# An ensemble learning method for variable selection

**Vincent Audigier**, Avner Bar-Hen

CNAM, MSDMA team, Paris

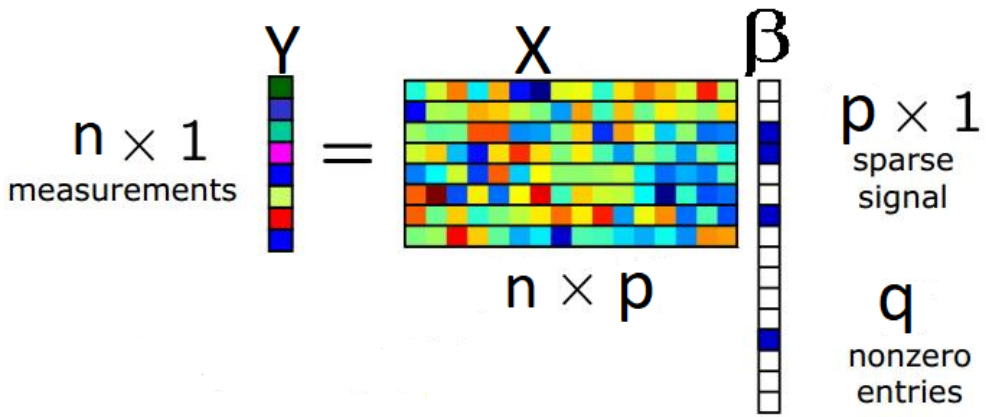
SBIM, November 2019

# Context

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon_{n\times 1} \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2\mathbb{I})$$

$\beta$ **sparse** with $q$ non-zeros

$X$ is **continuous**




**Goal: select the set of features $X_j$ that are related to $Y$**

# Stepwise

## Principle

1. start from the null model
2. at step $\ell$ ($\ell > 0$):
   - add to the model $\ell - 1$ the explanatory variable given the smallest RSS
   - for a given Type-I error level $\alpha$, remove from the model $\ell - 1$ a non-significant variable according to the student test (if any)
3. continue until convergence

## Properties

- no control of the Type-I error
- time consuming
- not tailored for high dimensional settings
- not tailored for missing values

# Lasso (Tibshirani, 1996)

Principle

- search the regression coefficients

$$argmin_{\beta} \underbrace{\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{RSS} + \underbrace{\lambda \sum_{j=1}^{p} |\beta_j|}_{\ell_1 \ penalty}$$

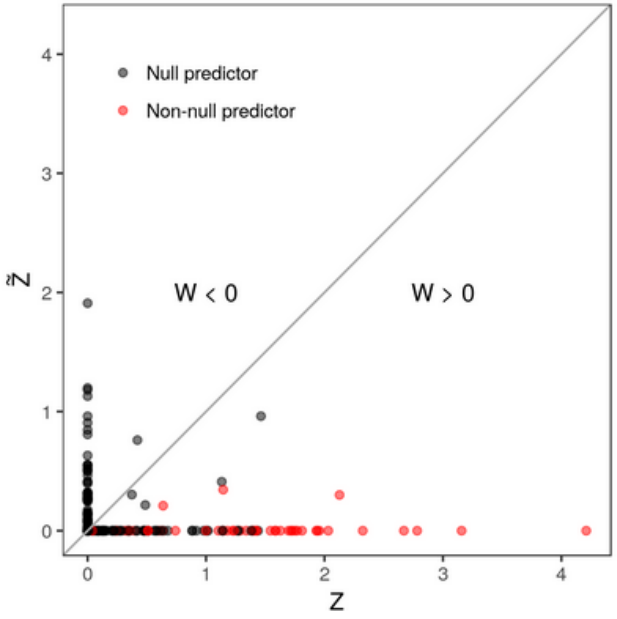
- the $\ell_1$ penalty forces some of the coefficient estimates to be exactly equal to 0 when $\lambda$ is sufficiently large (chosen by cv)

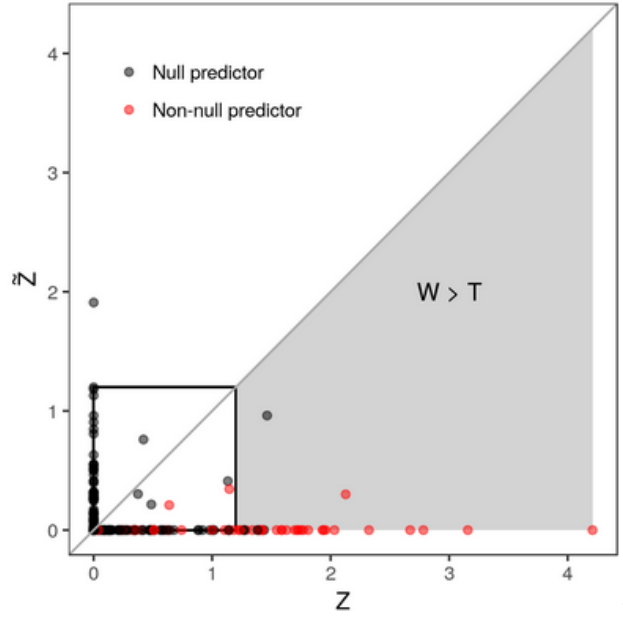
Properties

- no control of the Type-I error
- tailored for high-dimensional settings
- not tailored for missing values

# Knockoff (Barber and Candès, 2015)

1. Manufacture $p$ knockoff variables designed to **mimic the correlation structure** found within the $p$ original features
2. **Measure the feature importance** of knockoff ($\tilde{Z}$) and original variables ($Z$) according to a selection method (e.g. lasso)
3. **Compute the difference** between features importance of the original variable and its knockoff ($W = Z - \tilde{Z}$)
4. Among variables with positive difference, **filter variables** by keeping those with a measure of importance
$Z > T = min\{t : \frac{\#\{j:W_j \leq -t\}}{\#\{j:W_j > t\}} < \alpha\}$



# Properties

- control of the Type-I error
- tailored for high dimensional settings (Barber and Candes, 2016)
- not tailored for missing values

# Challenges

- **Stability** of the selected subset

  → Ensemble method (Genuer et al., 2010; Meinshausen and Bühlmann, 2010)

- **High dimensional data**

  → Shrinkage or preliminary sreening step (Tibshirani, 1996; Wasserman and Roeder, 2009; Barber and Candes, 2016)

- **Missing data**

  → Multiple imputation (Zhao and Long, 2017)

**Issues frequently occur simultaneously!**

# Outline

# Algorithm

**Create regression instances**

- Sample $k$ variables among the $p$ variables
- Handle missing values (e.g. by stochastic imputation)
- Apply a selection procedure (e.g. stepwise) to decide which of the $k$ variables are significantly related to $Y$.
- Iterate the process $B$ times ($\Rightarrow$ $B$ regression instances).

**Aggregate the regression instances**

- $r_j = \dfrac{\#\text{ times the variable } X_j \text{ is selected}}{\#\text{ time } X_j \text{ is present in the instances}}$
- Conclude that $X_j$ is significantly related to $Y$ if $r_j > 0.95$.

# Properties

- Improves stability repeating the selection procedure
- Overcomes the high dimensional setting by sampling a subset of variables
- Handles missing values by standard methods (by single stochastic imputation)
- The procedure inherits from the properties of the selection method used (consistency, error rates, ...)
- Several parameters to tune

  **k** number of sampled variables
  **B** number of regression instances

# Choice of $k$: complete data

**k**: number of sampled variables

- if the selection method does not handle high dimensional data, choose $k < n$
- $k$ large increases instability
- choose $k$ **not too small**...

With all explanatory variables

$$\mathbb{V}(Y \mid X) = \sigma^2$$

After a draw of $k$ variables $(X_{j_1}, \ldots, X_{j_k})$

$$\mathbb{V}(Y \mid X_{j_1}, \ldots, X_{j_k}) = \mathbb{V}(\beta_{j_{k+1}} X_{j_{k+1}} + \ldots + \beta_{j_p} X_{j_p}) + \sigma^2$$

... **since missing significant variables increase the noise** in the regression scheme

# Choice of $B$

$B$: number of iterations = number of regression instances

- We want all variables are selected, with high probability, in at least $\tilde{B}$ regression instances
- $Z_j$ number of regression instances that contains $X_j$.
- $Z_j \sim \mathcal{B}(B, k/p)$
- $\mathbb{P}(\min_{j=1,\dots,p} Z_j < \tilde{B}) < p \exp\left(-\left(1 - \frac{p\tilde{B}}{Bk}\right)^2 Bk/(2p)\right)$

$\Rightarrow$ $B$ and $p$ strongly related

# Outline

# Simulation design

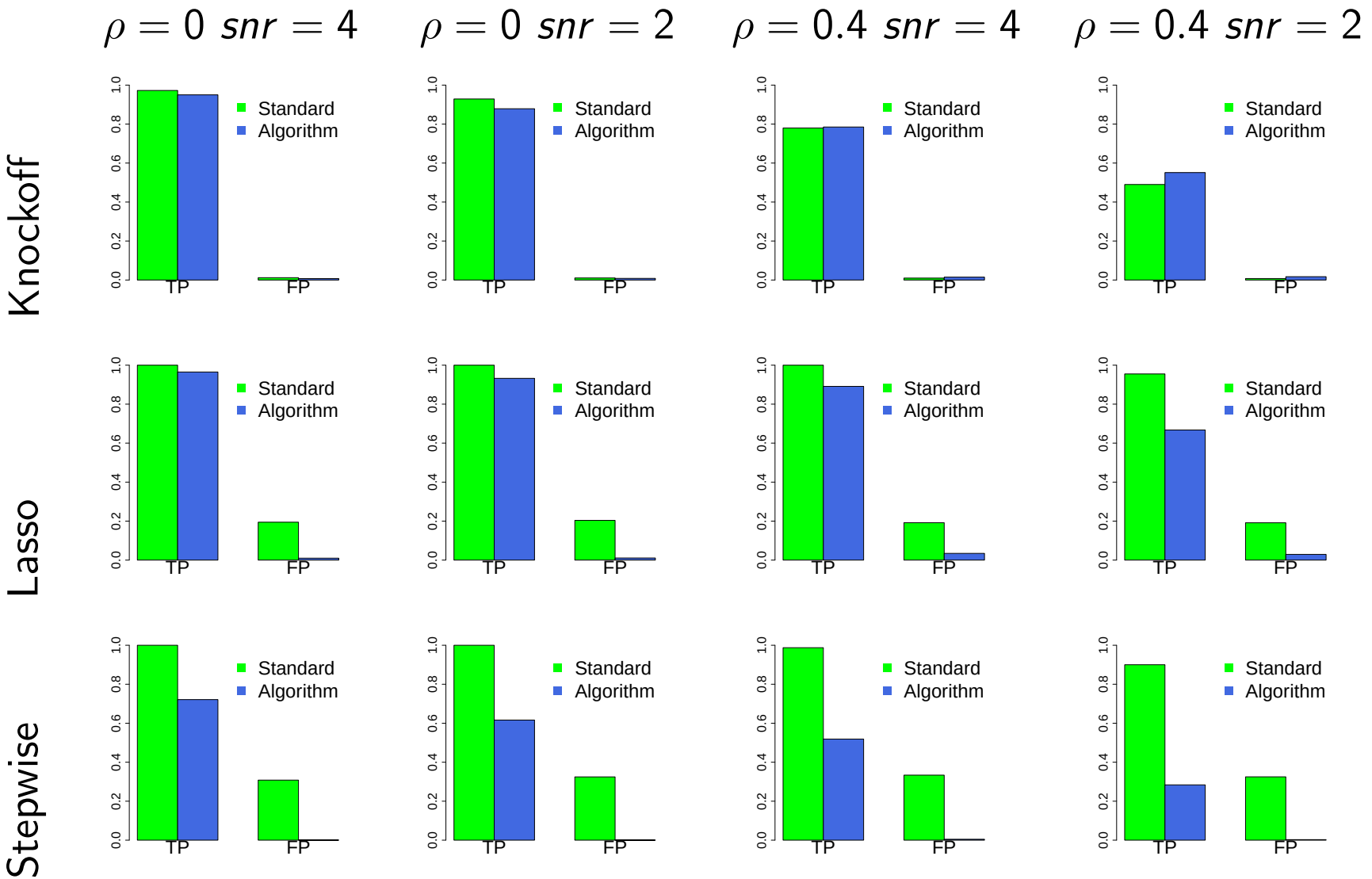
- $n = 200$ observations
- $p = 100$ or $p = 300$ variables
- Number of non-zeros: $q = 8$
- SNR = 2 ou 4 SNR: $\left(\frac{\text{Var}(Y)-\text{Var}(\varepsilon)}{\text{Var}(\varepsilon)}\right)$
- $\text{Cor}(X_i, X_j) = 0$ or $\text{Cor}(X_i, X_j) = 0.4$
- Missing values on covariates:
  - No
  - MCAR: $\mathcal{B}(0.2)$
  - MAR: Probit $\phi(\beta_0 + Y)$, with $\beta_0$ chosen so that 20% of values are missing (in expectation)

# Simulation design

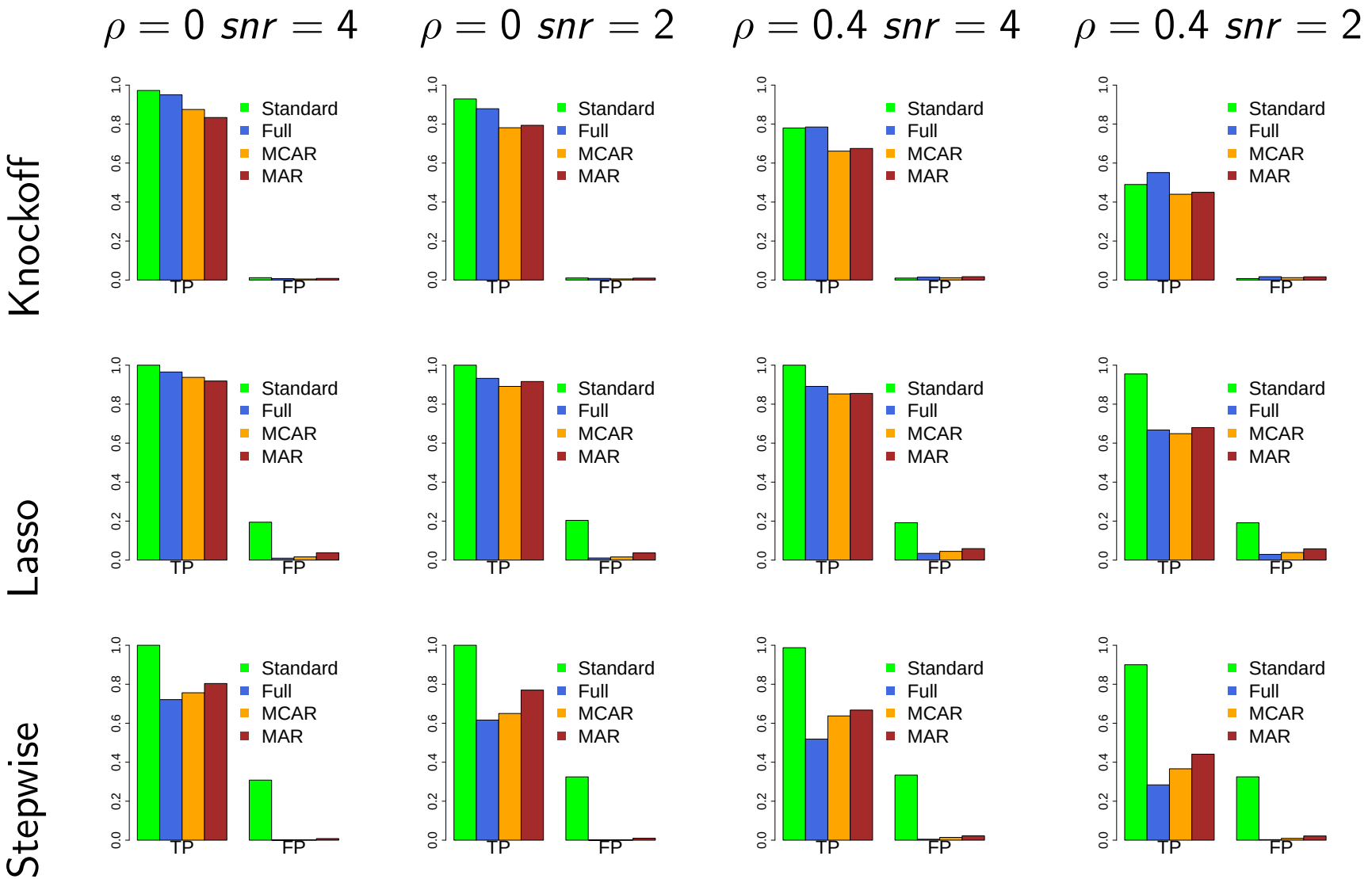
- For each of the 24 configurations, we generate 100 data sets
- Variable selection is performed according to
  - Knockoff
  - Lasso
  - Stepwise (with screening)
  - The algorithm (with Knockoff, Lasso or Stepwise)

  NB: only the algorithm is applied on incomplete data
- Performances are assessed using
  - TPR over the 100 generated data sets (1-FNR)
  - FPR over the 100 generated data sets (1-TNR)

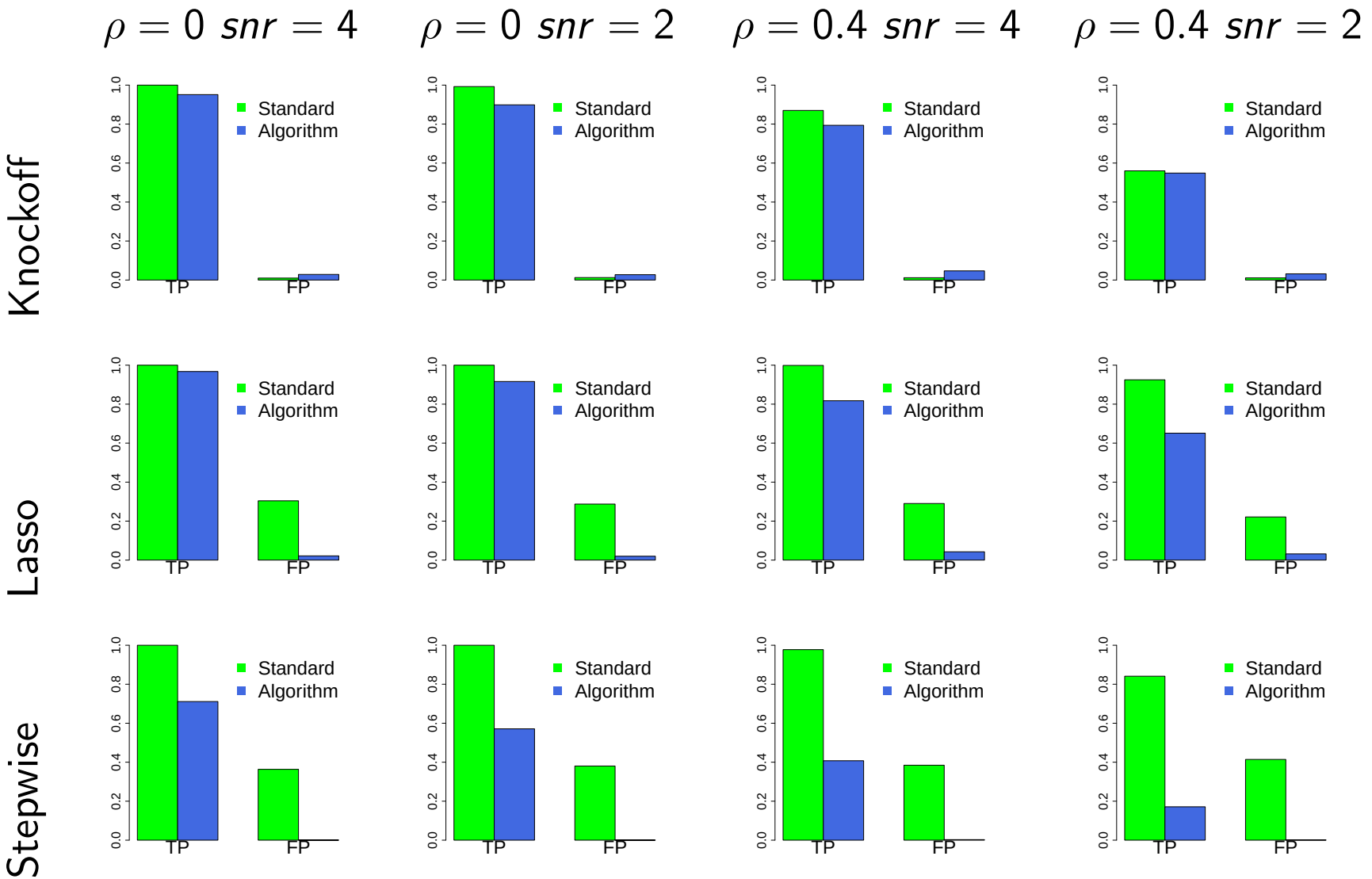
# Results: $n > p$ complete data



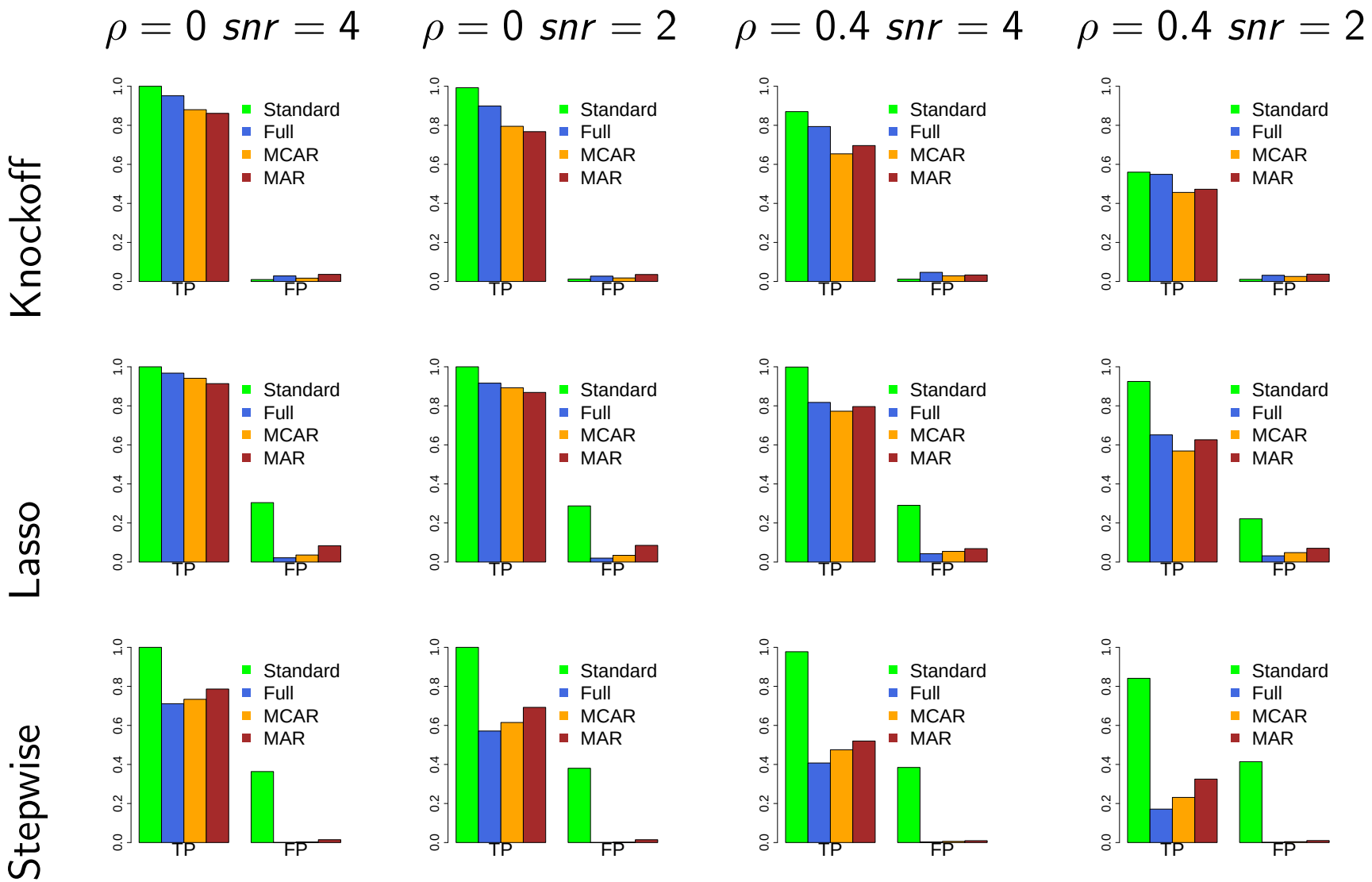
# Results: $n > p$ incomplete data



# Results: $n < p$ complete data



# Results: $n < p$ incomplete data



# Comparison to MIRL (Liu et al., 2016)

Input: A sample of n observations with p predictor variables with possible missing

↓

Step 1. Impute the sample m times and standardize all variables

↓

Step 2. For each imputed data set, generate B bootstrap samples
a) Compute lasso-OLS coefficients for each bootstrapped sample
b) Compute the importance measure for each of the p variables as the absolute value of the simple average of the $m \times B$ coefficients

↓

Step 3.
a) For each $m \times B$ dataset, randomly select $p/2$ variables with probability proportional to importance measures; compute lasso-OLS coefficients for $p/2$ variables and set the rest to be 0
b) Average $m \times B$ coefficients to get initial MIRL estimates

↓

Step 4.
Compute empirical selection probabilities and chose probability threshold by cross-validation, under which the MIRL estimates are set to be zero

# Comparison to MIRL (Liu et al., 2016)

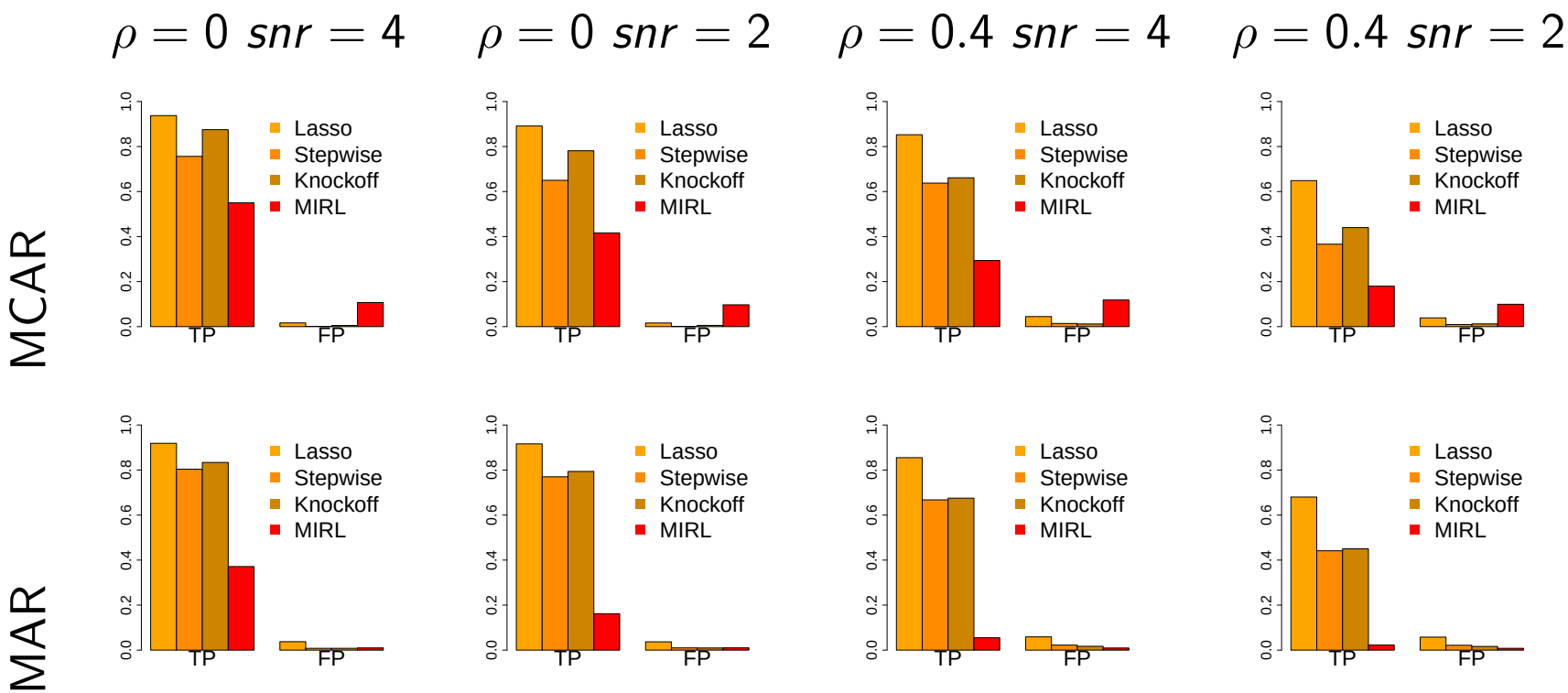



Table: Comparison between the proposed procedure (using Lasso, Stepwise or Knockoff) and MIRL: low dimensional setting with missing values

# Influence of $k$

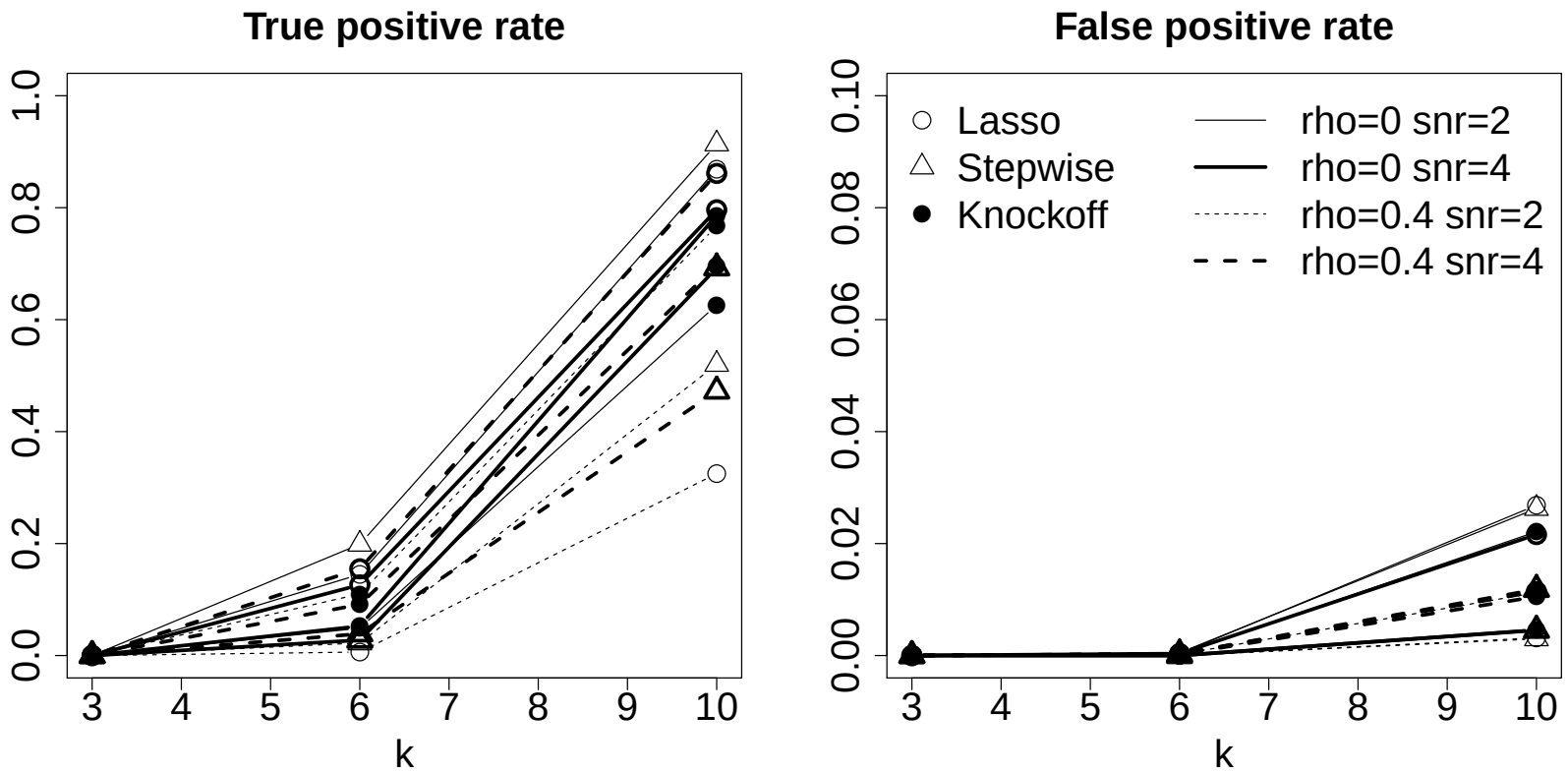



Figure: Influence of $k$: high dimensional setting with missing at random values according to $\rho$ and *snr* for 3 variable selection methods

# Influence of $B$

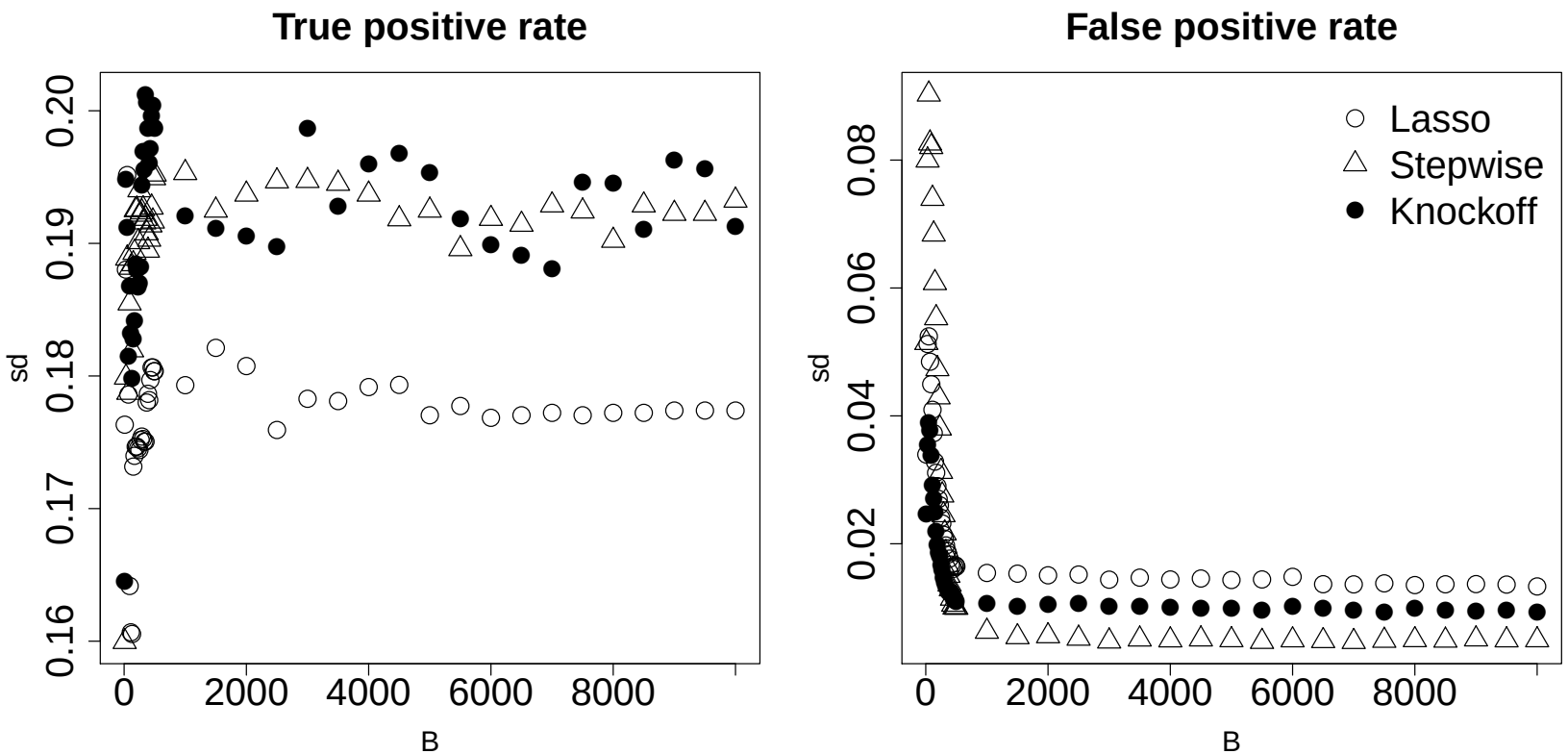



Figure: Influence of $B$: high dimensional setting with missing at random values. Standard deviation of the true positive rate (on the left) and false positive rate (on the right) over the 100 generated data sets for $\rho = 0$ and $snr = 4$

# Conclusions

A new variable selection procedure

- Very simple for applying any variable selection method in complex setting
- Relevant even in a low dimensional setting without missing data
- Providing a variable importance measure

Some limits

- Several tuning parameters
- MAR assumption less plausible with less variables

Perspective

- cross-validation for tuning
- R package

# References I


R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, 2015.

R.F. Barber and E. J. Candes. A knockoff filter for high-dimensional selective inference. *ArXiv e-prints*, 2016.

Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 31(14):2225–2236, 2010.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Larry Wasserman and Kathryn Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A): 2178, 2009.

Yize Zhao and Qi Long. Variable selection in the presence of missing data: imputation-based methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5):e1402–n/a, 2017. ISSN 1939-0068. doi: 10.1002/wics.1402. URL http://dx.doi.org/10.1002/wics.1402. e1402.

Ying Liu, Yuanjia Wang, Yang Feng, and Melanie M Wall. Variable selection and prediction with incomplete high-dimensional data. *The annals of applied statistics*, 10(1):418, 2016.

A. Bar-Hen and V. Audigier. An ensemble learning method for variable selection: application to high dimensional data and missing values. *ArXiv e-prints*, August 2018.