The revisited knockoffs method for variable selection in L_1-penalised regressions.

Anne Gégout-Petit[a], Aurélie Gueudin-Muller[a] & Clémence Karmann[a]

[a] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France, Inria BIGS Team ; anne.gegout-petit@univ-lorraine.fr, aurelie.gueudin@univ-lorraine.fr, clemence.karmann@inria.fr*

Abstract. We consider the problem of variable selection in regression models. In particular, we are interested in selecting explanatory covariates linked with the response variable and we want to determine which covariates are relevant, that is which covariates are involved in the model. In this framework, we deal with L_1-penalised regression models. To handle the choice of the penalty parameter to perform variable selection, we develop a new method based on the knockoffs idea. This revisited knockoffs method is general, suitable for a wide range of regressions with various types of response variables. Besides, it also works when the number of observations is smaller than the number of covariates and gives an order of importance of the covariates. Finally, we provide many experimental results to corroborate our method and compare it with other variable selection methods.

Keywords. Variable selection; knockoffs; regression; Lasso penalisation.

1 Introduction

Regression methods are really helpful to analyse dependencies between a variable, named the response, and one or several explanatory covariates. This is one of the reasons why they are widely used and studied in statistical analysis [7]. Many models have been introduced including the well-known linear regression for a continuous response variable or logistic regression for a binary response variable. Indeed, many data sets involve this last situation such as the occurence of a disease in medicine or voting intentions in econometrics. Another type of data is nominal data (that is unordered categorical data) like housing types or food choice of predators. The situation is a bit more complicated when the response is ordered categorical (ordinal), e.g. different stages of cancer, pain scales, place ratings on Google or data collected from surveys (0: never, 10: always). Logistic regression can naturally be extended to the case where the response is nominal. For such data, many authors [1, 11, 10, 16] provided models based on odds ratios such as cumulative link models, adjacent-categories logit models or continuation-ratio logit models. The choice of one of these models depends on the kind of problem. In this paper, we concentrate on a restricted but large spectrum of regression models including all regression models mentioned above.

Although prediction and interpretation provide major challenges in regression motivations, another important issue is to identify the influential explanatory covariates, that is variable selection. Selection problems often arise in many fields including biology [20]. For example, in microarray cancer diagnosis [22], a primary goal is to understand which genes are relevant. For cost and time reasons, it can also be convenient for biologists to restrict their studies to a smaller subset of explanatory covariates (genes, bacteria populations...). Accordingly, the sparsity assumption (that is, a few number of relevant explanatory covariates) is frequently suitable and adequate, even crucial for interpretation. Indeed, with

a large number of covariates, it is also useful for interpretation to determine a smaller subset of covariates that have the strongest effects. Besides, when the number of covariates is larger than the number of observations or when covariates are highly correlated, standard regression methods become inappropriate.

Lasso penalisation, or L_1-penalisation, introduced by Tibshirani [17] offers an attractive solution to these issues. That includes a L_1 penalty in the estimation of the regression coefficients in order to perform variable selection by optimising a convex criterion. The regularisation resulting from Lasso penalisation shrinks down to zero the coefficients of explanatory covariates that have the lowest effects and leads to sparse solutions and more interpretable models, making Lasso one of the most popular penalisation [8, 21, 12].

Moreover, using Lasso induces the critical choice of the penalty parameter which controls the number of selected covariables. This choice is major because two close values of the penalty parameter can often lead to very different scientific conclusions. Many general techniques have been proposed in the literature but they do not have the same purposes. For instance, K-fold cross validation emphasises prediction, the validation step involving computing the prediction error and aiming at minimising this. Furthermore, cross validation is often quite greedy and tends to overfit the data [18]. Other techniques, like StARS [9], can be adapted to a regression framework and aim at 'overselecting', that is selecting a larger set of covariates which contains the relevant ones, allowing false positive detections. Some frameworks such as gene regulatory networks require this choice: indeed, false positive detections can then be eliminated by further biological experiments whereas omitted interactions cannot be recovered after that. On the contrary, one can prefer selecting a set of covariates included in the set of true covariates to avoid false positive detections that is 'underselecting'. This constraint comes from the fact that after selection, the relevant covariates have to be studied by scientists through new experiments. But new experiments are generally expensive or time-consuming and it would be a waste to involve noisy covariates. In this paper, we focus on variable selection in the former case. Compared with our goals, we propose an intuitive and general method for automatic variable selection, inspired from the knockoffs idea of Barber and Candes [5, 6]. This method uses a matrix of knockoffs of the covariates built by swapping the rows of the matrix of covariates. This knockoffs matrix is thus random and aims at determining if a covariable belongs to the model using a decision rule based on change detection methods. One of the major advantages is that it can be performed in a wide range of regression frameworks including when the number of covariates is much larger than the number of observations. We will see that our method does not lead to a choice of the penalty parameter. Nevertheless, it provides an order of importance on the covariates allowing to select covariates according to the target.

In this paper, we address the problem of variable selection in L_1-penalised regressions. Our goal is to determine which covariates are relevant and which are noisy. We achieve it by proposing a new method of type knockoffs. The rest of the paper is organised as follows. In Section 2, we first introduce the background and describe the knockoffs method for variable selection. We also describe briefly our R package `kosel` in which the revisited knockoffs method is implemented. In Section 3, we give many illustrations and results of our method on simulated data. Furthermore, we propose a way to exploit randomness of our procedure.

2 Revisited knockoffs method

2.1 Background

Consider we have p explanatory $\mathbb{R}$-valued covariates $\overrightarrow{X} := (X_1, X_2, \ldots, X_p)$ and a response variable Y linked with $\overrightarrow{X}$ by m equations of the type:

$$f_k(\mu_k(Y|X)) = \alpha_k + \beta_1 X_1 + \ldots + \beta_p X_p, \ k = 1, \ldots, m, \tag{1}$$

where f_k is a deterministic function, $\mu_k(Y|X)$ parameters of the distribution of Y given X and $\alpha_k, \beta_1, \ldots, \beta_p$ real coefficients. Note that the vector of regression coefficients $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_p)$ does not depend on k.

This framework is quite general and emcompasses many regression models such as generalised linear models [2] (linear regression, logistic regression, Poisson regression, multinomial regression), ordinal logistic regression models [1] (cumulative logit models with proportional odds [15, 19, 3], adjacent-categories logit models, continuation-ratio logit models) or cumulative link models [1]. Indeed, for generalised linear models, $m = 1$, $\mu_1(Y|X) = \mathbb{E}(Y|X)$ and f_1 is the link function (identity, log, logit...) of the corresponding model. For ordinal logistic regression models, f_k = logit and $\mu_k(Y|X) = \mathbb{P}(Y \leq k|X)$ (cumulative), $\mu_k(Y|X) = \mathbb{P}(Y = k|Y = k \text{ or } Y = k+1, X)$ (adjacent), $\mu_k(Y|X) = \mathbb{P}(Y > k|Y \leq k, X)$ (continuation). Notice that these last models only allow identical effects of the covariates, which implies that the regression coefficients β_i do not depend on the modality k of the response variable Y. In particular, this framework includes models for many types of response variable such as binary, continuous, ordinal or categorical.

In this framework, covariates $X_l, \ l = 1, \ldots, p$ have to be linked to the response variable Y through a linear expression so that the conditional dependence between Y and X_l given $X_1, \ldots, X_{l-1}, X_{l+1}, \ldots, X_p$ can be measured through the regression coefficient β_l. More precisely, $\beta_l = 0$ means that X_l and Y are independent conditionally on the other covariates $X_k, \ k = 1, \ldots, l-1, l+1, \ldots p$. We are thus interested in the nullity of the regression coefficients β_l to select the relevant covariates. Moreover, we make sparsity assumption, that is a relatively small number of covariates play an important role. This implies that only a few covariates are relevant and thus, only a few regression coefficients β_l are non-null. This sparsity assumption is convenient for scientists to restrict their studies to a smaller subset of covariates, namely in high-dimensional settings. Instead of checking the nullity of each coefficient β_l by performing statistical tests, we add a L_1-penalisation on the coefficients $\boldsymbol{\beta}$ in the estimation of the coefficients of the model. Coefficients are usually estimated by solving the optimisation problem:

$$\underset{(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \ L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}), \tag{2}$$

where $L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X})$ is a function of the coefficients relative to the model (like log-likelihood), depending on the observations $\mathbf{Y}$ and $\mathbf{X}$ of the response variable Y and the vector of covariates $\overrightarrow{X}$ respectively. Instead of estimating the coefficients by (2), we add a Lasso penalisation on the coefficients vector $\boldsymbol{\beta}$ which leads to solve the following optimisation problem:

$$\underset{(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\operatorname{argmax}} \ \big\{ L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, \mathbf{X}) - \lambda ||\boldsymbol{\beta}||_1 \big\}, \tag{3}$$

where $\lambda > 0$ is the penalty parameter.

Usually, all penalisation methods require the choice of the (positive) penalty parameter, also referred as tuning or regularisation parameter. We then need to tune the penalty parameter λ (involved in the optimisation problem (3)) which controls the number of selected covariates: the larger λ is, the fewer the selected covariates are. On the contrary, values of λ closed to 0 lead to the full model, that is the model with all the covariates. We remind that our goal is to select only relevant covariates and thus, to avoid false positive detections (the wrongly detected covariates).

With regard to our problems and goals, we propose a new method, inspired from the knockoffs process used by Barber and Candes [5] in the linear Gaussian regression setting. Actually, this method does not lead to a choice of the penalty parameter λ but it puts the explanatory covariates in order from the most relevant to the least. Furthermore, it suits any regression of the type presented in (1) including when the number n of observations is smaller than the number p of covariates. Obviously, in the linear Gaussian model, it is much more pertinent to use the procedure described in [5] because of their theoretical guarantees. Even if their procedure initially held for $n > p$, they subsequently extended it thanks to a preliminary screening step [6]. In what follows, we present the principle of our revisited knockoffs method.

2.2 Principle and generalities

Let $\mathbf{X}$ denote the $n \times p$ matrix of the n observations of the p-vector $\overrightarrow{X} = (X_1, \ldots, X_p)$ of covariates, called the design matrix. The principle, given in [5], is to use a matrix $\tilde{\mathbf{X}}$ of knockoffs (of the covariates X_i) whose covariance structure is similar to $\mathbf{X}$ but independent from $\mathbf{Y}$. The goal is to determine if a covariate X_i is relevant by studying if it enters the model before its knockoff $\tilde{X}_i$, that is if X_i enters the model for a larger value of the penalty parameter λ. Indeed, as the knockoff matrix is independent from $\mathbf{Y}$, if a covariate enters the model after its knockoff, we can rightfully suspect that this covariate does not belong to the model.

We mainly differ from the method proposed by [5] in the construction of the knockoffs. In their paper, they propose a sophisticated construction of the knockoff filter using linear algebra tools. This construction allows to control the false discovery rate (FDR) –the expected fraction of false discoveries among all discoveries– in the linear Gaussian model whenever there are at least as many observations as covariates. This difference in the construction of the knockoffs makes our method suitable for the setting $n < p$ and for a larger set of regression models. Nevertheless, theoretical guarantees about the control of the false discovery rate do not hold anymore.
We construct our knockoff matrix $\tilde{\mathbf{X}}$ by randomly swapping the n rows of the design matrix $\mathbf{X}$. This way, the correlations between the knockoffs remain the same as the original variables but the knockoffs are not linked to the response $\mathbf{Y}$. Note that this construction of the knockoffs matrix also makes the procedure random. Then, in the same way as [5], we perform the regression of $\mathbf{Y}$ on the $n \times 2p$ augmented matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$, i.e. the columnwise concatenation of $\mathbf{X}$ and $\tilde{\mathbf{X}}$. Let us note $\hat{\boldsymbol{\beta}}(\lambda)$ the estimated regression

coefficients of the λ-penalised regression of $\mathbf{Y}$ on the augmented matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$:

$$\Big(\hat{\boldsymbol{\alpha}}(\lambda), \hat{\boldsymbol{\beta}}(\lambda)\Big) := \underset{(\boldsymbol{\alpha},\boldsymbol{\beta})}{\operatorname{argmax}} \left\{L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{Y}, [\mathbf{X}, \tilde{\mathbf{X}}]) - \lambda ||\boldsymbol{\beta}||_1\right\}.$$

For each variable of the augmented design, that is for each covariate and its corresponding knockoff, we consider $T_i := \sup\{\lambda > 0,\ \hat{\beta}_i(\lambda) \neq 0\},\ i \in \{1, \ldots, p, p+1, \ldots, 2p\}$. Statistics T_i correspond to the largest value of λ for which the covariate X_i if $i \in \{1, \ldots, p\}$ or its knockoff $\tilde{X}_{i-p}$ if $i \in \{p+1, \ldots, 2p\}$ first enters the model. At this stage, we hope that T_i is large for the relevant covariate, that is for $X_i, i \in \{1, \ldots, p\}$ such that $\beta_i \neq 0$ and small for the knockoffs variables $X_i := \tilde{X}_{i-p},\ i \in \{p+1, \ldots, 2p\}$ or for the noisy covariates $X_i,\ i \in \{1, \ldots, p\}$ such that $\beta_i = 0$. This yields us a $2p$-vector $(T_1, ..., T_p, \tilde{T}_1, ..., \tilde{T}_p)$ where $\tilde{T}_i$ denotes T_{i+p}. Then, we consider, for all $i \in \{1, ..., p\}$, $W_i := \max(T_i, \tilde{T}_i) \times \begin{cases} (+1) & \text{if } T_i > \tilde{T}_i \\ (-1) & \text{if } T_i \leq \tilde{T}_i \end{cases}$.

Statistics W_i allo to determine if a covariate enters the model before or after its knockoff. A negative value for W_i means that the covariate X_i enters the model after its knockoff and we eliminate it. On the contrary, a positive value for W_i means that the covariate X_i enters the model before its knockoff and is more likely to belong to the model. But covariates X_i whose statistic W_i is positive are not necessarily relevant: we hope that W_i is large for most of relevant covariates and small for the other ones. Thus, we are interested in the largest positive values of the p-vector of statistics W which moreover indicates that the corresponding covariate enters the model early, that is for a large value of λ. Statistics W_i allows in fact to sort the covariates according to their importance: the larger W_i is, the more relevant the associated covariate X_i is.
This then implies defining a threshold s for W_i over which we will keep the corresponding covariates in the estimated model. On the whole, we will choose the estimated model $\hat{S}$ such that:

$$\hat{S} := \{X_i : W_i \geq s\}.$$

2.3 Choice of the threshold

The second major difference with Barber and Candes [5] lies in the choice of the threshold s. They provide in fact a data-dependent threshold that shows attractive results relative to the false discovery rate in the Gaussian setting. Unfortunately, these results do not hold in general, out of the linear Gaussian model. In our method, we make the assumption that there is a breakdown in the distributions between the W_i corresponding to the covariates X_i belonging to the model and the other ones (see Figure 1). Figure 1 illustrates that distributions of W_i depend on whether X_i is relevant or not. To generate Figure 1, we have simulated a set of data under a linear Gaussian regression model with $p = 20$ independent Gaussian covariates. Only the five first ones were linked to Y:

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$ and $\epsilon \sim \mathcal{N}(0, 1)$. In our knockoffs procedure applicated to this data set, only statistics $W_1, W_2, W_3, W_4, W_5, W_6, W_7, W_{13}, W_{14}, W_{16}, W_{19}$ associated to the covariates $X_1, X_2, \ldots, X_{14}, X_{16}$ and X_{19} have a positive value. For example, the covariate X_1 entered the model for $\lambda = 1.002$ (thus, $T_1 = 1.002$) and entered the model before its knockoff $\tilde{X}_1$. This means that the knockoff variable $\tilde{X}_1$ entered the model for $\lambda < 1.04$ and implies that $\tilde{T}_1 < 1.002$. W_3 takes the largest value among all the statistics

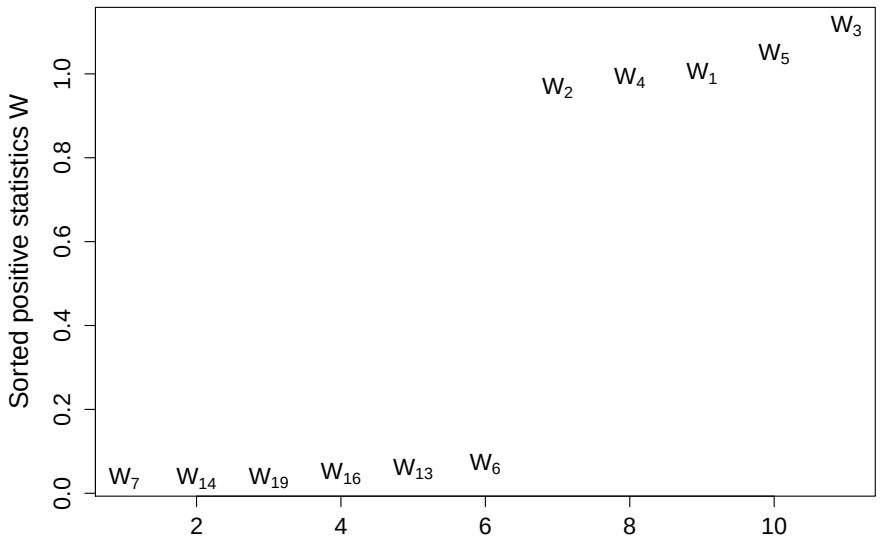


Figure 1: Example of positive statistics W_i sorted in ascending order. Linear Gaussian regression model with $n = 500$ observations of $p = 20$ covariates. Only covariates X_1, X_2, X_3, X_4 and X_5 belong to the model (regression coefficients are set to $\beta = (1, 1, 1, 1, 1, 0, \ldots, 0)$).

$W_i,\ i = 1, \ldots, 20$, which implies that X_3 is the covariate the most likely to belong to the model. We can clearly observe a breakdown between the values of the five first covariates and the other ones.

Consequently, we present two automatic ways to define the threshold s by using two change detection methods: the method proposed by Auger and Lawrence [4, 13, 14] and the CUSUM method for mean change detection. Let $W_{(i)},\ i = 1, \ldots, w$, denote the sorted w positive statistics $W_i,\ i = 1, ..., w$, that is $0 < W_{(1)} \leq W_{(2)} \leq \ldots \leq W_{(w)}$ and $e_j = W_{(j+1)} - W_{(j)}$ for all $j = 1, \ldots w-1$ the $w-1$ gaps between these sorted statistics. Remark that w, the number of positive statistics W_i, is random ($w = 11$ on Figure 1). We propose two automatic thresholds defined as:

- the minimum of the two thresholds obtained by applying these two change detection methods directly on the statistics $W_{(i)},\ i = 1, \ldots, w$ sorted in ascending order,
- the minimum of the two thresholds obtained by applying these two change detection methods on the gaps $e_j,\ j = 1, \ldots, w-1$.

Let us name the first one 'W-threshold' and the second one 'gaps-threshold' for the sake of simplicity.

2.4 R package `kosel`

Our procedures have been implemented in a R package, called `kosel` (for knockoffs selection), available on the CRAN. Our package includes three functions: `ko.glm`, `ko.ordinal`

and `ko.sel`.

The two first functions construct the knockoffs matrix and return the p-vector of statistics W for L_1-penalised regressions models respectively implemented in the R functions `glmnet` and `ordinalNet` from the packages of the same name. `glmnet` emcompasses generalised linear models whereas `ordinalNet` includes ordinal regression models such as cumulative link, adjacent or continuation-ratio or stopping-ratio. By default, a seed is used so that the knockoffs matrix remains the same (and thus, the resulting statistics vector W). But this can be changed with the option `random = TRUE` to exploit the randomness of the procedure (see Subsection 3.3 for further details).

The third function `ko.sel` deals with the choice of the threshold. It uses the statistics vector W obtained by one of the two other functions and returns the p-binary vector of estimation and the threshold s. Three choices are proposed: `method = 'stats'` and `method = 'gaps'` respectively correspond to the 'W-threshold' and 'gaps-threshold' while `method = 'manual'` allows the user to choose its own threshold. The option `print = TRUE` displays the positive statistics W_i sorted in ascending order like in Figure 1. For `method = 'manual'`, they are automatically displayed so that the user can choose its threshold. For the 'W-threshold' (`method = 'stats'`) and 'gaps-threshold' (`method = 'gaps'`), option `print = TRUE` also displays an horizontal line corresponding to the threshold.

3 Simulation studies

3.1 Settings

We now describe experimental results to study the efficiency of our procedure. For that, we have performed different simulations with various regressions: linear Gaussian regression, logistic regression and cumulative logit regression (with proportional odds). Covariates $\overrightarrow{X}$ are simulated as Gaussian such that $\mathbb{E}(X_k) = 0$ and $\text{var}(X_k) = 1$ for all $k = 1, \ldots, p$ and such that X_i and X_j are dependent conditionally on the other covariates $X_k,\ k \in \{1, \ldots, p\} \setminus \{i, j\}$ with probability 0.2. The design matrix $\mathbf{X}$ of covariates has been simulated with the R function `huge.generator` from the package `huge`, for a random graph structure. We have then simulated the observations of the response variable Y as:

$$
\begin{aligned}
Y &= \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon, && \text{(linear regression)} \\
\text{logit}(\mathbb{P}(Y = 1|X)) &= \alpha_1 + \beta_1 X_1 + \ldots + \beta_p X_p, && \text{(logistic regression)} \\
\text{or logit}(\mathbb{P}(Y \leq k|X)) &= \alpha_k + \beta_1 X_1 + \ldots + \beta_p X_p, k = 1, 2, && \text{(cumulative logit regression)}
\end{aligned}
$$

where $\epsilon \sim \mathcal{N}(0, 1)$ is Gaussian noise, the vector of regression coefficients $\boldsymbol{\beta}$ is sparse and given below and intercepts α_1 and α_2 are chosen so that the response variable $\mathbf{Y}$ takes enough values in each of its modalities ($\{0, 1\}$ for logistic regression and $\{0, 1, 2\}$ for cumulative logit regression). These regressions have been respectively performed with the R functions `glmnet` and `ordinalNet` of the eponymous R packages.

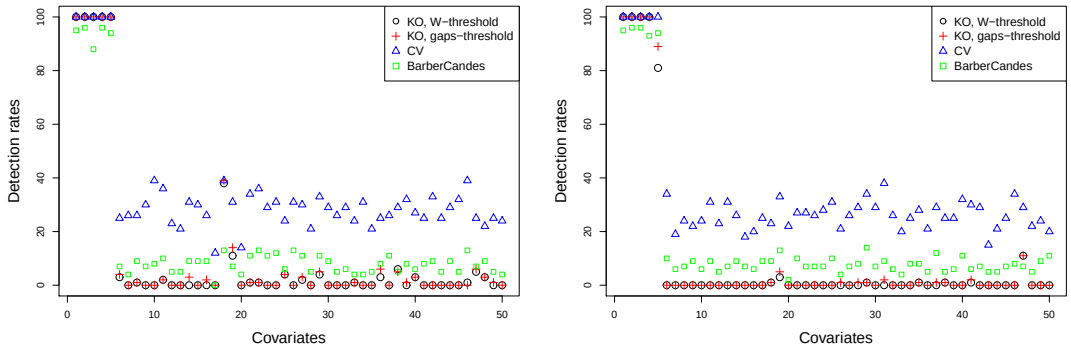
We present detection rates of each covariates on $B = 100$ repetitions for different settings. The detection rate of the covariate X_l is the number of times among the 100 repetitions where the estimated model included X_1. First, we have simulated $n = 200$ observations of $p = 50$ covariates for pedagogical reasons and next, $n = 1000$ observations of

$p = 2000$ covariates to illustrate results in a higher-dimensional framework. For $p = 2000$ covariates, results are presented as boxplots of detection rates according to the regression coefficient β in order to improve readability.
In addition, we compare our results with results obtained by cross validation. Cross validation has been performed with the R functions `cv.glmnet` for linear and logistic regressions and `ordinalNetTune` with 'logLik' tune method for cumulative logit regression. For $p = 50$, we also compare our results in the linear Gaussian setting with results obtained by Barber and Candes' procedure. Their procedure is implemented in the function `knockoff.filter` from the R package `knockoff`. We do not perform this comparison for $p = 2000$ because their procedure is not applicable due to the too few observations.

3.2 Efficiency and comparisons

3.2.1 p = 50

In the first place, we present results for $n = 200$ observations of $p = 50$ covariates. Regression coefficients are set to $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ and $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$. Covariates $\mathbf{X}$ are the same for each different regression. But they are different according to the regression coefficients $\boldsymbol{\beta}$. In other words, for a fixed value of $\boldsymbol{\beta}$, the design matrix $\mathbf{X}$ is the same for each of the three regression models. But the response variable $\mathbf{Y}$ is simulated according to the regression model and is therefore different. The knockoffs matrix is also different.

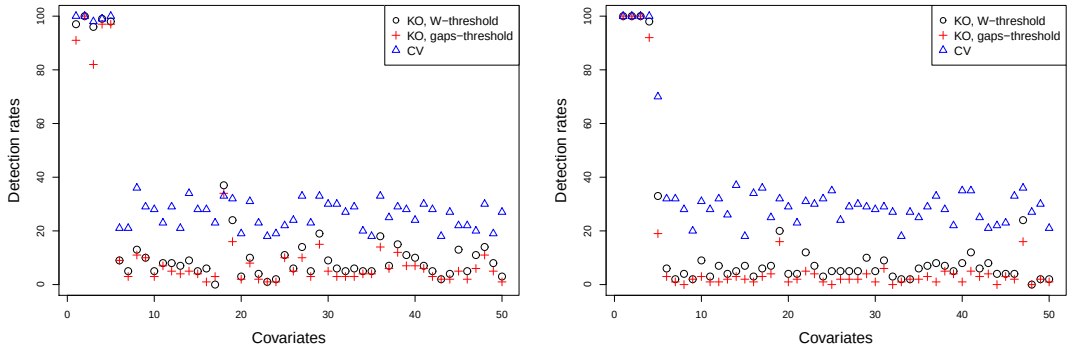


(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$.

(b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

Figure 2: Detection rates of each covariate for the four methods: revisited knockoffs W-threshold and gaps-thresholds, cross validation and Barber and Candes' knockoffs. Linear Gaussian regression model with $n = 200$ observations of $p = 50$ covariates with regression coefficients $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) and $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.

Results and comments. Figures 2, 3 and 4 show detection rates for cross validation and for the revisited knockoffs method after thresholding with the W-threshold and with the gaps-threshold. These detection rates are illustrated on Figures 2, 3 and 4 for respectively linear, logistic and cumulative logit regressions. First, we can note that our procedure is efficient for each of these regressions: the difference of detection rates of the first five



(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. (b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

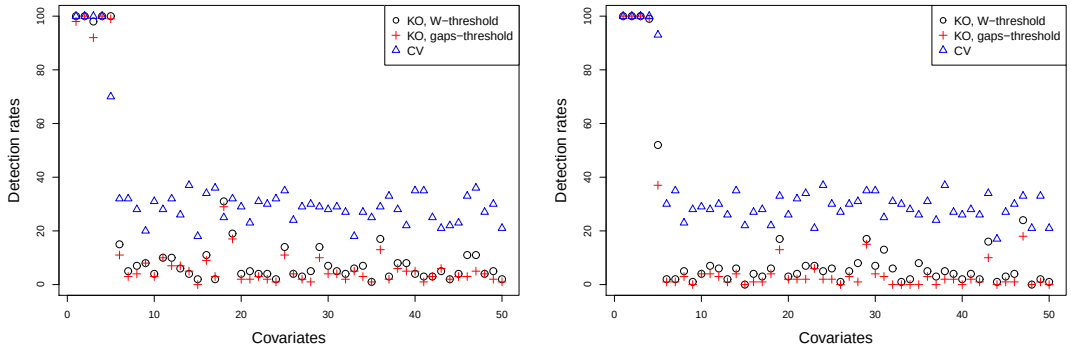
Figure 3: Detection rates of each covariate for the three methods: revisited knockoffs W-threshold and gaps-thresholds and cross validation. Logistic regression model with $n = 200$ observations of $p = 50$ covariates with regression coefficients $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) and $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.

covariates and the rest of them is really clear, regardless of the regression coefficients or the choice of the threshold. For linear regression, these two thresholds give similar results whereas for logistic and cumulative logit regressions, gaps-threshold tends to give slightly lower detection rates than W-threshold, for both relevant and noisy covariates.

In comparison, cross validation provides considerably higher detection rates: although the first five covariates, especially X_4 and X_5, can be more detected, noisy covariates are also much more detected than with our procedure. For example, for logistic regression in Figure 3, noisy covariates are almost all detected less than 20% with our procedures whereas they are detected between 20% and 40% with cross validation. In practice, using cross validation would give more false positive detections than our procedures.

Figure 2 also show detection rates obtained by Barber and Candes' knockoffs in the linear Gaussian regression model. To perform their procedure, we have to choose a target false discovery rate. In practice, we want it to be small but too small values lead to an infinite threshold and thus an empty estimated model. By default, the FDR is set to 0.1 but we set it to 0.4 to avoid too many empty estimated models. For the two different configurations of $\boldsymbol{\beta}$, we have obtained 4 empty estimated models on 100 repetitions. Because of that, detection rates of the noisy covariates tend to be a bit higher than ours. For the same reason, the five first ones are a bit less detected (close to 96%, which corresponds to the number of repetitions for which the threshold was not infinite). For $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$, X_5 is yet better detected.

All of these three figures illustrate also that detection rates depend on the regression coefficient β: the higher β is, the more the associated covariate is detected. Indeed, for $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$, we can observe that the covariate X_5 tends to be less detected than the four first ones. Furthermore, we can notice that some noisy covariates are more detected. For example, this is the case for the noisy covariates $X_{18}, X_{19}, X_{29}, X_{36}, X_{39}$ or X_{47} for $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ and for all kind of regressions (since the design matrix is the same). This is probably due to the dependence structure of $\overrightarrow{X}$. In particular, these covariates are dependent to three of the first five covariates conditionally on the other



(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. (b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

Figure 4: Detection rates of each covariate for the three methods: revisited knockoffs W-threshold and gaps-thresholds and cross validation. Cumulative logit model with $n = 200$ observations of $p = 50$ covariates with regression coefficients $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$ (a) and $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ (b). Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.

ones. Similar phenomenon occurs for $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$ with noisy covariates X_{19}, X_{41} and X_{47}.
Finally, our procedure seems to be quite efficient regardless of the regression model. Nevertheless, results are a little bit better for linear regression. In this case, we remind that Barber and Candes' procedures [5, 6] also provide theoretical guarantees.

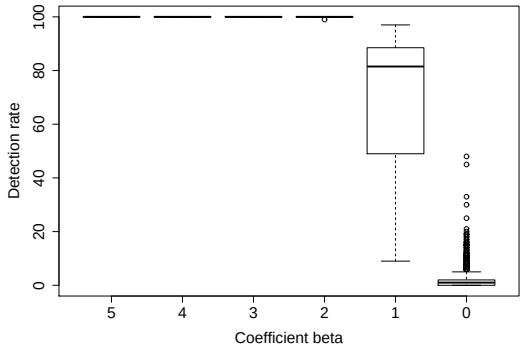
3.2.2 p = 2000

We present now results for $n = 1000$ observations of $p = 2000$ covariates to illustrate that our procedure is suitable with thousands of covariates. Regression coefficients are set to

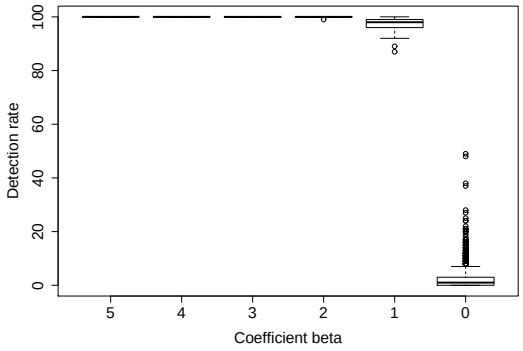
$$\beta_k = \begin{cases} 5, \text{ if } 1 \leq k \leq 20, \\ 4, \text{ if } 21 \leq k \leq 40, \\ 3, \text{ if } 41 \leq k \leq 60, \\ 2, \text{ if } 61 \leq k \leq 80, \\ 1, \text{ if } 81 \leq k \leq 100, \\ 0, \text{ otherwise.} \end{cases}$$

. In the same way as for $p = 50$ (Subsection 3.2.1), covariates $\mathbf{X}$ are the same for each different regression. But they are different according to the regression coefficients $\boldsymbol{\beta}$. In other words, for a fixed value of $\boldsymbol{\beta}$, the design matrix $\mathbf{X}$ is the same for each of the three regression models. But the response variable $\mathbf{Y}$ is simulated according to the regression model and is therefore different. The knockoffs matrix is also different.

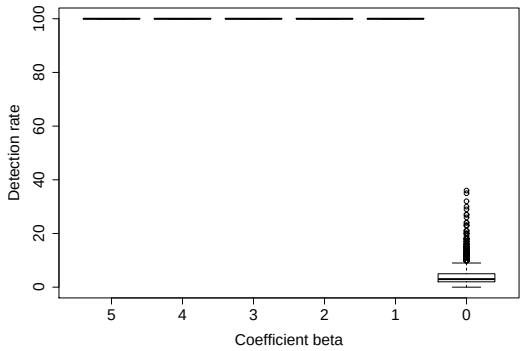
Results and comments. Figures 5, 6 and 7 each contain four graphics: three of them are boxplots of detection rate of each of the 6 groups of covariates according to their regression coefficient β. These detection rates are respectively obtained with the revisited knockoffs methods and cross validation. In order to compare our method with cross validation for the noisy covariates, we present detection rates of the noisy covariates (these for which $\beta = 0$) obtained with the knockoffs method (with gaps-threshold) in function of detection rates obtained with cross validation in the last graphic.
Results in the linear regression framework are presented in Figure 5. Comparing the



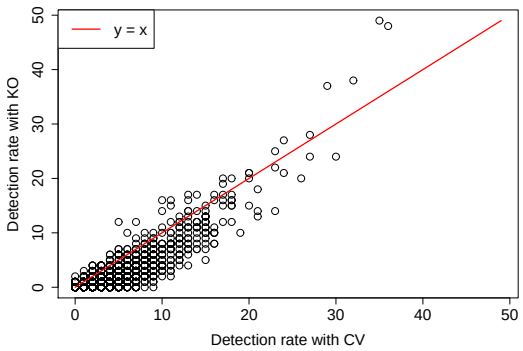
(a) Revisited knockoffs with W-threshold.



(b) Revisited knockoffs with gaps-threshold.

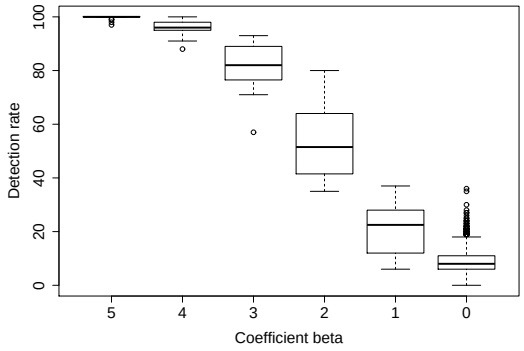


(c) Cross validation.

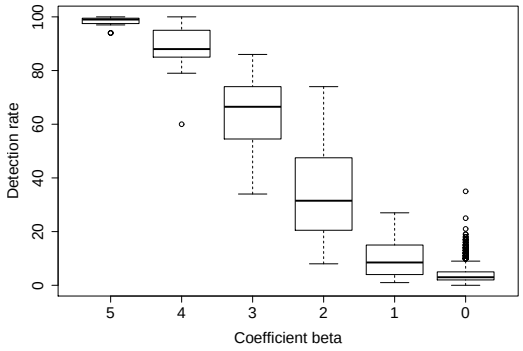


(d) Comparison of detection rates of noisy covariates for the revisited knockoffs method with gaps-threshold and for cross validation.

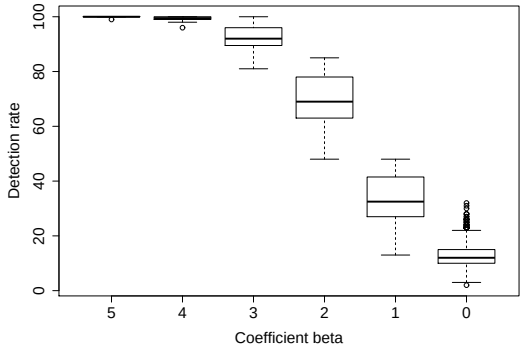
Figure 5: Boxplots of detection rates of each covariate according to their regression coefficient β for the three methods: revisited knockoffs W-threshold (a), gaps-thresholds (b) and cross validation (c). Linear Gaussian regression model with $n = 1000$ observations of $p = 2000$ covariates. Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.



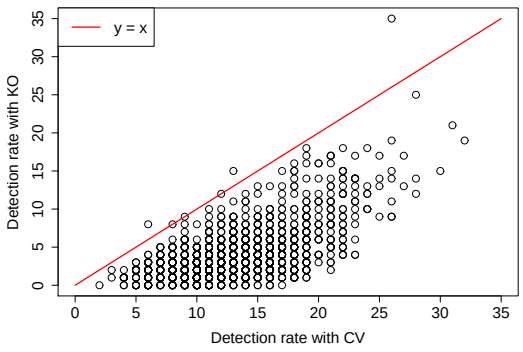
(a) Revisited knockoffs with W-threshold.



(b) Revisited knockoffs with gaps-threshold..

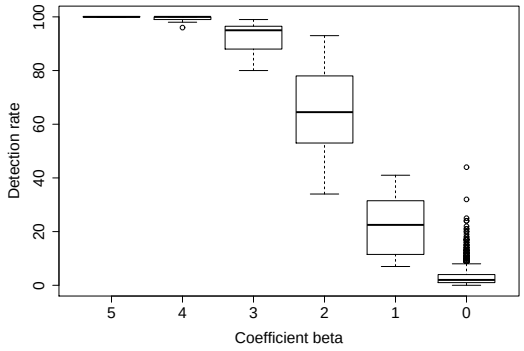


(c) Cross validation.

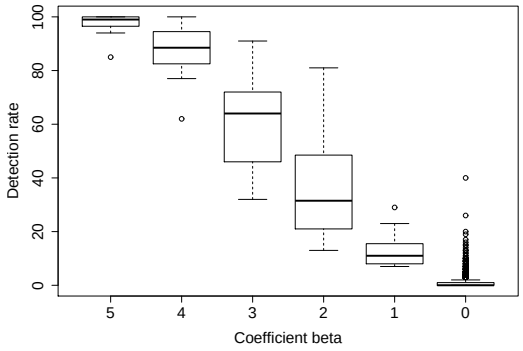


(d) Comparison of detection rates of noisy covariates for the revisited knockoffs method with gaps-threshold and for cross validation.

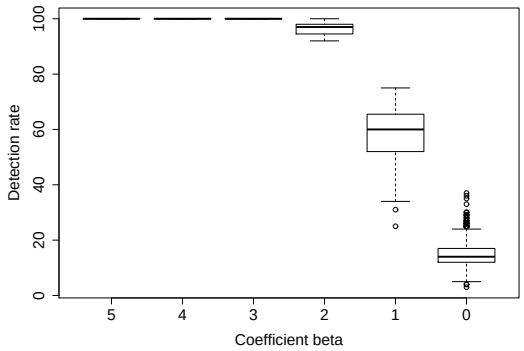
Figure 6: Boxplots of detection rates of each covariate according to their regression coefficient β for the three methods: revisited knockoffs W-threshold (a), gaps-thresholds (b) and cross validation (c). Logistic regression model with $n = 1000$ observations of $p = 2000$ covariates. Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.



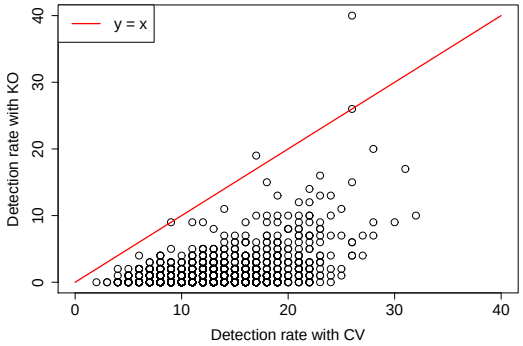
(a) Revisited knockoffs with W-threshold.



(b) Revisited knockoffs with gaps-threshold..



(c) Cross validation.



(d) Comparison of detection rates of noisy covariates for the revisited knockoffs method with gaps-threshold and for cross validation.

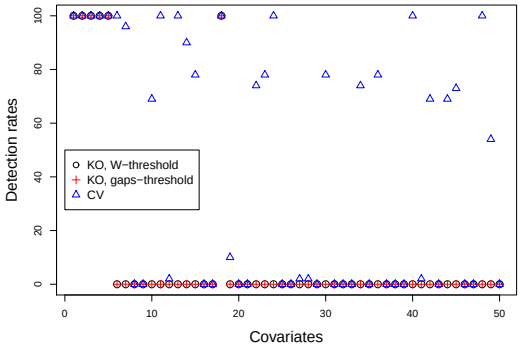
Figure 7: Boxplots of detection rates of each covariate according to their regression coefficient β for the three methods: revisited knockoffs W-threshold (a), gaps-thresholds (b) and cross validation (c). Cumulative logit model with $n = 1000$ observations of $p = 2000$ covariates. Covariates are dependent Gaussian with a random structure. The number of i.i.d. repetitions is $B = 100$.

three boxplots, we can remark that detection rates with revisited knockoffs method with W-threshold are lower than with gaps-threshold. More specifically, relevant covariates whose $\beta = 1$ are detected between 10 and 95% with W-threshold whereas they are detected more than 80% with gaps-threshold. Detection rates with W-threshold are lower and more widespread for covariates whose regression coefficient $\beta = 1$ comparing to the other relevant covariates. Cross validation leads to better detection rates for the relevant covariates. However, Figure 5d illustrates that most of the noisy covariates have higher detection rates with cross validation than with our procedure. Thus, cross validation gives more false positive detections.
Results for logistic and cumulative logit regressions are respectively presented in Figures 6 and 7. As for $p = 50$, we can note on boxplots that detection rates depend on the regression coefficient β: for all the three methods, detection rates are decreasing according to β that is, the higher β is, the more the associated covariates are detected. We also observe this on graphics 5a and 5b for linear regression, although it is less pronounced. As for linear regression, even if cross validation gives better detection rates for the relevant covariates, it also gives more fase positive detections for the noisy covariates as illustrated in graphics 6d and 7d. This phenomenon is even stronger for these two regression models where almost all of the noisy covariates are more detected with cross validation than with our procedure. Contrary to linear regression, detection rates obtained by the knockoffs method with W-threshold are higher than with gaps-threshold and they are higher for both relevant and noisy covariates.
Although detection rates are better for linear regression, our procedures lead to satisfying results for the three regression models. Even though relevant covariates are not always enough detected, detection rates of noisy covariates are also often very low, especially in comparison with cross validation. On the whole, it seems to be appropriate for sparse models regardless to the regression model and particularly when the goal is to avoid false positive detections. Notice that the threshold (W or gaps-threshold) to be used to avoid false positive detections may vary according to the regression model.

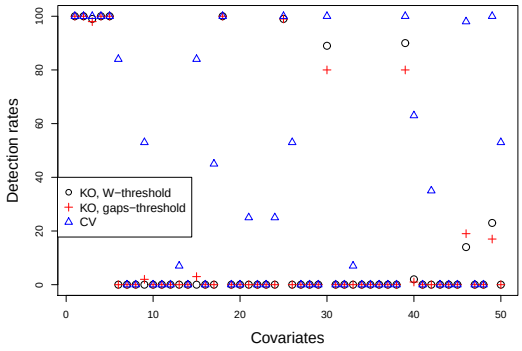
3.3 Randomness of the procedure

Note that revisited knockoffs procedure and cross validation are both random (which is not the case for Barber and Candes' procedure). Indeed, the former is random in the construction of the knockoffs matrix whereas randomness of cross validation lies in the choice of the folds. Hence, applying several time one of these methods leads to different results. To conclude this section, we compare detection rates obtained by these three methods on the same sample of data. This sample includes $n = 200$ observations of $p = 50$ covariates and consists in the first sample of the $B = 100$ samples used in the Subsection 3.2.1. Dependence structure of the vector $\overrightarrow{X}$ is thus also the same as in Subsection 3.2.1.

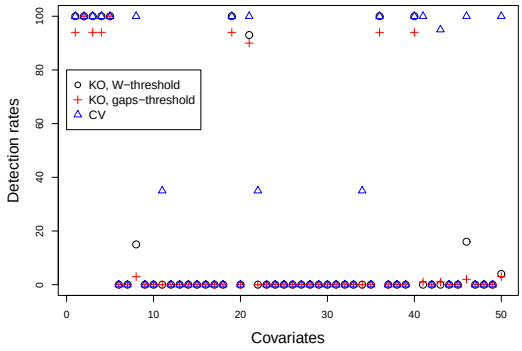
Results and comments. Figure 8 displays detection rates of each covariate using randomness of the three procedures: revisited knockoffs method with W and gaps-thresholds and cross validation. Detection rates are obtained on 100 repetitions on the same sample for the three methods and for the three regression models: linear, logistic and cumulative logit regressions. Regression coefficients are set to $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. Thus, only the first five covariates belong to the model. We can notice that these first five covariates are almost always 100% detected except with gaps-threshold for cumulative logit regression (for which they are still detected more than 95%). Noisy covariates, that is covariates



(a) Linear regression.



(b) Logistic regression.



(c) Cumulative logit regression.

Figure 8: Detection rates of each covariate for the three methods: revisited knockoffs method with the W- and gaps-thresholds (see Subsection 2.3) and cross validation. Detection rates are obtained on 100 repetitions on the same sample of $n = 200$ observations of $p = 50$ covariates with regression coefficients $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$. Covariates are dependent Gaussian with a random structure.

$X_6, \ldots, X_{50}$, are always less detected by our procedures than by cross validation. However, some of them are wrongly highly detected: X_{18} for linear regression, X_{18}, X_{25}, X_{30} and X_{39} for logistic regression or X_{19}, X_{21}, X_{36} and X_{40} for cumulative logit regression. This is probably due to the sample (in which the dependence structure is more or less pronounced). It should be recalled that this dependence structure is the same as in Subsubsection 3.2.1 for $p = 50$. In comparison with Figures 2, 3 and 4 of Subsubsection 3.2.1, we can see that these covariates have also higher detection rates. For all these covariates, cross validation gives higher detection rates than our procedures. Cross validation gives also much higher detection rates for some other noisy covariates. For instance, $X_6, X_{11}, X_{13}, X_{24}, X_{40}$ and X_{48} are always detected for linear regression whereas revisited knockoffs with W and gaps threshold never detect them. For cumulative logit model, X_8, X_{41}, X_{46} and X_{50} are always detected whereas our procedures detect them less than 15%.

In practice, with real data, this randomness opens up to further ways to perform variable selection.

4 Discussion

In this paper, we proposed a method for variable selection in regression models based on the construction of a matrix of knockoffs of the covariates. This method is quite intuitive and suitable for many types of regressions, including when the number of observations is much smaller than the number of covariates. Two different thresholds can be chosen, leading to two procedures, which have been implementend in the R package `kosel`. We have seen that these procedures both turn out to be very pertinent and efficient as the many and diverse simulations exemplify. Our two procedures are particularly appropriate when the goal is to avoid false positive detections. Indeed, even if there are false negative detections, there are also a very small rate of false positive detections. Simulations show also that efficiency of our procedures depends on the regression model. In general, we can try the two thresholds and choose results according to its target. Furthermore, randomness of our procedures provides other techniques to perform variable selection.
Nonetheless, in the case of linear Gaussian regression, Barber and Candes' procedures also offer theoretical guarantees.
In addition, our procedures give better results than cross validation with regard to false positive detections, even when we make use of randomness. However, if we aim at overselecting, it is more appropriate to use other techniques such as cross validation.

References

[1] Alan Agresti. *Analysis of ordinal categorical data*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2010.

[2] Alan Agresti. *Categorical data analysis*. Wiley series in probability and statistics. Wiley, 3ed. edition, 2013.

[3] JA Anderson and PR Philips. Regression, discrimination and measurement models for ordered categorical variables. *Applied statistics*, pages 22–31, 1981.

[4] Ivan E. Auger and Charles E. Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.*, 51(1):39–54, 1989.

[5] Rina Foygel Barber and Emmanuel J. Candès. Controlling the false discovery rate via knockoffs. *Ann. Statist.*, 43(5):2055–2085, 2015.

[6] Rina Foygel Barber and Emmanuel J Candes. A knockoff filter for high-dimensional selective inference. *arXiv preprint arXiv:1602.03574*, 2016.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[8] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[9] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1432–1440. Curran Associates, Inc., 2010.

[10] Ivy Liu and Alan Agresti. The analysis of ordered categorical data: an overview and a survey of recent developments. *Test*, 14(1):1–73, 2005. With discussion and a rejoinder by the authors.

[11] Peter McCullagh. Regression models for ordinal data. *J. Roy. Statist. Soc. Ser. B*, 42(2):109–142, 1980.

[12] Mee Young Park and Trevor Hastie. L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):659–677, 2007.

[13] Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, Gilles Celeux, and Jean-Jacques Daudin. *A statistical approach for CGH microarray data analysis*. PhD thesis, INRIA, 2004.

[14] Franck Picard, Stéphane Robin, E Lebarbier, and J-J Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 63(3):758–766, 2007.

[15] Gary Simon. Alternative analyses for the singly-ordered contingency table. *Journal of the American Statistical Association*, 69(348):971–976, 1974.

[16] Arun Sai Suggala, Eunho Yang, and Pradeep Ravikumar. Ordinal graphical models: A tale of two approaches. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3260–3269, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[17] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

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

[19] O Dale Williams and James E Grizzle. Analysis of contingency tables having ordered response categories. *Journal of the American Statistical Association*, 67(337):55–63, 1972.

[20] Tong Tong Wu, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[21] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563, 2006.

[22] Ji Zhu and Trevor Hastie. Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3):427–443, 2004.