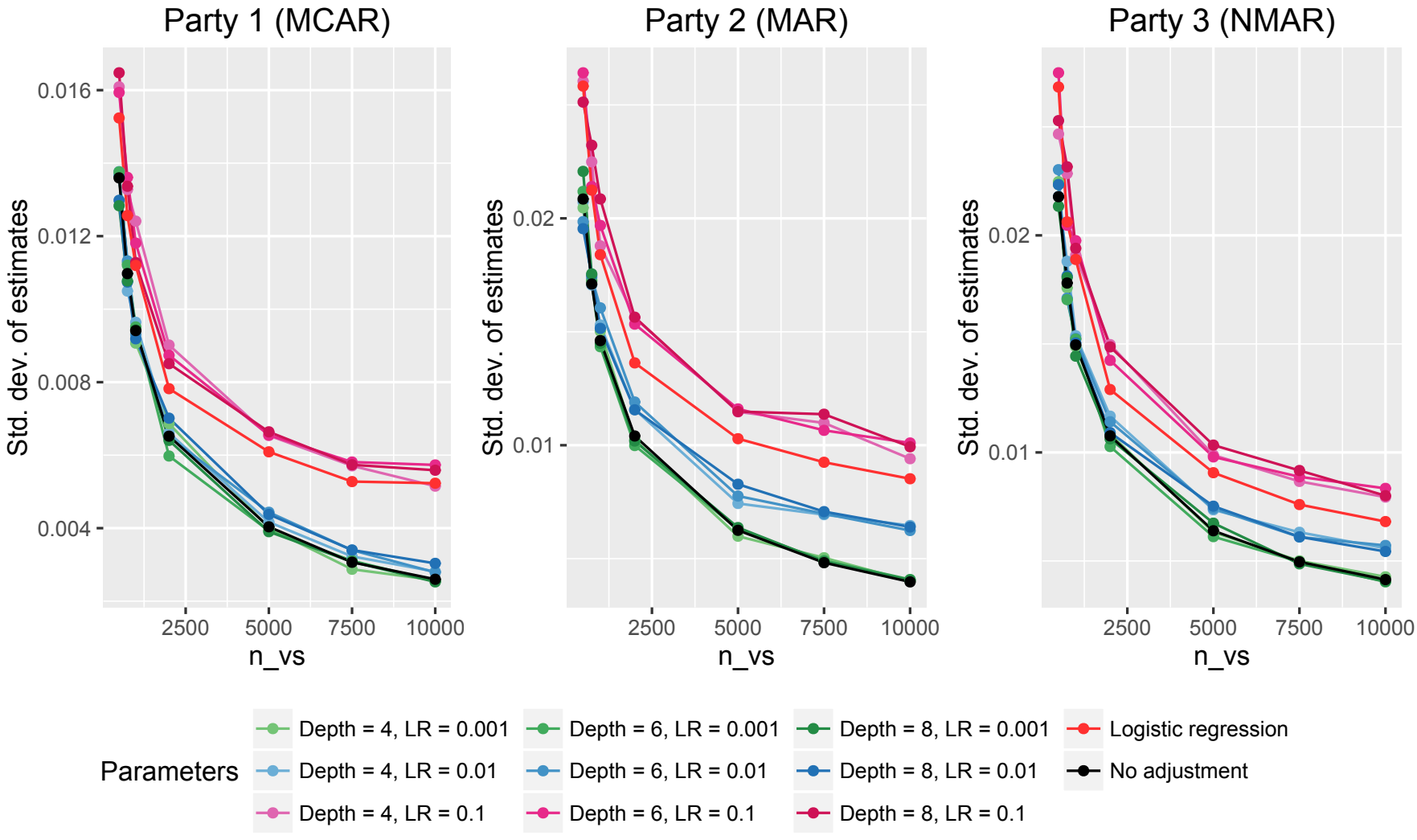
PSA with Random Forest in artificial data



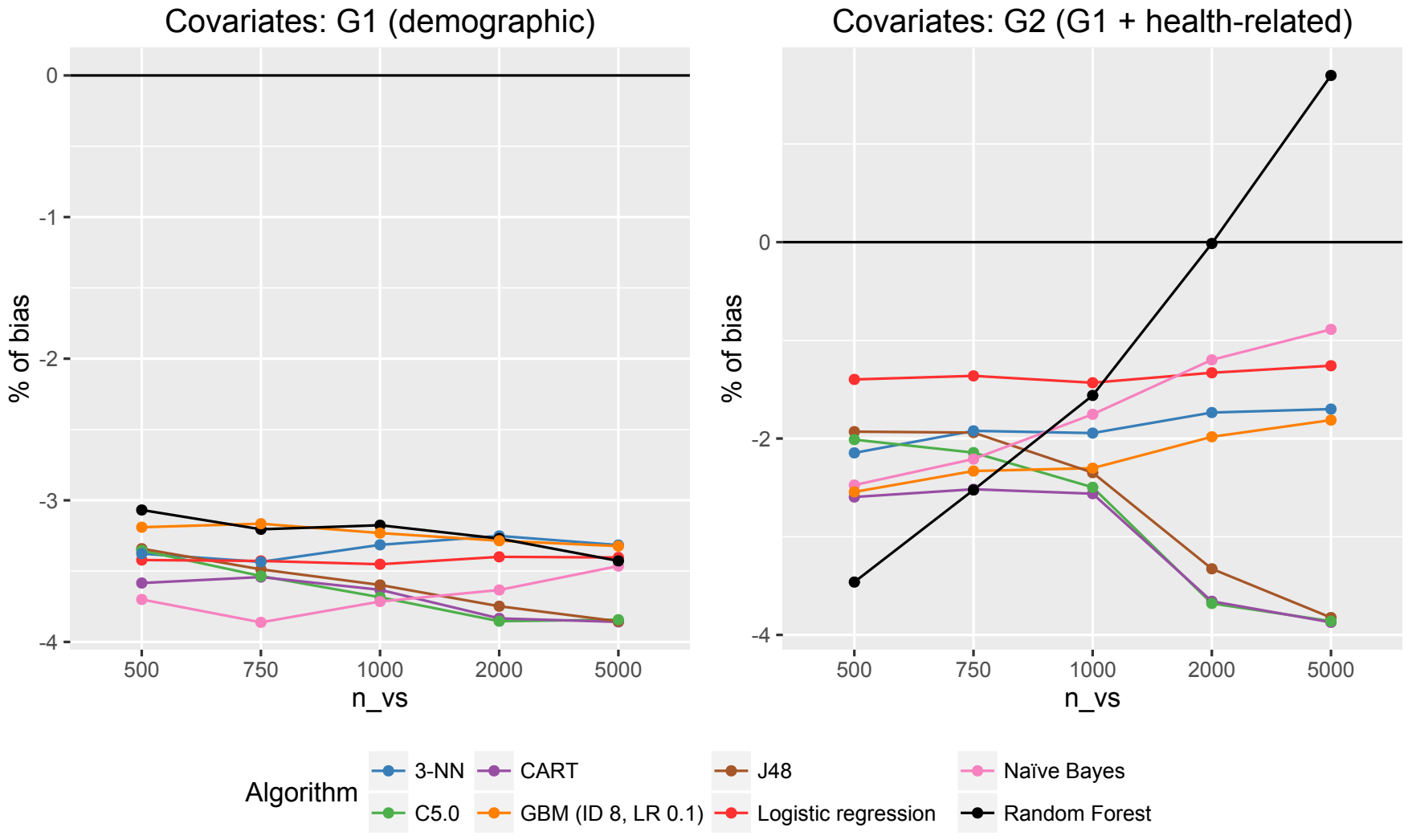
PSA with GBM in artificial data



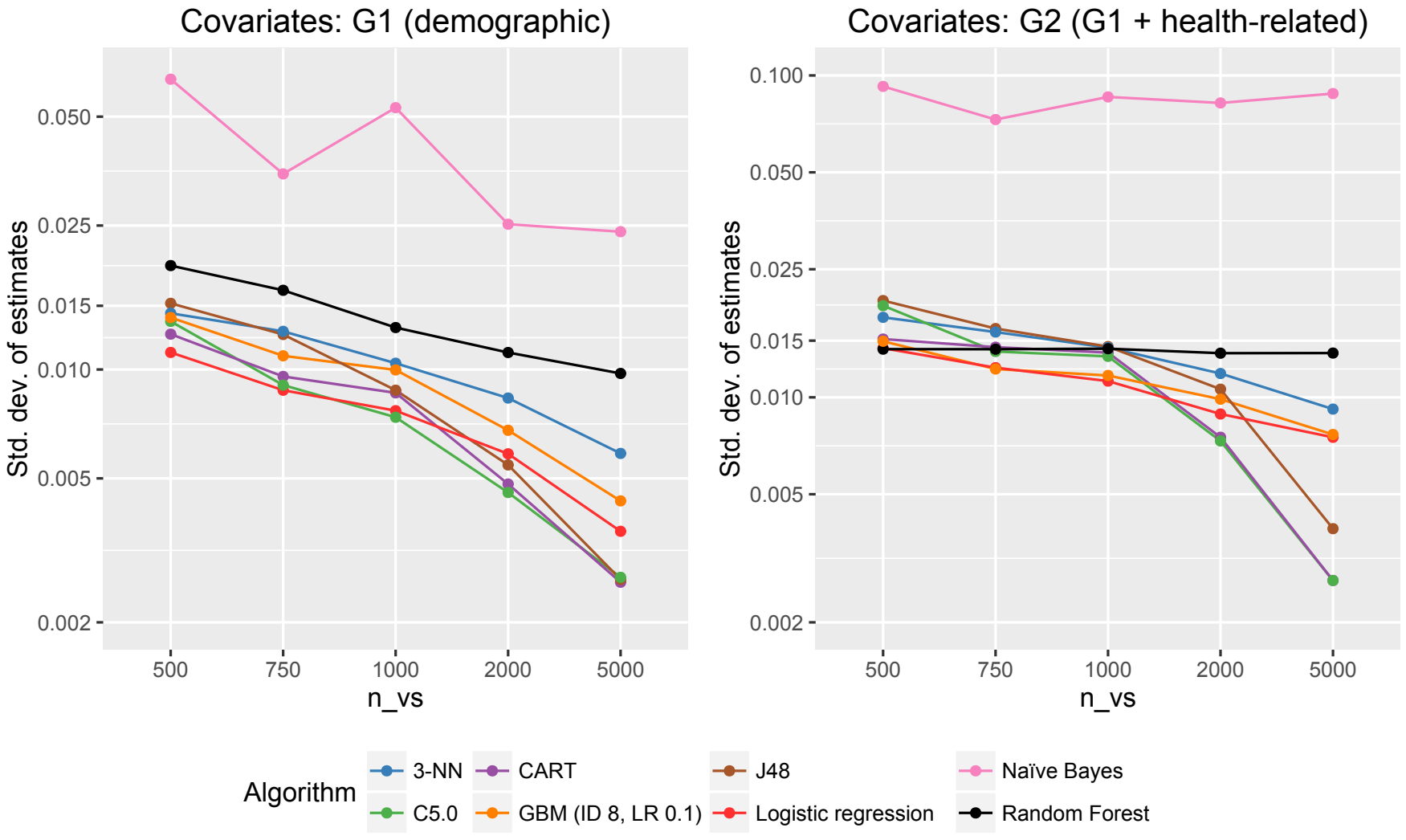
PSA with GBM in artificial data



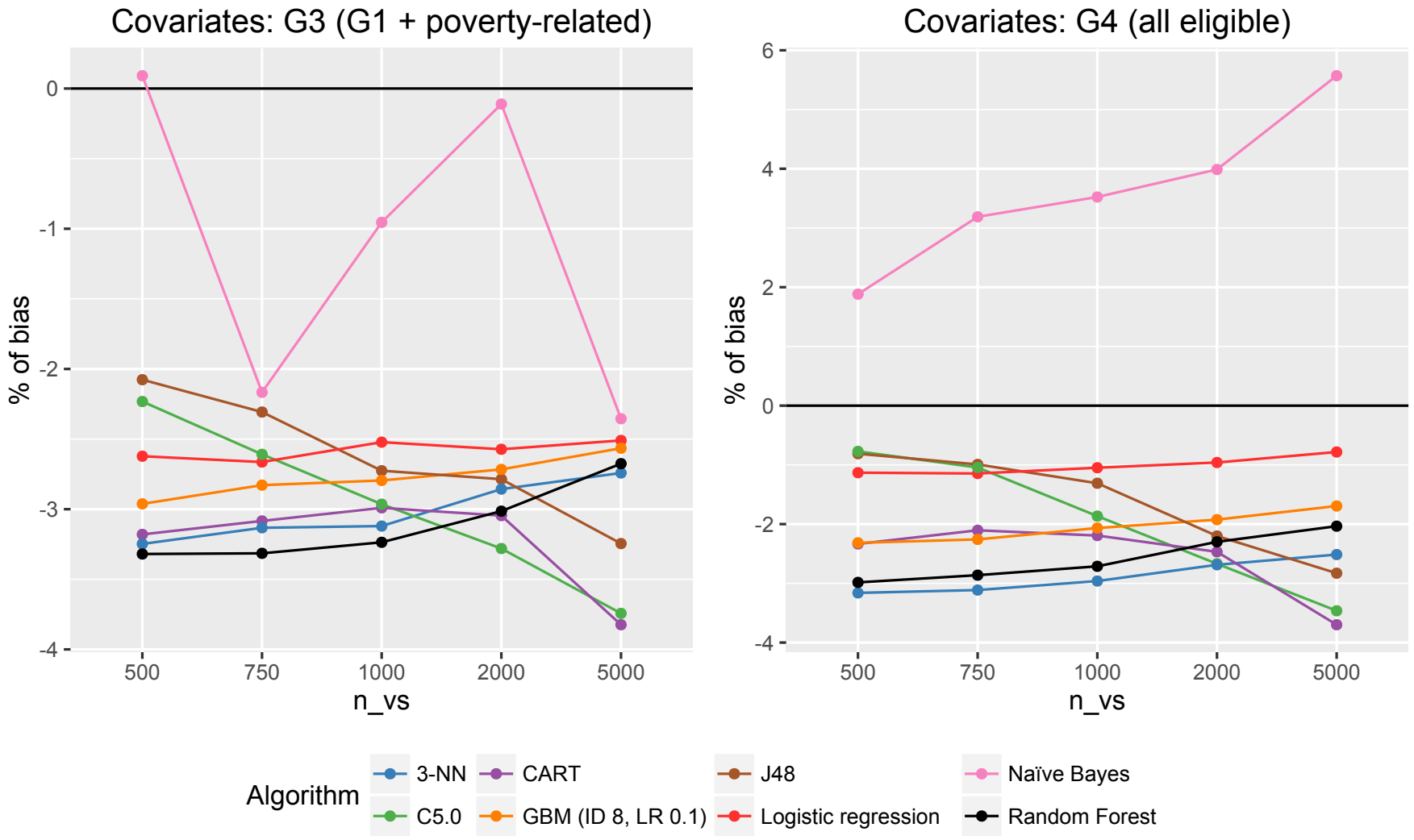
PSA in self reported health with real data



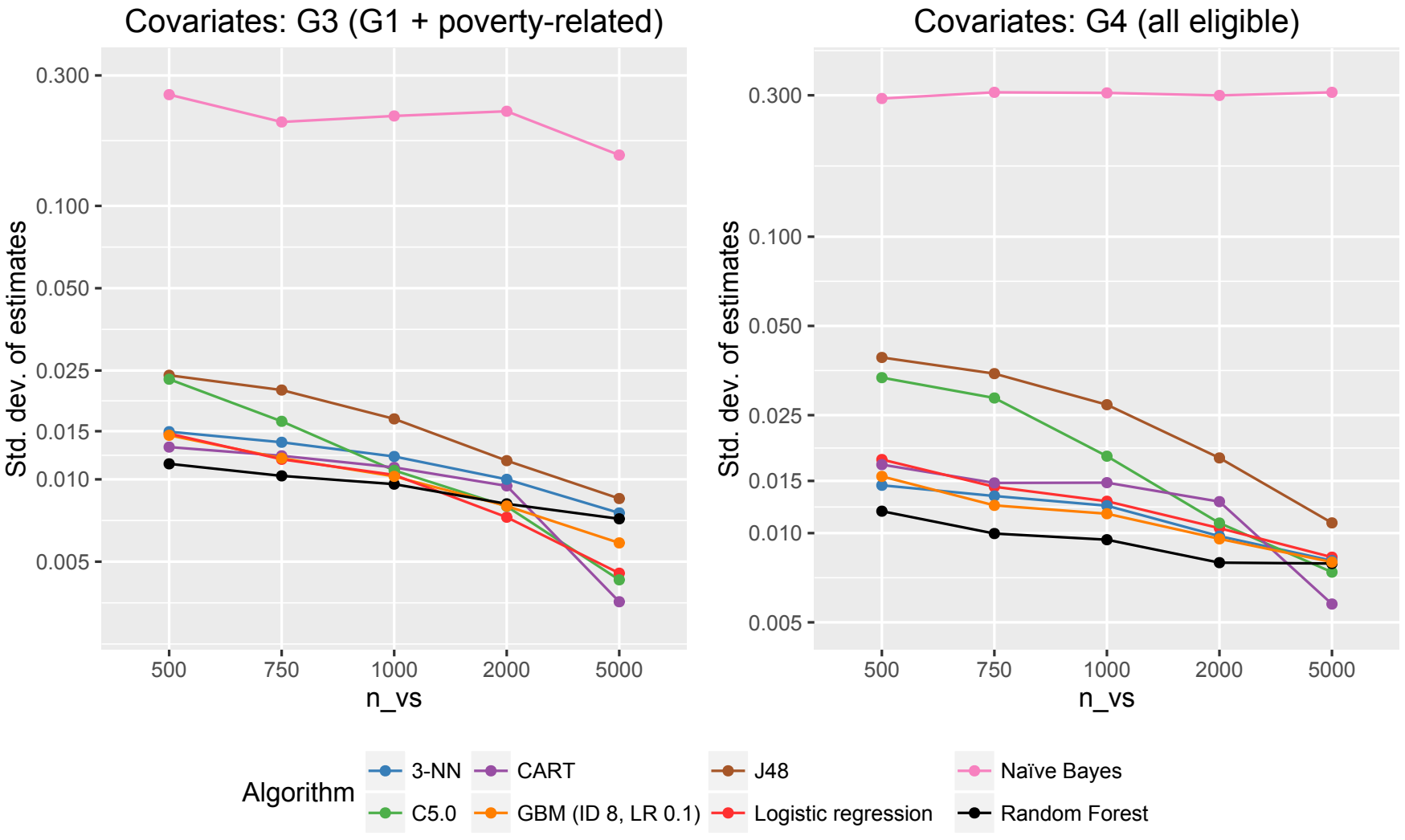
PSA in self reported health with real data



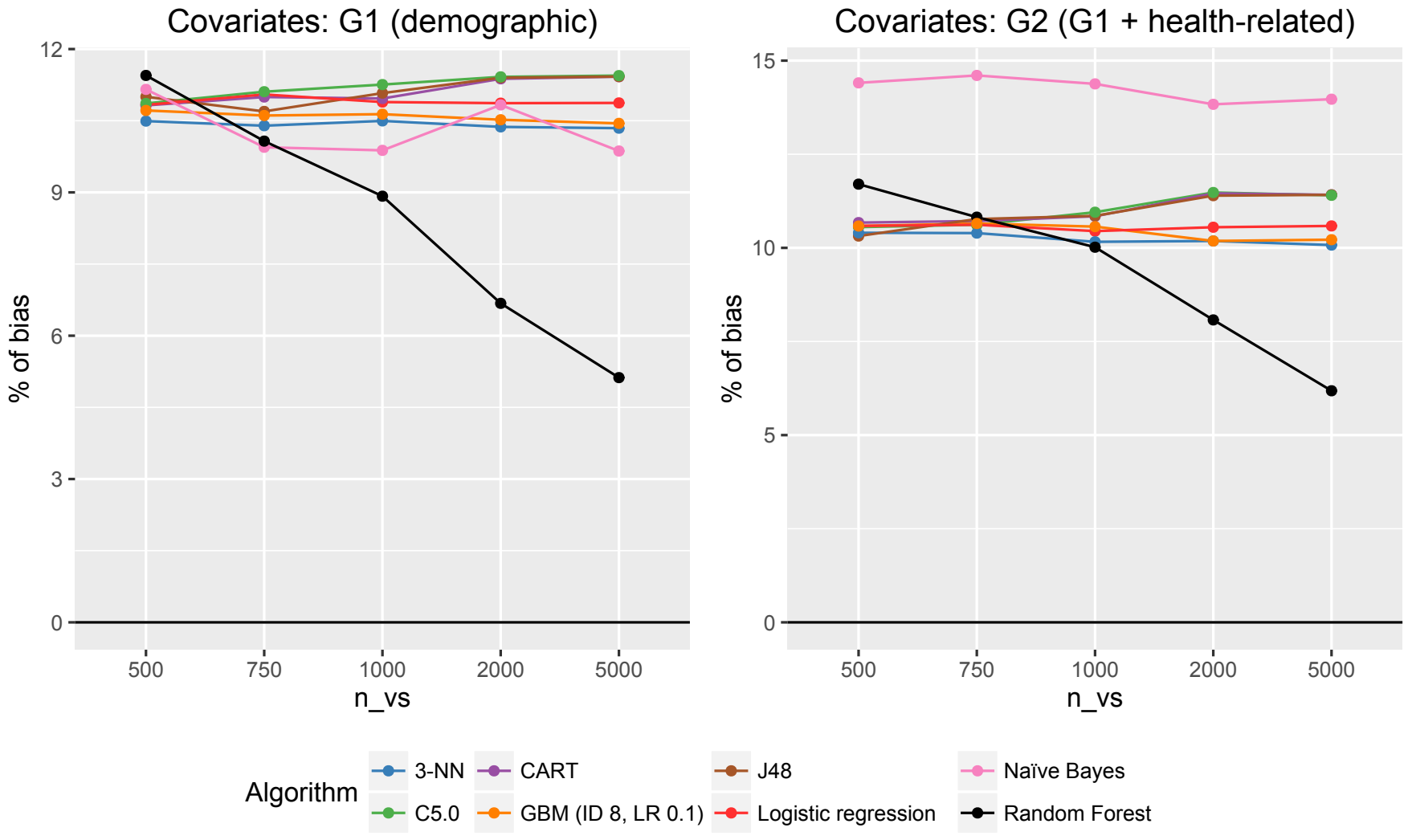
PSA in self reported health with real data



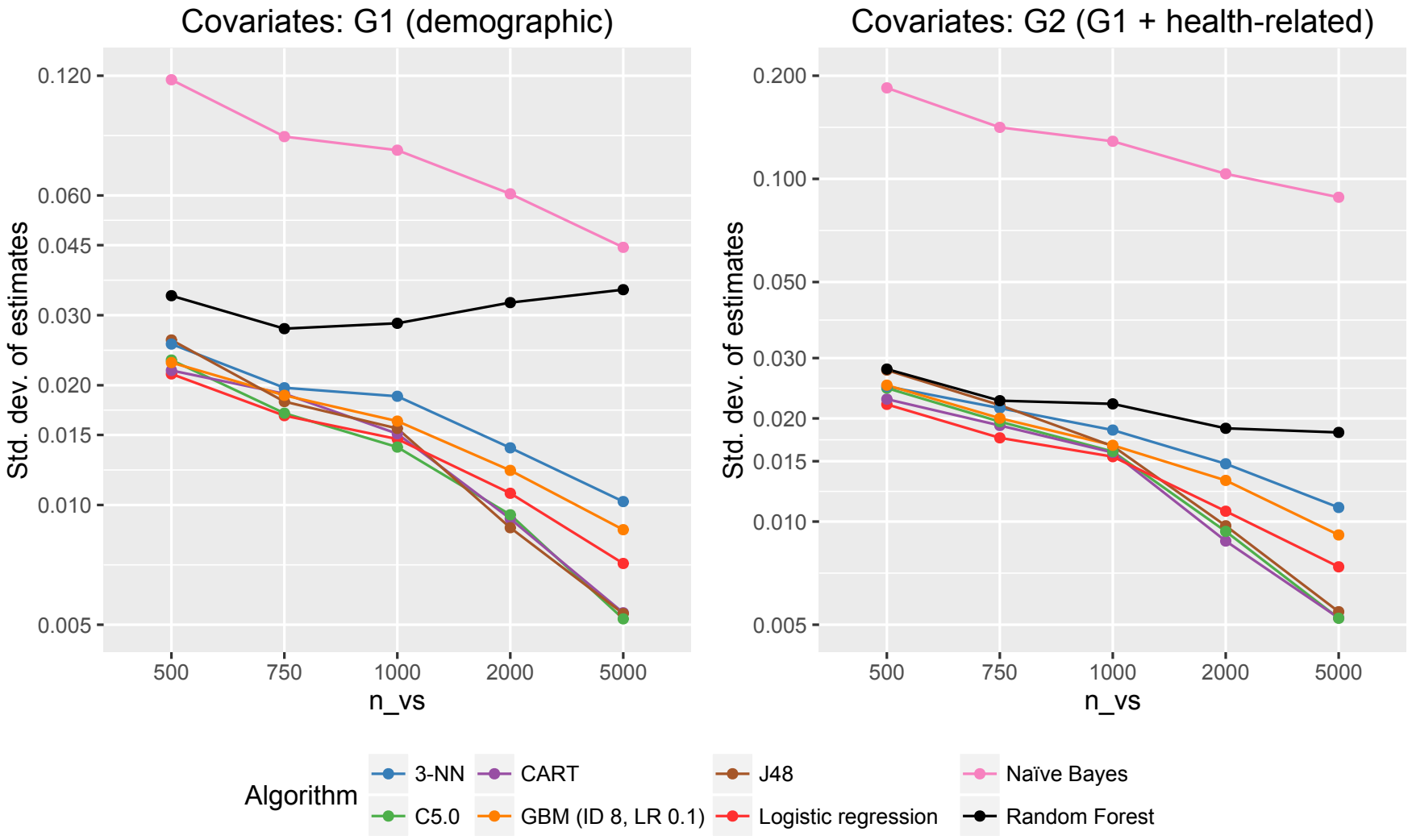
PSA in self reported health with real data



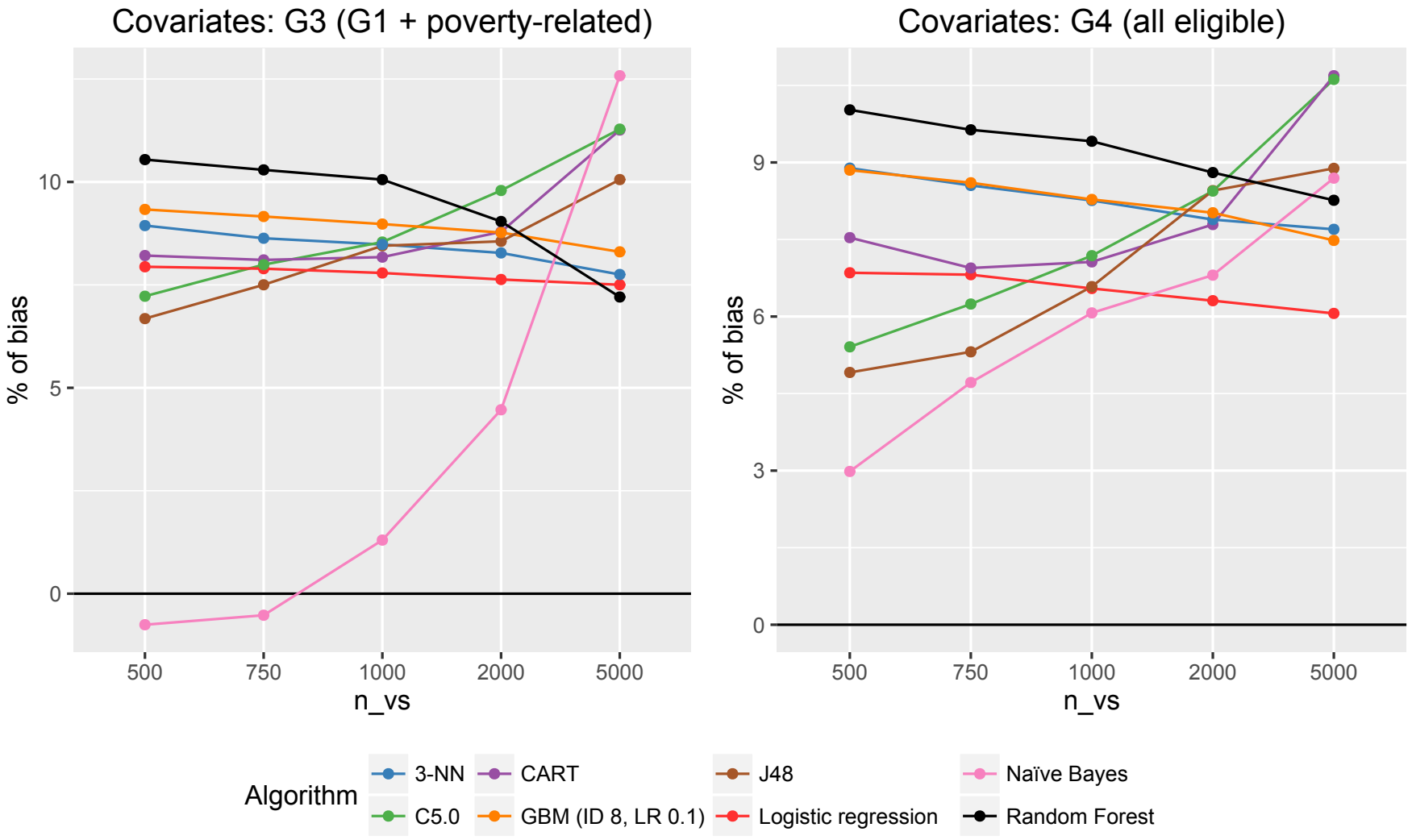
PSA in household members with real data



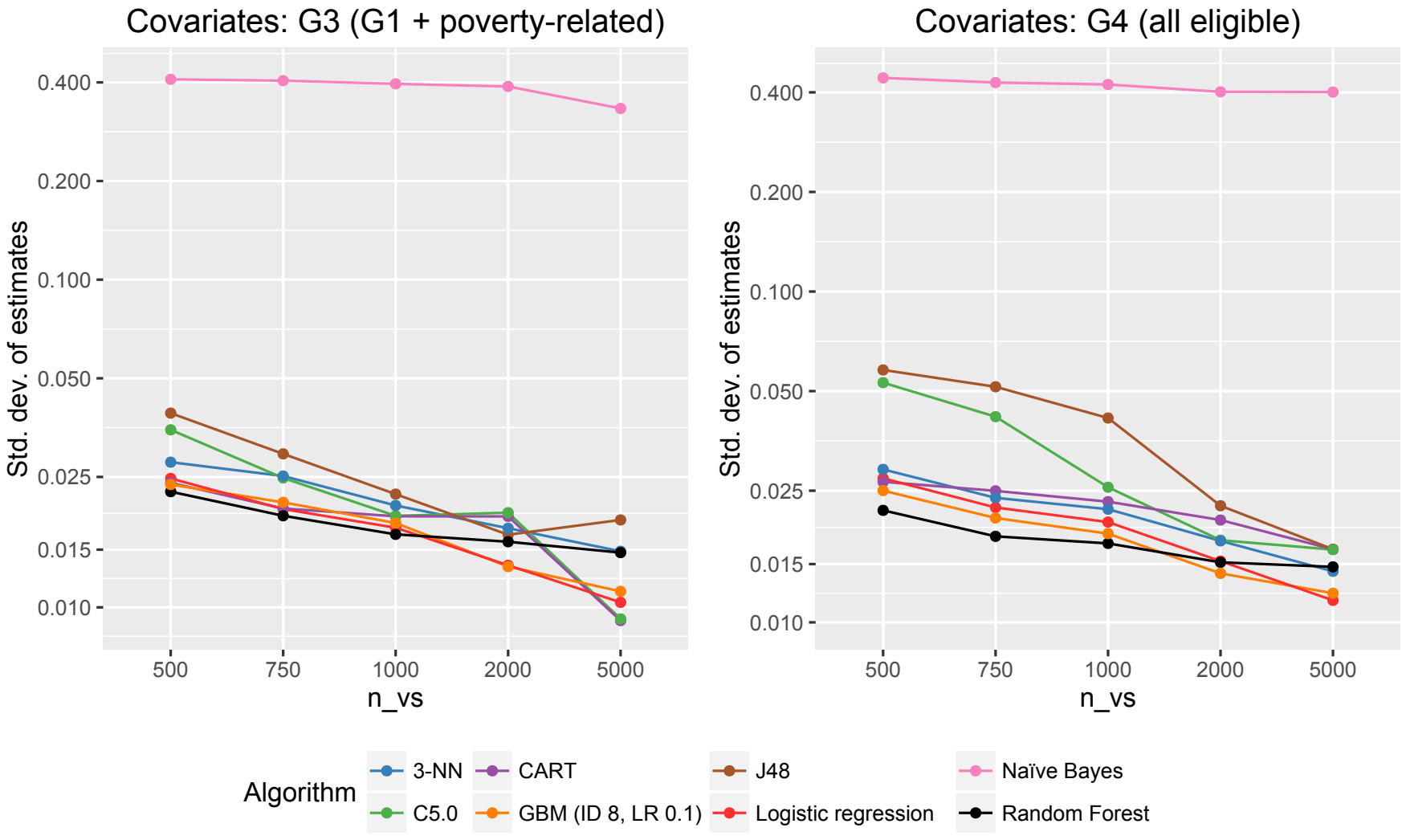
PSA in household members with real data



PSA in household members with real data



PSA in household members with real data



- Decision trees can be an alternative if both samples are balanced. The increase in estimators' variance caused by this approach has been observed in literature (Wyss et al., 2014).
- Nearest neighbors provide better results in low-dimensional contexts. This is a well-known attribute of k-NN classifiers (see Beyer et al., 1999).
- PSA with Naïve Bayes is unstable if covariates present rare classes (p. e. numeric variables) or the dimensionality of the dataset increases. Naïve Bayes is generally unable to deal with numeric variables and relies on independence assumptions which are hardly met as the number of variables increase (Barber, 2012; García et al., 2015).

- If controlling for overfitting, Random Forest is a solid alternative, specially if the target variable is influenced by Internet access. However, as in Pirracchio et al. (2014), an increase in variance could be attributed to its use.
- PSA with GBM presents very good properties, specially in terms of estimators' variance. These findings are in line with McCaffrey et al. (2013).
- **However, the choice on which algorithm should be used remains on the dataset and its attributes.**

- If non-volunteering is MCAR, CART and k-NN provide good results under any conditions, although most of the algorithms (except for Random Forests) perform well as n_{vs} increases.
- If non-volunteering is MAR or NMAR, best choices are k-NN and GBM if dimensionality is low. If assessing for right covariates, J48 and C5.0 are good alternatives if n_{vs} is small. Random Forests can cause a high impact in bias reducing but they tend to overfit.

- Data preprocessing and sample balancing are key steps when estimating efficient propensity scores with ML algorithms. This point has been discussed in literature (Linden and Yarnold, 2017).
- Parameter tuning, except for GBM, has not been proved to be a decisive factor in PSA success at removing bias. Up to the authors' knowledge, only Setoguchi et al. (2008) and Lee et al. (2010) addressed parameter tuning for PSA with ML.
- The limitations of the experiment on which parameter tuning was considered (artificial data, low dimensionality, etc.) must be taken into account in further research.

- Barber, D. (2012). Bayesian Reasoning and Machine Learning. Cambridge University Press, New York.
- Bethlehem, J. (2010). Selection Bias in Web Surveys. Int. Stat. Rev., 78(2), 161-188.
- Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U. (1999). When is “nearest neighbor” meaningful?. In International conference on database theory (pp. 217-235). Springer, Berlin, Heidelberg.
- Cavuto, S., Bravi, F., Grassi, M. C., Apolone, G. (2006). Propensity score for the analysis of observational data: an introduction and an illustrative example. Drug Dev. Res., 67(3), 208-216.
- D’Agostino, R. B. Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat. Med., 17, 2265-2281.
- Dever, J. A., Rafferty, A., Valliant, R. (2008). Internet Surveys: Can Statistical Adjustments Eliminate Coverage Bias?. Surv. Res. Methods, 2(2), 47-62.
- Ferri-García, R., Rueda, M. M. (2018). Efficiency of Propensity Score Adjustment and calibration on the estimation from non-probabilistic online surveys. SORT-Stat. Oper. Res. T. (accepted).
- García, S., Luengo, J., Herrera, F. (2015). Data Preprocessing in Data Mining. Switzerland: Springer International Publishing.

- Glynn, R. J., Schneeweiss, S., Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin. Pharmacol. Toxicol., 98(3), 253-259.
- Lee, S. (2006). Propensity Score Adjustment as a Weighting Scheme for Volunteer Web Panel Surveys. J. Off. Stat., 22(2), 329-349.
- Lee, B. K., Lessler, J., Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Stat. Med., 29(3), 337-346.
- Lee, S., Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociol. Method. Res., 37(3), 319-343.
- Linden, A., Yarnold, P. R. (2017). Using classification tree analysis to generate propensity score weights. J. Eval. Clin. Pract., 23(4), 703-712.
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat. Med., 32(19), 3388-3414.
- Pirracchio, R., Petersen, M. L., van der Laan, M. (2014). Improving propensity score estimators' robustness to model misspecification using super learner. Am. J. Epidemiol., 181(2), 108-119.
- Rosenbaum, P. R., Rubin, D. B. (1983). The Central Role of the PRopensity Score in Observational Studies for Causal Effects. Biometrika, 70(1), 41-55.

- Schonlau, M., Couper, M. (2017). Options for Conducting Web Surveys. Stat. Sci., 32(2), 279-292.
- Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol. Drug. Saf., 17(6), 546-555.
- Valliant, R., Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. Sociol. Method. Res., 40(1), 105-137.
- Watkins, S., Jonsson-Funk, M., Brookhart, M. A., Rosenberg, S. A., O'Shea, T. M., Daniels, J. (2013). An Empirical Comparison of Tree-Based Methods for Propensity Score Estimation. Health. Serv. Res., 48(5), 1798-1817.
- Westreich, D., Lessler, J., Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J. Clin. Epidemiol., 63, 826-833.
- Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., Stürmer, T. (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bCART, and the covariate-balancing propensity score. Am. J. Epidemiol., 180(6), 645-655.
- Zhao, P., Su, X., Ge, T., Fan, J. (2016). Propensity score and proximity matching using random forest. Contemp. Clin. Trials Commun., 47, 85-92.