Supplementary file for the article "Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection"

Perrine Lacroix, Marie-Laure Martin

Electronic Journal of Statistics
ISSN: 1935-7524

Supplementary file for the article "Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection"

Perrine Lacroix

Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
Université Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
e-mail: perrine.lacroix@universite-paris-saclay.fr

and

Marie-Laure Martin

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120, Palaiseau, France
Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
Université Paris Cité, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
e-mail: marie-laure.martin@inrae.fr

Abstract: This supplementary file is an appendix of the article entitled "Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection" [3]. This file contains three sections. Section 1 contains theoretical and empirical justifications about the estimator choice of unknown parameters β^*, D_m^* and σ^2 involved in the theoretical FDR bounds in Theorem 3.2. It is a complementary work to Subsection 4.1. Section 2 contains plots as a complement to Section 1 for the bounds $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ for each of the four scenarios described in Table 4 of Section 7. Section 3 contains studies and plots evaluating the robustness of the model collections; studies and tables of results of the random model collection constructions and results of the variable selection procedure applications for scenarios (ii), (iii) and (iv) described in Table 4 of Section 7. This last section is a complementary work to Subsections 4.3 and 4.4 of [3].

Keywords and phrases: Ordered variable selection, Prediction, FDR, High-dimension, Gaussian regression, Hyperparameter calibration.

Contents

1. Estimation of the theoretical FDR

This section completes Subsection 4.1 of [3] and is devoted to the study of the theoretical upper bound terms of the FDR in Theorem 3.2 for a practical point of view.

The FDR bounds of Theorem 3.2 depend on the P_r, the $\underline{f}_r(K, \beta^*, \sigma^2)$ and the $\overline{f}_r(K, \beta^*, \sigma^2)$ quantities. Concerning the P_r quantities, they do not depend on the data as soon as r is given. They can be estimated once and for all without any dataset. For each $1 \leq r \leq q$, P_r is estimated by generating 5000 independent standard Gaussian vectors $\left(Z_k\right)_{k\in\{r+1,\cdots,q\}}$ and by counting for each vector the number of times that $Z_k^2 < K(\ell - r)$ for each $\ell \in \{r+1, \cdots, q\}$.
Concerning the $\underline{f}_r(K, \beta^*, \sigma^2)$ and $\overline{f}_r(K, \beta^*, \sigma^2)$ quantities, they depend on β^* and σ^2, both unknown.

We present the slope heuristic principle and an analyze of the $\hat{\sigma}^2$, obtained by the slope heuristics, is processed. Then, a large simulation study is performed to justify the choice of $\hat{\beta}_{\hat{m}(4)}$ to estimate β^* in the upper bound of the FDR. Both analyses are crucial since both estimators $\hat{\sigma}^2$ and $\hat{\beta}_{\hat{m}(4)}$ are proposed as inputs to the algorithm.

The slope heuristic to estimate σ^2. The slope heuristic principle, introduced in [2], is that when D_m is large enough, the empirical values of least squares $\frac{1}{n}||Y - X\hat{\beta}_m||_2^2$ are almost equal to $-\frac{1}{2n}K\sigma^2 D_m$ plus an additive constant independent of n and m. Hence, it is possible to estimate σ^2 from the dataset by the multiplicative coefficient of the affine behavior between the empirical values of least squares and $-\frac{K}{2n}D_m$ for D_m large enough. We use the function capushe of the R package capushe (version 1.1.1) [1] with parameters set to the default values.

Some substitutes of β^*. According to [2], $\hat{\beta}_{\hat{m}(K)}$ is a good estimator of β^* in a predictive point of view when K is equal or close to 2. We propose to test the estimators $\hat{\beta}_{\hat{m}(\tilde{K})}$ for $\tilde{K} \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, \log(n)\}$ to replace β^* in the lower and upper bounds $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$.
To determine the best constant $\tilde{K}$ among $\{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, \log(n)\}$, we evaluate all $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ on the sets $\mathcal{D}$ from the four scenarios described in Section 7 of [3]. To take into account the randomness of $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, the model collection generation

and model selection given by Equation (2.2) of [3] are processed on a new data set independent of $\mathcal{D}$ for the four scenarios.
To evaluate the error by replacing $b(K, \beta^*, \sigma^2)$ and $B(K, \beta^*, \sigma^2)$ with their estimation $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, we propose to evaluate the relative changes defined by : $\forall K > 0$,

$$\frac{b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2) - b(K, \beta^*, \sigma^2)}{b(K, \beta^*, \sigma^2)}$$

for the lower bound and by :

$$\frac{B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2) - B(K, \beta^*, \sigma^2)}{B(K, \beta^*, \sigma^2)}$$

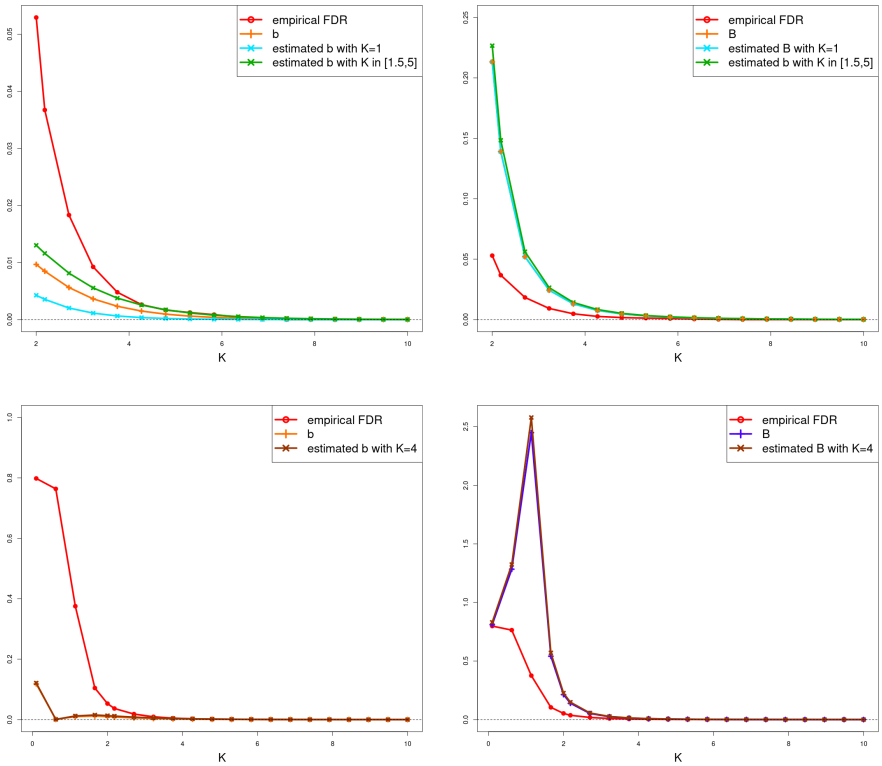
for the upper bound. To ensure that $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ values are larger than the $B(K, \beta^*, \sigma^2)$ values and so larger than the FDR ones, positive relative change values and as close to 0 as possible are expected. Concerning the lower bounds, negative relative change values are expected to ensure that $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ values are smaller than $B(K, \beta^*, \sigma^2)$ values and so smaller than the FDR ones. To take into account randomness of the $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ terms, we evaluate for all K the relative standard deviation, defined by the standard deviation divided by the mean, by calculated the variance of bounds $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ evaluated on 100 new data sets generated independently of $\mathcal{D}$. The relative standard deviation values are expected to be as close to 0 as possible.
Figures S-1-S-3 are plotted from the *toy data set*. In Figure S-1, the empirical estimation of the $\mathrm{FDR}\big(\hat{m}(K)\big)$ and the quantities $b(K, \beta^*, \sigma^2)$, $B(K, \beta^*, \sigma^2)$, $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are plotted on a grid of positive K. Relative changes and relative standard deviations for the lower bounds $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and upper bounds $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are plotted in Figure S-2 and S-3. The graphs of all others $\mathcal{D}$ of the 4 scenarios described in Table 4 of [3] are provided in Section 2.

The lower bounds : For $\tilde{K} > 1$, the relative change values are positive until achieving more than 2 for large K (Figure S-2 (top)) and the estimated lower bounds curves can be larger than the theoretical one. The relative standard deviation functions increase quickly whatever the value of $\tilde{K}$ suggesting that fluctuations around the mean are not negligible (Figure S-3 (top)).

The upper bounds : For $\tilde{K} > 1$, the relative change functions are always positive and do not exceed 0.11 meaning that the $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ curves are close to $B(K, \beta^*, \sigma^2)$ for all $K > 0$ (Figure S-2 (bottom)). For data sets $\mathcal{D}$ other than the *toy data set* (Figures S-4, S-7, S-10 and S-13), the relative change values are always small but can be negative. However, it happens very rarely for $\tilde{K} \geq 4$ and in this case, values are low enough (smaller than -0.02%) to ensure that the empirical FDR estimation curves do not exceed the $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$

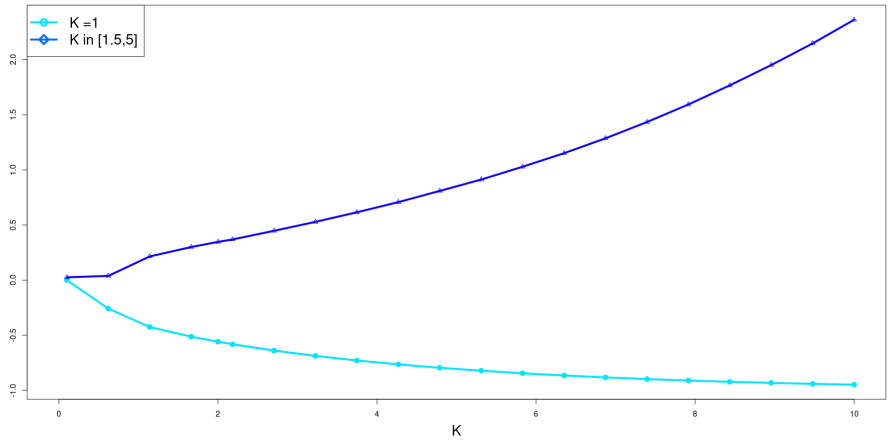
terms. Concerning the relative standard deviation functions (Figures S-3 (bottom), S-5, S-8, S-11 and S-14), the larger $\tilde{K}$, the smaller the values, except for the scenario (ii) with the third configuration where values increase after $\tilde{K} \geq 4.5$. For $\tilde{K} \geq 3.5$, the relative standard deviation values are around 0.2 for all the scenarios except for scenario (ii) with the second configuration (can achieve 0.8) and with the third configuration (can achieve 1). Thus, for a value of $\tilde{K} \in \{3.5, \log(n), 4, 4.5, 5\}$ and eventually except for the two extreme scenarios, fluctuations around the mean are small, meaning that the upper bound estimations are stable.

To conclude, we drop the lower bound to implement our data-driven algorithm for hyperparameter calibration since $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions can be larger than the theoretical FDR one. To control the FDR, only an upper bound control is sufficient. The best results for $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are obtained with the hyperparameter $\tilde{K} = 4$, where the relative change values are almost always positive, small enough to guarantee that the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ are larger than the theoretical FDR, and the relative standard deviation values are the smallest ones whatever the scenarios. So, the estimator used in our algorithm to replace β^* in the upper bound of the FDR is $\hat{\beta}_{\hat{m}(4)}$. A natural estimator of D_{m^*} is $D_{\hat{m}(4)}$. The value of the hyperparameter $\tilde{K} = 4$ is not surprising since the value of $D_{\hat{m}}$ has to be small enough in Equation (3.5) of [3] to get an upper bound $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ larger than the theoretical upper one. So, the penalization function has to be large enough in Equation (2.2) of [3].

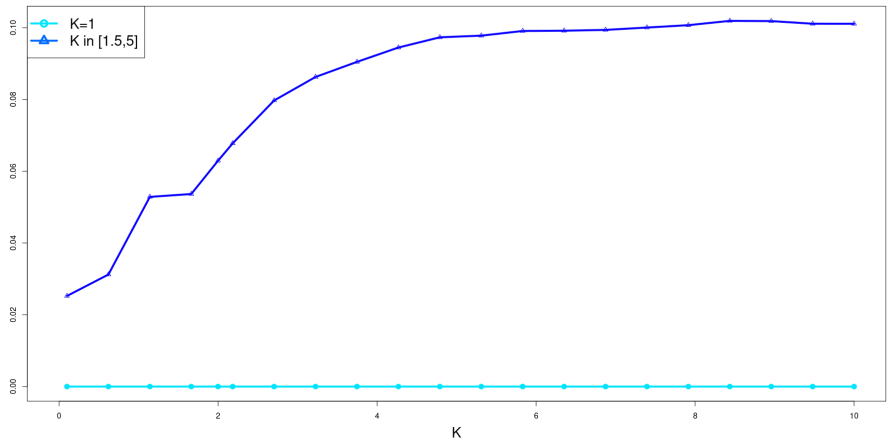


Lower bound Upper bound

Figure S-1: Top : Comparison of the empirical values of FDR, the functions $b(K, \beta^*, \sigma^2)$ (left) and $B(K, \beta^*, \sigma^2)$ (right) for the orthogonal design matrix X and the functions $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ (left) and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ (right) with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$. The terms $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ are calculating from a single data set, independent of those used for the empirical estimations; for a better readability, we plot curves only for $K \geq 2$. Bottom : Same comparison and estimation only with $\tilde{K} = 4$.

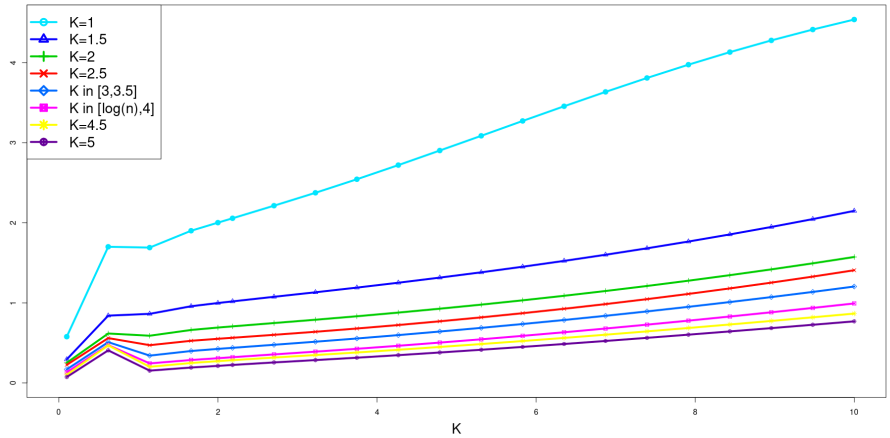


Lower bound

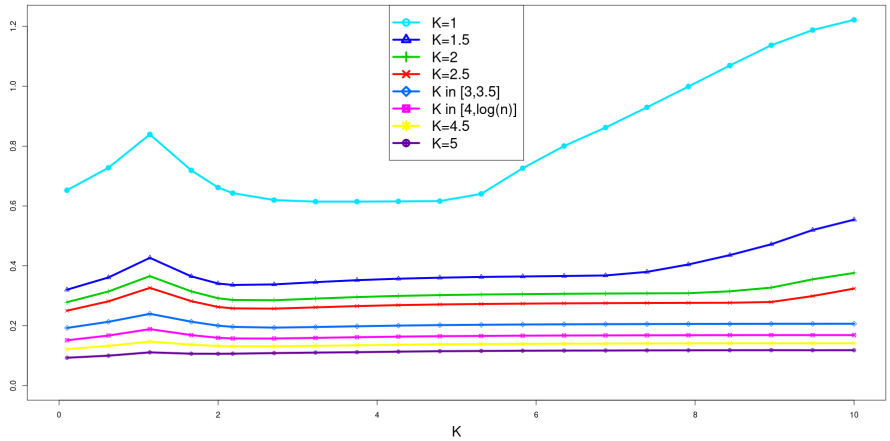


Upper bound

Figure S-2: Curves of the relative change values between the function $b(K, \beta^*, \sigma^2)$ (top) and $B(K, \beta^*, \sigma^2)$ (bottom) and the functions $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ (top) and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ (bottom) with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$, where estimators are calculated from a single data set.



Lower bound



Upper bound

Figure S-3: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $b(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ and $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K.

2. Complementary graphs for ordered variable selection

This section completes Subsection 4.1 of [3].

In this section, graphs for scenarios (ii) to (iv) described in Table 4 of [3] are provided. Relative changes and relative standard deviations for the $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ bounds when $\tilde{K} \in \{1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, \log(n)\}$ are plotted in Figures S-4 and S-5 for scenario (i), in Figures S-7 and S-8 for scenario (ii), in Figures S-10 and S-11 for scenario (iii) and in Figures S-13 and S-14 for scenarios (iv). The empirical difference in predictions and the empirical $\mathrm{FDR}\big(\hat{m}(K)\big)$ functions, the estimated difference in predictions (Equation 4.2 of [3]) and the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions for a grid of values of $K > 0$ are plotted on Figure S-6 for scenario (i), Figure S-9 for scenario (ii), Figure S-12 for scenario (iii) and Figure S-15 for scenario (iv).

When we focus on the scenario (i), the higher the D_{m^*} value is, the smaller the empirical FDR is but the larger the empirical PR for large K is. Moreover, the relative change functions decreases when D^*_m increases, as well as the relative standard deviation ones which remain smaller than 0.5. This can be explained since the higher D_{m^*}, the smaller the number of non active variables, so the smaller the number of the selected non active variables and the smaller FDR value. In the opposite trend, the empirical PR increases with D_{m^*} since penalization tends to select too few variables, especially even K moves away from 2.

As expected, concerning the scenario (ii), when coefficients are smaller than the amplitude of the noise (the second configuration), values of the relative change for the $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ bounds explode (until 10^5) and the relative standard deviation values increase until exceed 1. The best results are obtained for the first β^* configuration of the scenario (ii), but results still remain reasonable with the third one. The relative standard deviation functions increase after $\tilde{K} \geq 4$ whereas in all other scenarios, functions always decrease when $\tilde{K}$ increases. When permutations of the ten first variables are processed from the nested model collection (Equation 2.1 of [3]), $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ values begin to diverge from those of $B(K, \beta^*, \sigma^2)$ for the second and the third configurations where the coefficients of β^* are close to each other. Unlike the other scenarios, $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ and $B(K, \beta^*, \sigma^2)$ values are both larger than the empirical FDR ones for the second configuration.

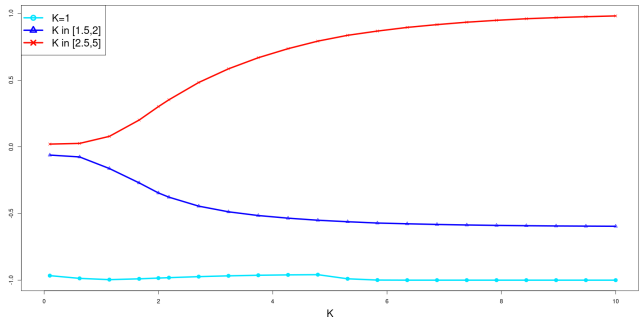
For configuration (iii) and when permutations of the first twelve and fifteen variables are processed, FDR values are the highest all along the collections compared to all other scenarios (similar to scenario (iv) and $\sigma^2 = 4$) and so, distinction between active and non active variables is more difficult. Proportions of active variables in models of size 5, 10, 15 and 20 fall to 0.6 with the third configuration and 0.8 with the second configuration for which the discrimination between active and non active variables is naturally less obvious.

As for the scenario (iii), we observe unsurprisingly that the higher the value of n, the smaller the relative change, the smaller the relative standard deviation, and the tighter the confidence interval of PR (< 0.04 for $n = 30$). However,

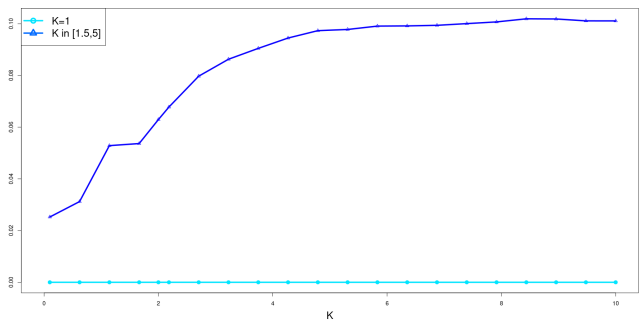
we note that the computational time to estimate the bounds was significantly higher for $n = 300$.

Lastly, concerning the scenario (iv), as expected, the higher the noise amplitude, the larger the confidence interval for the PR (< 0.45 for $\sigma^2 = 4$), the higher the relative change (which equals 0 when $\sigma^2 = 0.1$ and around 2 when $\sigma^2 = 4$) and the higher the relative standard deviation. However, values remain reasonable even for $\sigma^2 = 4$ excepted for the empirical PR values which are always larger than 5. For $\sigma^2 = 4$ and when permutations of the first twelve and fifteen variables are processed, FDR values are the highest all along the collections compared to all other scenarios (similar to scenario (ii) configuration (iii)) and so, distinction between active and non active variables is more difficult. Unlike the other scenarios, the Bolasso provides the highest values of proportion of active variables in models of size 5, 10, 15 and 20 when $\sigma^2 = 0.1$. Proportions fall to 0.8 when $\sigma^2 = 4$.

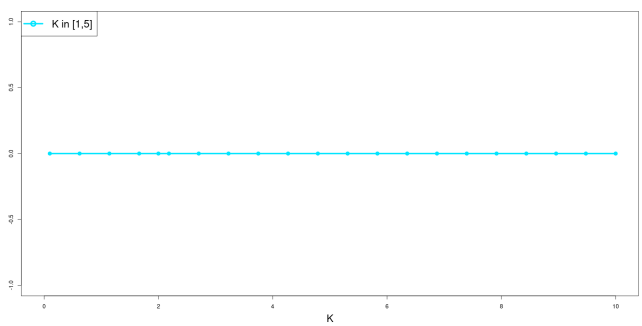
2.1. Scenario (i)



$|\beta^*| = 1$

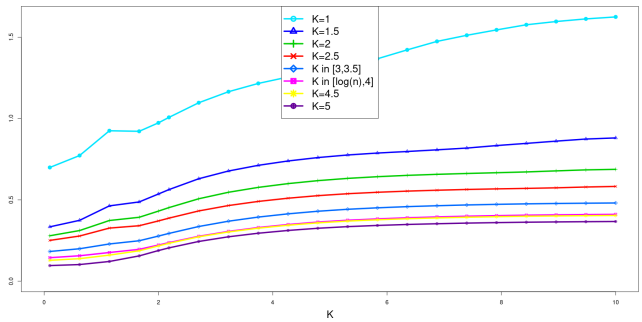


$|\beta^*| = 10$

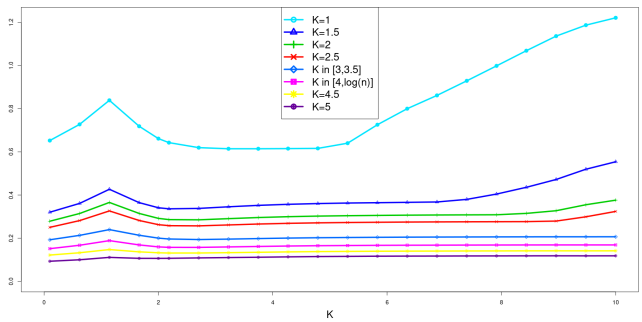


$|\beta^*| = 20$

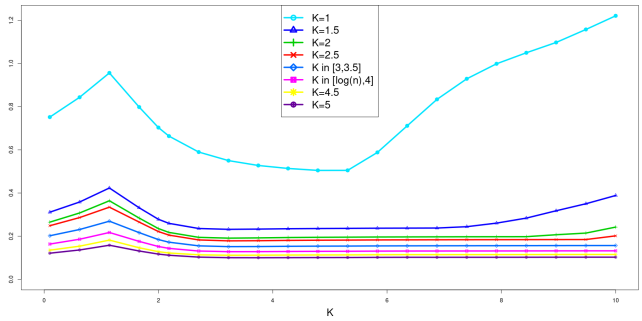
Figure S-4: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from a single data set. Top: for $|\beta^*| = 1$. Middle: for $|\beta^*| = 10$. Bottom: for $|\beta^*| = 20$.



$|\beta^*| = 1$

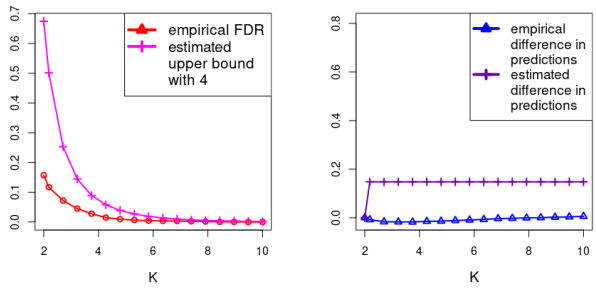


$|\beta^*| = 10$



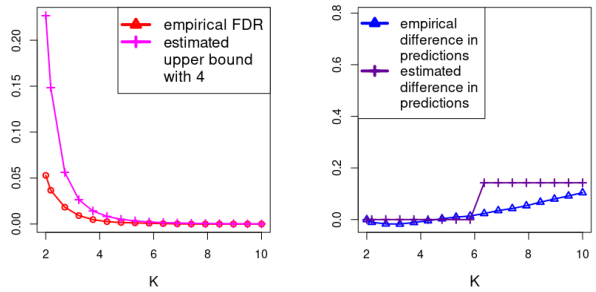
$|\beta^*| = 20$

Figure S-5: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. Top: for $|\beta^*| = 1$. Middle: for $|\beta^*| = 10$. Bottom: for $|\beta^*| = 20$.



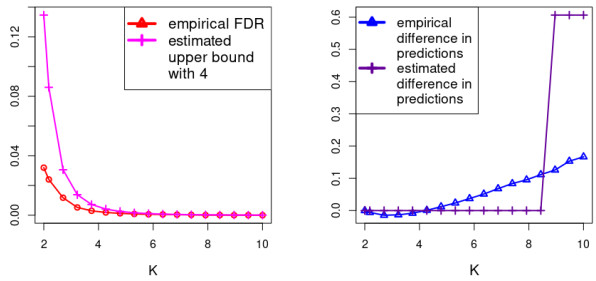
FDR diff-PR

$|\beta^*| = 1$



FDR diff-PR

$|\beta^*| = 10$

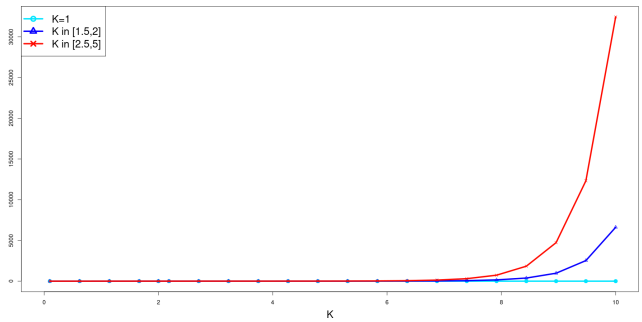


FDR diff-PR

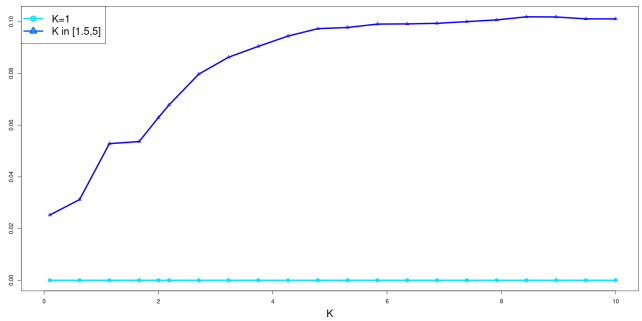
$|\beta^*| = 20$

Figure S-6: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and of $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for $K \geq 2$ for the *toy data set*. Top: for $D_m = 1$. Middle: for $D_m = 10$. Bottom: for $D_m = 20$.

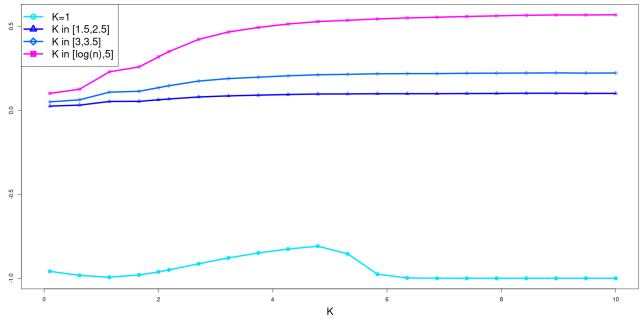
2.2. Scenario (ii)



$\beta^*_{10} = \frac{2}{10}$

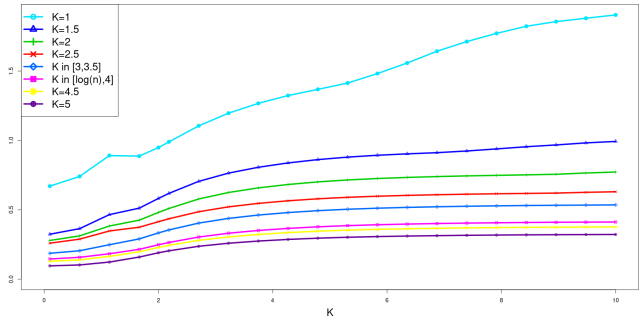


$\beta^*_{10} = 2$ and distant coefficients.

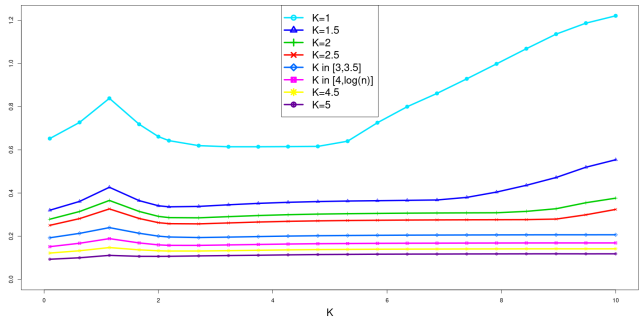


$\beta^*_{10} = 2$ and close coefficients.

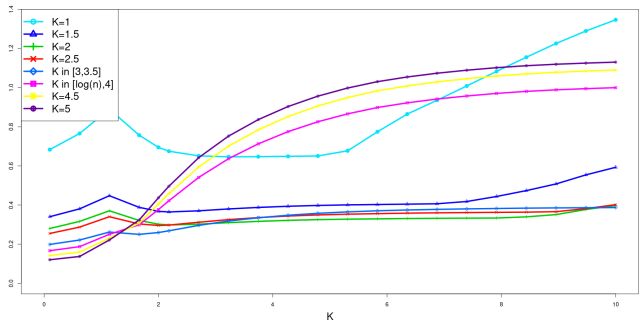
Figure S-7: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from a single data set. Top: for $\beta^*_{10} = \frac{2}{10}$. Middle: for $\beta^*_{10} = 2$ and distant coefficients. Bottom: for $\beta^*_{10} = 2$ and close coefficients.



$\beta^*_{10} = \frac{2}{10}$

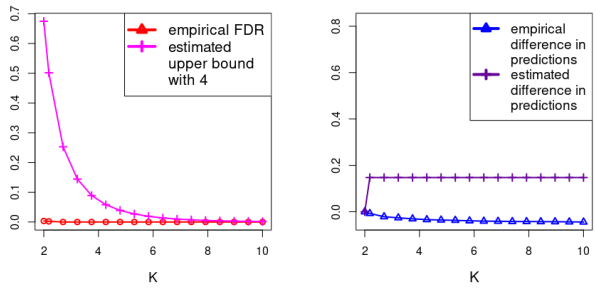


$\beta^*_{10} = 2$ and distant coefficients.



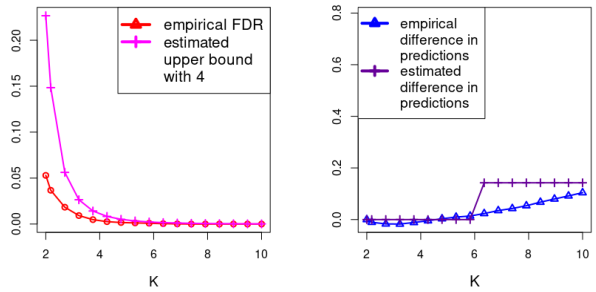
$\beta^*_{10} = 2$ and close coefficients.

Figure S-8: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. Top: for $\beta^*_{10} = \frac{2}{10}$. Middle: for $\beta^*_{10} = 2$ and distant coefficients. Bottom: for $\beta^*_{10} = 2$ and close coefficients.



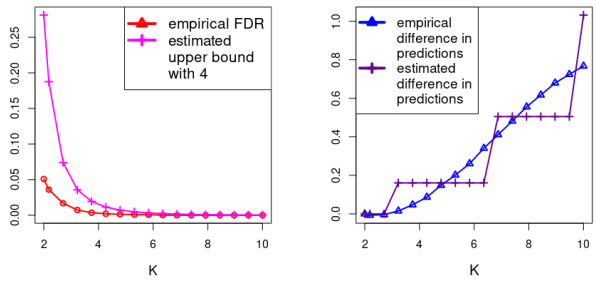
FDR diff-PR

$\beta^*_{10} = \frac{2}{10}$



FDR diff-PR

$\beta^*_{10} = 2$ and distant coefficients

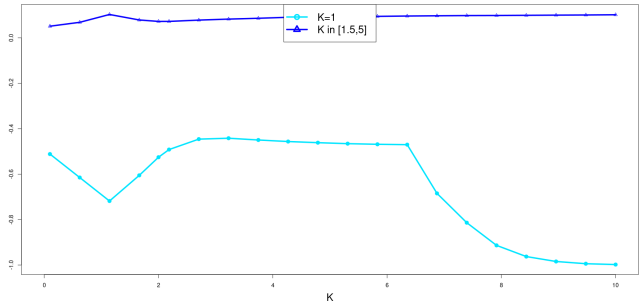


FDR diff-PR

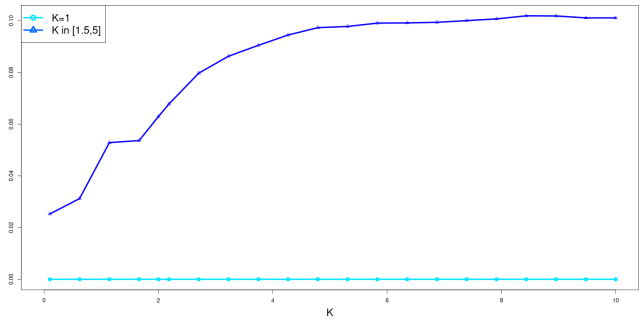
$\beta^*_{10} = 2$ and close coefficients

Figure S-9: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and of $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for $K \geq 2$ for the *toy data set*. Top: for $\beta^*_{10} = \frac{2}{10}$. Middle: for $\beta^*_{10} = 2$ and distant coefficients. Bottom: for $\beta^*_{10} = 2$ and close coefficients.

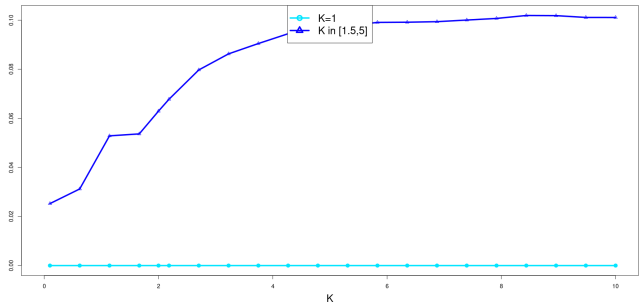
2.3. Scenario (iii)



$n = 30$

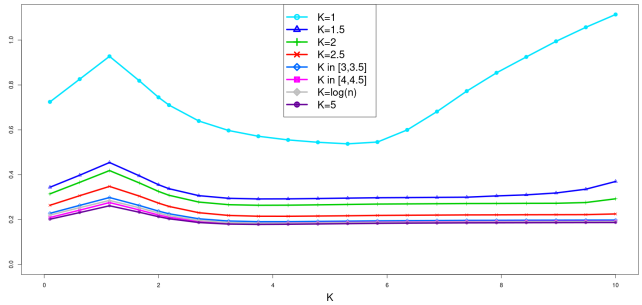


$n = 50$

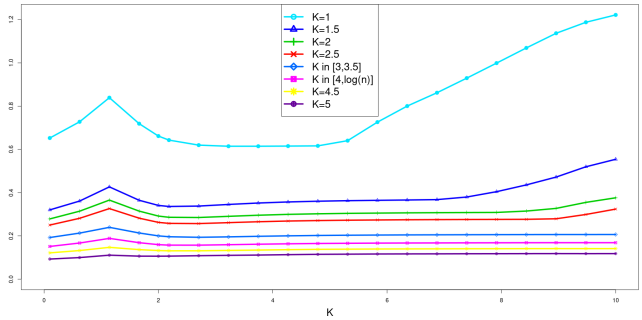


$n = 300$

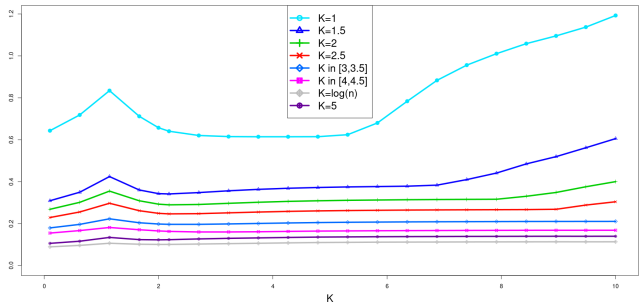
Figure S-10: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from a single data set. Top: for $n = 30$. Middle: for $n = 50$. Bottom: for $n = 300$.



$n = 30$

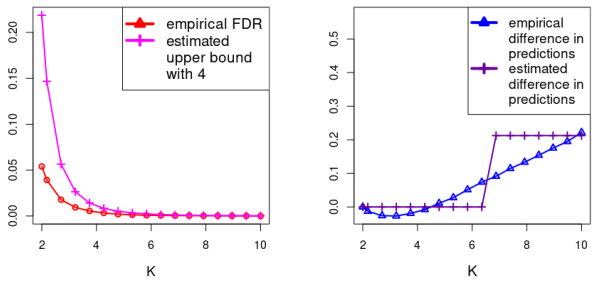


$n = 50$



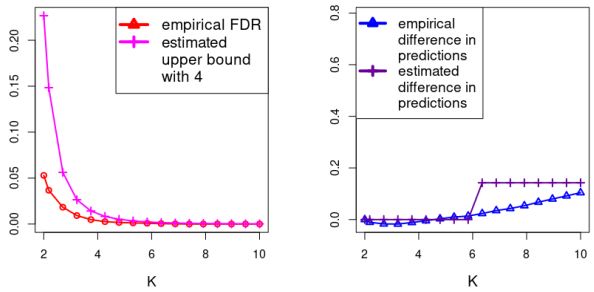
$n = 300$

Figure S-11: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. Top: for $n = 30$. Middle: for $n = 50$. Bottom: for $n = 300$.



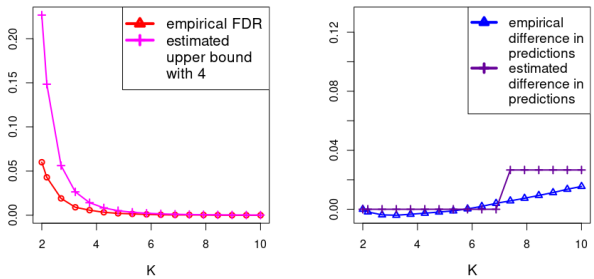
FDR diff-PR

$n = 30$



FDR diff-PR

$n = 50$

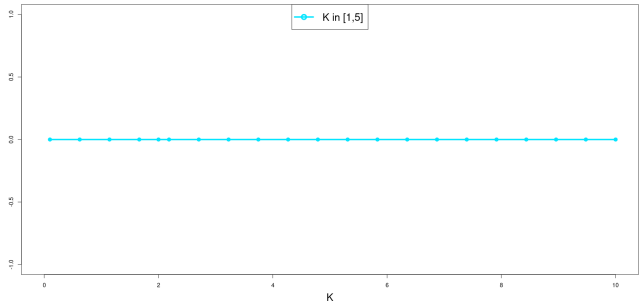


FDR diff-PR

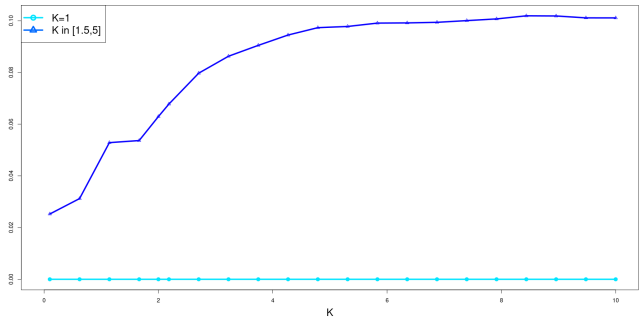
$n = 300$

Figure S-12: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and of $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for $K \geq 2$ for the *toy data set*. Top: for $n = 30$. Middle: for $n = 50$. Bottom: for $n = 300$.

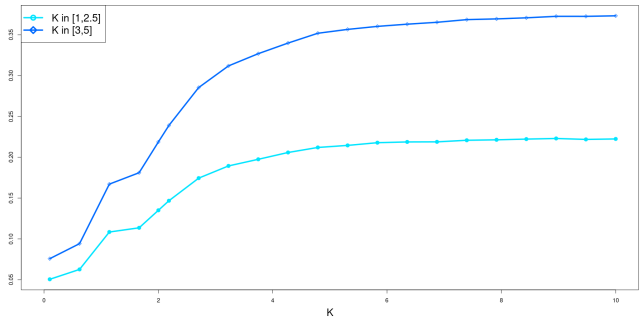
2.4. Scenario (iv)



$\sigma^2 = 0.1$

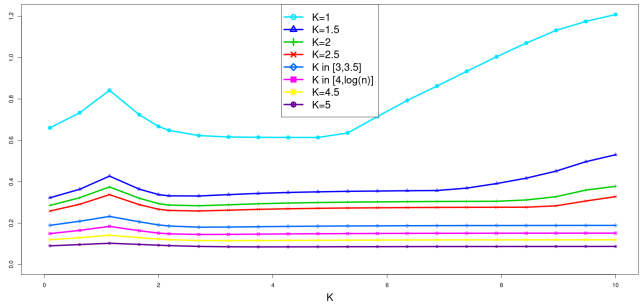


$\sigma^2 = 1$

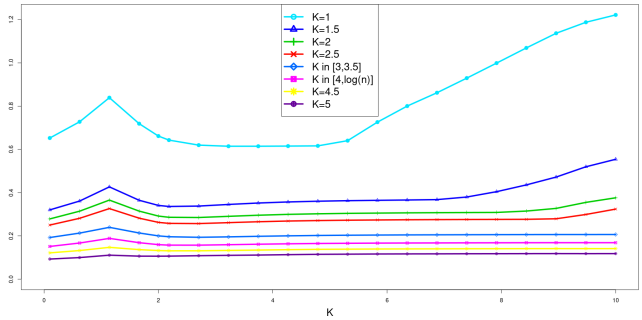


$\sigma^2 = 4$

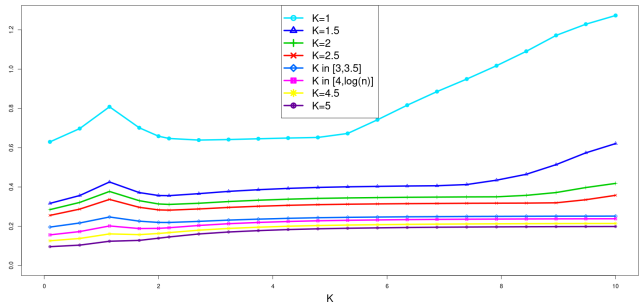
Figure S-13: Curves of the relative change values between the functions $B(K, \beta^*, \sigma^2)$ and the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ with respectively $\hat{\beta}_{\hat{m}(1)}$, $\hat{\beta}_{\hat{m}(1.5)}$, $\hat{\beta}_{\hat{m}(2)}$, $\hat{\beta}_{\hat{m}(2.5)}$, $\hat{\beta}_{\hat{m}(3)}$, $\hat{\beta}_{\hat{m}(3.5)}$, $\hat{\beta}_{\hat{m}(4)}$, $\hat{\beta}_{\hat{m}(4.5)}$, $\hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ where estimators are calculating from a single data set. Top: for $\sigma^2 = 0.1$. Middle: for $\sigma^2 = 1$. Bottom: for $\sigma^2 = 4$.



$\sigma^2 = 0.1$

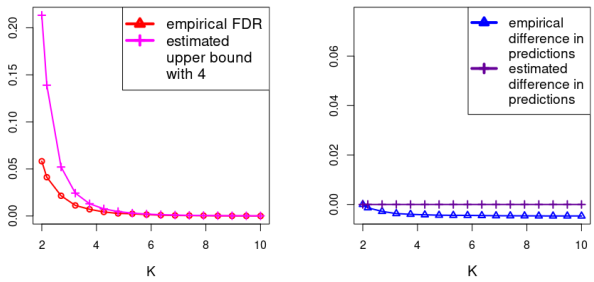


$\sigma^2 = 1$



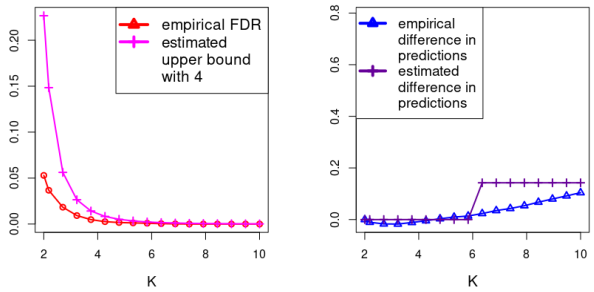
$\sigma^2 = 4$

Figure S-14: Curves of the relative standard deviation (standard deviation normalized by the mean) of the functions $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ obtained from 100 data sets. With each one, $\hat{\beta}_{\hat{m}(1)}, \hat{\beta}_{\hat{m}(1.5)}, \hat{\beta}_{\hat{m}(2)}, \hat{\beta}_{\hat{m}(2.5)}, \hat{\beta}_{\hat{m}(3)}, \hat{\beta}_{\hat{m}(3.5)}, \hat{\beta}_{\hat{m}(4)}, \hat{\beta}_{\hat{m}(4.5)}, \hat{\beta}_{\hat{m}(5)}$ and $\hat{\beta}_{\hat{m}(\log(n))}$ are calculated given $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$, variance of the 100 $B(K, \hat{\beta}_{\hat{m}(\tilde{K})}, \hat{\sigma}^2)$ functions and then the relative standard deviation with respect to K. Top: for $\sigma^2 = 0.1$. Middle: for $\sigma^2 = 1$. Bottom: for $\sigma^2 = 4$.



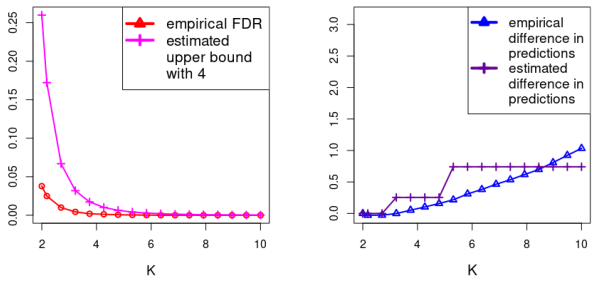
FDR diff-PR

$\sigma^2 = 0.1$



FDR diff-PR

$\sigma^2 = 1$



FDR diff-PR

$\sigma^2 = 4$

Figure S-15: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and of $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for $K \geq 2$ for the *toy data set*. Top: for $\sigma^2 = 0.1$. Middle: for $\sigma^2 = 1$. Bottom: for $\sigma^2 = 4$.

3. Complementary studies and graphs for non-ordering variable selection

This section completes Subsection 4.3 of [3] about the assessment of our approach to non-ordered variable selection by considering scenarios described in Table 4 of [3].

In particular, graphs for scenarios (ii) to (iv) described in Table 4 are provided. The empirical FDR, the theoretical FDR, the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$, the diff-PR$\big(\hat{m}(K)\big)$ and the $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ functions, calculated on the three perturbed model collections described in Subsection 4.3.1 are plotted in Figures S-16- S-18. They are devoted to study the robustness to variable ordering (Subsection 3.1). Table S-1 contains the proportion of active variables in models of size 5, 10, 15 an 20 for random collections built with Bolasso, SLOPE, random forest and the knockoff method. It is devoted to compare methods for the reconstruction of variable ordering (Subsection 3.2). Lastly, Tables S-2- S-4 contain the dimension, PR and FDR of the selected models obtained by LinSelect, the 50-fold CV, the knockoff method and our algorithm, respectively applied on the nested model collection (Equation 2.1 of [3]), the random collection built with Bolasso and the random collection built with the knockoff method. They are devoted to compare our algorithm and the three considered variable selection methods (Subsection 3.3).
All the R scripts are available at `https://github.com/PerrineLacroix/Trade_off_FDR_PR`.

3.1. Robustness to variable ordering

Figures S-16 to S-18 show similar results to Subsection 4.3.1 of [3] about the robustness to variable ordering. This confirms that being able to discriminate between active and non-active variables is crucial. For scenario (ii), $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ values begin to diverge from those of $B(K, \beta^*, \sigma^2)$ for the second and the third configurations where the coefficients of β^* are close to each other. Concerning the second configuration, $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ and $B(K, \beta^*, \sigma^2)$ values are both larger than the empirical FDR ones which is expected. When permutations of the first twelve and fifteen variables are processed, FDR values are, in most cases, even higher along the collections than for the *toy data set*, especially for scenario (ii) configuration (iii) and for scenario (iv) with $\sigma^2 = 4$ for which distinction between variables is more difficult; the PR values increase faster than for the *toy data set*. A meticulous study on the choice of the parameters γ and α is required to get low values of both PR and FDR.

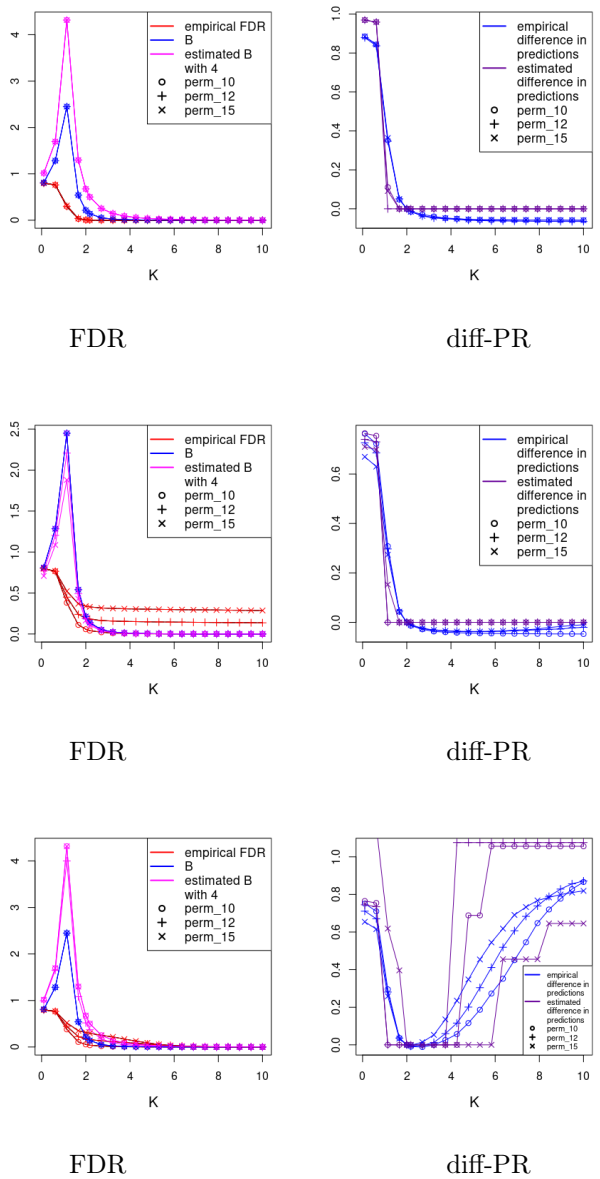


Figure S-16: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \beta^*, \sigma^2)$ functions (blue), the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for the *toy data set* and for the three perturbed collections. Top: for $\beta^*_{10} = \frac{2}{10}$. Middle: for $\beta^*_{10} = 2$ and distant coefficients. Bottom: for $\beta^*_{10} = 2$ and close coefficients.

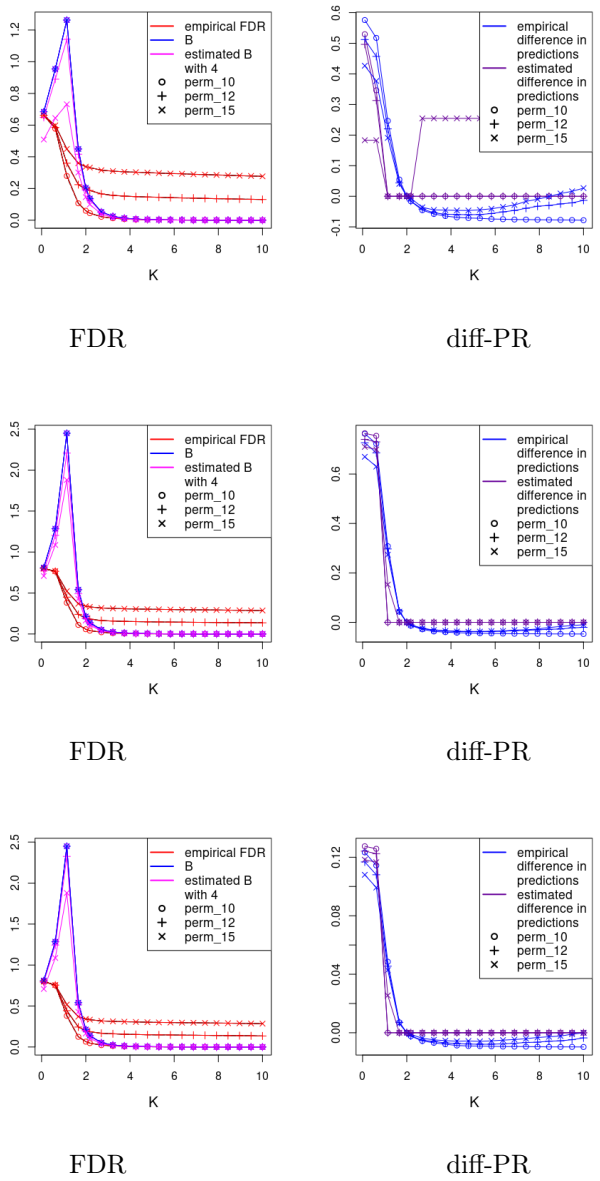


Figure S-17: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \beta^*, \sigma^2)$ functions (blue), the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for the *toy data set* and for the three perturbed collections. Top: for $n = 30$. Middle: for $n = 50$. Bottom: for $n = 300$.

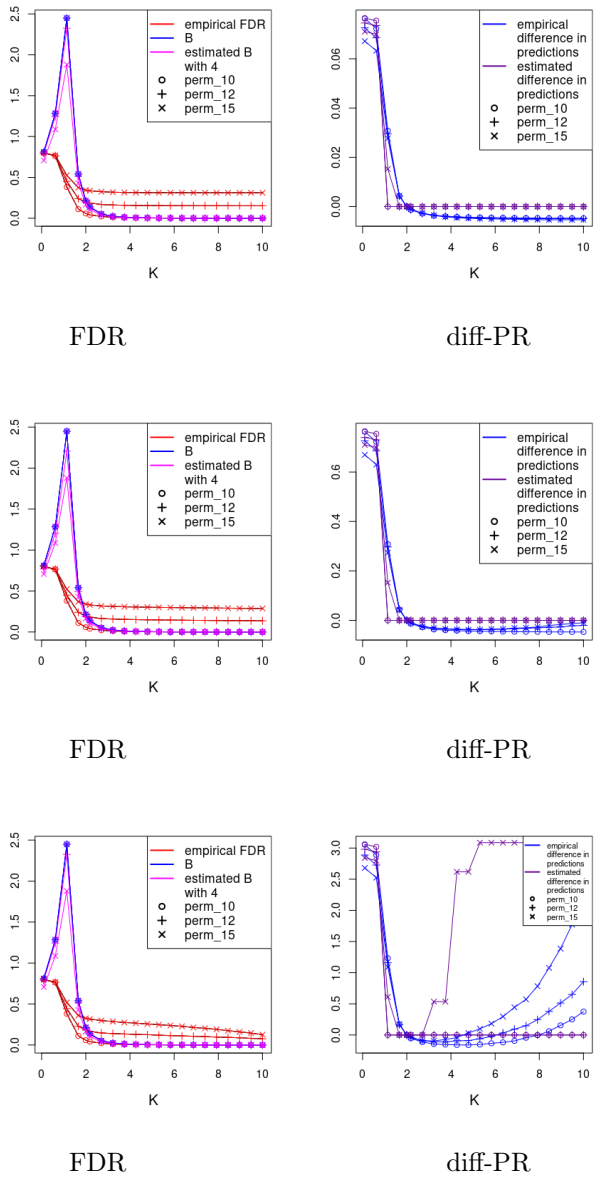


Figure S-18: Curves of the empirical functions $\mathrm{FDR}\big(\hat{m}(K)\big)$ (red) and diff-$\mathrm{PR}\big(\hat{m}(K)\big)$ (blue), of the $B(K, \beta^*, \sigma^2)$ functions (blue), the $B(K, \hat{\beta}_{\hat{m}(4)}, \hat{\sigma}^2)$ functions (pink) and $\widehat{\text{diff-PR}}\big(\hat{m}(K)\big)$ (violet) for the *toy data set* and for the three perturbed collections. Top: for $\sigma^2 = 0.1$. Middle: for $\sigma^2 = 1$. Bottom: for $\sigma^2 = 4$.

3.2. Random variable order

This subsection completes Subsection 4.3.2 of [3] about the reconstruction of variable ordering.

Table S-1 shows the proportion of active variables in models of size 5, 10, 15 and 20 for random collection built with Bolasso, SLOPE, random forest and the knockoff method. Among the random model collections, the knockoff method provides the highest values for all scenarios except scenario (iv) with $\sigma^2 = 0.1$ and some models of size 20 where Bolasso is the best method. Results deteriorate for specific scenarios : around 0.8 for scenario (iv) with $\sigma^2 = 4$, 0.6 for scenario (ii) with the third configuration and 0.8 for scenario (ii) with the second configuration for which the discrimination between active and non-active variables is naturally less obvious.

	Bolasso	SLOPE	random forests	the knockoff method
Scenario (ii) config. (ii)				
$D_m = 5$	0.25	0.24	0.24	0.26
$D_m = 10$	0.23	0.23	0.23	0.24
$D_m = 15$	0.34	0.34	0.34	0.34
$D_m = 20$	0.44	0.44	0.44	0.43
Scenario (ii) config. (iii)				
$D_m = 5$	0.63	0.60	0.63	0.74
$D_m = 10$	0.54	0.52	0.53	0.58
$D_m = 15$	0.69	0.68	0.69	0.69
$D_m = 20$	0.79	0.78	0.79	0.75
Scenario (iii), $n = 30$				
$D_m = 5$	0.96	0.95	0.95	*the knockoff method*
$D_m = 10$	1.00	1.00	1.00	*is not adapted*
$D_m = 15$	1.00	1.00	1.00	*to the* $n < p$ *case*
$D_m = 20$	1.00	1.00	1.00	
Scenario (iii), $n = 300$				
$D_m = 5$	0.98	0.91	0.92	1.00
$D_m = 10$	0.83	0.73	0.78	0.92
$D_m = 15$	0.92	0.86	0.91	0.96
$D_m = 20$	0.96	0.92	0.96	0.97
Scenario (iv), $\sigma^2 = 0.1$				
$D_m = 5$	1.00	1.00	1.00	1.00
$D_m = 10$	0.99	0.99	0.94	0.98
$D_m = 15$	1.00	1.00	0.98	0.98
$D_m = 20$	1.00	1.00	0.99	0.98
Scenario (iv), $\sigma^2 = 4$				
$D_m = 5$	0.86	0.85	0.82	0.96
$D_m = 10$	0.66	0.66	0.65	0.70
$D_m = 15$	0.78	0.77	0.77	0.78
$D_m = 20$	0.85	0.84	0.84	0.82

TABLE S-1

Active variable proportions in models of size 5, 10, 15 *and* 20 *for random collections built with Bolasso, SLOPE, random forest and the knockoff method for scenarios (ii)-(iv) of Table 4 of [3]. Values are the average over* 100 *iterations.*

3.3. Comparison with other variable selection methods

This subsection completes Subsection 4.2 and Subsection 4.4 of [3] by comparing Algorithm 1 and the three variable selection methods (presented below) on scenarios described in Table 4 and from the different considered model collections. With the nested model collection (Equation 2.1 of [3]) and with $\alpha = 0.05$ and $\gamma = 0.1$, algorithm 1 of [3] provides $K = 2.8$ for scenario (i) with $D_m^* = 20$ and $K = 3.3$ for all others except for scenario (i) with $D_m^* = 1$, for scenario (ii) with the second and the third configurations and for scenario (iv) with $\sigma^2 = 4$. Concerning these last four scenarios, the intersection of I_1 and I_2 is empty. The minimum of I_1 equals 4.8 for scenario (i) with $D_m^* = 1$ and for scenario (ii) with the second configuration, and equals 3.3 for scenario (ii) with the third configuration and for scenario (iv) with $\sigma^2 = 4$. To get a non-empty intersection, γ or α has to be higher. In all these examples, we observe that the value of K provided by taking $\min(I_1 \cap I_2)$ coincides with $\min(I_1)$, so increasing the value of γ does not change K. However, increasing the value of α provides smaller values for K. When $\alpha = 0.1$ and $\gamma = 0.1$, the intersection of I_1 and I_2 is empty and the obtained values of K from $\min(I_1)$ are 3.8 for scenario (i) with $D_m^* = 1$ and for scenario (ii) with the second configuration and 2.8 for scenario (ii) with the third configuration and for scenario (iv) with $\sigma^2 = 4$. Hence, these four cases are typical examples where the choice of K depends strongly on the chosen balance between PR and FDR. In all cases, we always notice that whatever the given balance, the K provided from algorithm 1 coincides with the one given by the trade-off between the two empirical quantities of PR and FDR.

When we compare our algorithm application with the three considered existing variable selection methods (Tables S-2-S-4), all observations mentioned in Subsection 4.4 of [3] remain valid over the different scenarios

	$D_{\hat{m}}$	PR($\hat{m}$)	FDR($\hat{m}$)
Scenario (ii) config. (ii)			
LinSelect	0.00	1.06	0.00
50-fold CV	23.15	1.47	0.45
Our algorithm	0.27	1.08	0.00
Scenario (ii) config. (iii)			
LinSelect	0.03	2.21	0.00
50-fold CV	18.65	1.78	0.33
Our algorithm	7.55	1.39	0.00
Scenario (iii), $n = 30$			
LinSelect	2.00	14.41	0.00
50-fold CV	12.00	3.57	0.20
Our algorithm	9.00	1.43	0.01
Scenario (iii), $n = 300$			
LinSelect	2.07	14.41	0.00
50-fold CV	11.81	3.57	0.20
Our algorithm	9.38	1.43	0.01
Scenario (iv), $\sigma^2 = 0.1$			
LinSelect	10.27	0.12	0.02
50-fold CV	28.18	0.35	0.47
Our algorithm	10.07	0.12	0.00
Scenario (iv), $\sigma^2 = 4$			
LinSelect	3.37	10.94	0.00
50-fold CV	25.13	7.74	0.44
Our algorithm	7.85	5.12	0.00

TABLE S-2

Results of the dimension, PR and FDR of the selected models obtained by LinSelect, the 50*-fold CV and our algorithm, applied on the nested model collection (Equation 2.1 of [3]) for the scenarios (ii), (ii) and (iv) described in Table 4 of [3]. Each value is the average over* 100 *independent iterations. PR and FDR of each selected model are the empirical quantities. Input parameters of our algorithm are fixed to* $\gamma = 0.1$ *and* $\alpha = 0.05$.

	$D_{\hat{m}}$	PR($\hat{m}$)	FDR($\hat{m}$)
Scenario (ii) config. (ii)			
LinSelect	0.02	1.06	0.00
50-fold CV	23.51	1.71	0.44
Our algorithm	6.80	1.50	0.05
Scenario (ii) config. (iii)			
LinSelect	0.06	2.22	0.00
50-fold CV	24.83	1.98	0.47
Our algorithm	13.86	1.85	0.25
Scenario (iii), $n = 30$			
LinSelect	1.47	15.89	0.00
50-fold CV	14.66	3.66	0.28
Our algorithm	12.65	1.79	0.19
Scenario (iii), $n = 300$			
LinSelect	11.58	1.17	0.13
50-fold CV	21.68	1.34	0.40
Our algorithm	15.20	1.11	0.30
Scenario (iv), $\sigma^2 = 0.1$			
LinSelect	11.24	0.14	0.08
50-fold CV	27.58	0.47	0.46
Our algorithm	13.60	0.15	0.24
Scenario (iv), $\sigma^2 = 4$			
LinSelect	3.44	12.91	0.01
50-fold CV	24.23	8.53	0.44
Our algorithm	13.93	6.73	0.25

TABLE S-3

Results of the dimension, PR and FDR of the selected models obtained by LinSelect, the 50*-fold CV and our algorithm, applied on the random collections built with Bolasso for the scenarios (ii), (ii) and (iv) described in Table 4 of [3]. Each value is the average over* 100 *independent iterations. PR and FDR of each selected model are the empirical quantities. Input parameters of our algorithm are fixed to* $\gamma = 0.1$ *and* $\alpha = 0.05$.

	$D_{\hat{m}}$	PR($\hat{m}$)	FDR($\hat{m}$)
Scenario (ii) config. (ii)			
LinSelect	0.04	1.07	0.00
50-fold CV	24.97	1.70	0.45
Knockoff	0.00	1.06	0.00
Our algorithm	4.31	1.43	0.00
Scenario (ii) config. (iii)			
LinSelect	0.07	2.22	0.00
50-fold CV	24.25	1.99	0.43
Knockoff	0.00	2.22	0.00
Our algorithm	9.42	1.84	0.07
Scenario (iii) $n = 30$			
LinSelect			*the knockoff method*
50-fold CV			*is not adapted*
Knockoff			*to the $n < p$ case*
Our algorithm			
Scenario (iii), $n = 300$			
LinSelect	9.84	1.07	0.03
50-fold CV	23.21	1.21	0.44
Knockoff	0.21	2.65	0.01
Our algorithm	13.39	1.10	0.23
Scenario (iv), $\sigma^2 = 0.1$			
LinSelect	10.07	0.31	0.04
50-fold CV	23.64	0.47	0.38
Knockoff	0.00	13.18	0.00
Our algorithm	21.02	0.16	0.32
Scenario (iv), $\sigma^2 = 4$			
LinSelect	4.44	10.17	0.00
50-fold CV	21.12	7.91	0.38
Knockoff	0.00	17.12	0.00
Our algorithm	10.58	6.62	0.10

TABLE S-4

Results of the dimension, PR and FDR of the selected models obtained by LinSelect, the 50-fold CV, the knockoff method and our algorithm, applied on the random collections built with the knockoff method for the scenarios (ii), (ii) and (iv) described in Table 4 of [3]. Each value is the average over 100 independent iterations. PR and FDR of each selected model are the empirical quantities. Input parameters of our algorithm are fixed to $\gamma = 0.1$ and $\alpha = 0.05$.

References

[1] Baudry, J. P., Maugis, C. and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing* **22** 455–470.

[2] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability theory and related fields* **138** 33–73.

[3] Lacroix, P. and Martin, M.-L. (2023). Trade-off between predictive performance and FDR control for high-dimensional Gaussian model selection. *arXiv preprint arXiv:2302.01831*.