Penalized polytomous ordinal logistic regression using cumulative logits. Application to network inference of zero-inflated variables

Aurélie Deveau, Anne Gégout-Petit, Clémence Karmann

Penalized polytomous ordinal logistic regression using cumulative logits. Application to network inference of zero-inflated variables.

Aurélie Deveau[b], Anne Gégout-Petit[a] & Clémence Karmann[a]

[a] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France ;*
anne.gegout-petit@univ-lorraine.fr, clemence.karmann@inria.fr
[b] *Université de Lorraine, INRA, UMR IAM, 54280 Champenoux, France ;*
aurelie.deveau@inra.fr

Abstract. We consider the problem of variable selection when the response is ordinal, that is an ordered categorical variable. In particular, we are interested in selecting quantitative explanatory variables linked with the ordinal response variable and we want to determine which predictors are relevant. In this framework, we choose to use the polytomous ordinal logistic regression model using cumulative logits which generalizes the logistic regression. We then introduce the Lasso estimation of the regression coefficients using the Frank-Wolfe algorithm. To deal with the choice of the penalty parameter, we use the stability selection method and we develop a new method based on the knockoffs idea. This knockoffs method is general and suitable to any regression and besides, gives an order of importance of the covariates. Finally, we provide some experimental results to corroborate our method. We then present an application of this regression method for network inference of zero-inflated variables and use it in practice on real abundance data in an agronomic context.

Keywords. Multiclass logistic regression, ordinal data, Lasso, knockoffs variable selection, cumulative logit, network inference.

1 Introduction

Regression methods are really helpful to analyze dependencies between a variable, named the response, and one or several explanatory covariates. This is one of the reasons of their wide use and study in statistical analysis [23]. Many models have been introduced to respond to natural demands including the well-known linear regression for continuous response variables or logistic regression for binary response variables. 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. This is named the multinomial logistic regression and it has been particularly studied namely by [2]. In the case of ordinal data, a solution could be to forget the ordinal nature of the variable and to treat it as nominal. But this leads to poor models since the order of the values strongly matters. For such data, many authors [1, 13, 12, 21] 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 focus on a particular case of the cumulative

link models: the polytomous ordinal logistic regression using cumulative logits that uses cumulative probabilities [20, 26, 3].

Although prediction and interpretation provide major challenges in regression motivations, another important issue is to identify the influential explanatory variables, that is variable selection. Selection problems often arise in many fields including biology [27]. For example, in microarray cancer diagnosis [29], 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 variables (genes, bacteria populations...). Accordingly, the sparsity assumption (that is, a few number of relevant explanatory variables) 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 variables that have the strongest effects. Besides, when the number of variables is larger than the number of observations or when variables are highly correlated, standard regression methods become inappropriate.

Lasso penalization introduced by Tibshirani [22] offers an attractive solution to these issues. That includes a L_1 penalty in the estimation of the coefficients in order to perform variable selection by optimizing a convex criterion. The regularization resulting from Lasso penalization shrinks down to zero the coefficients of predictors that have the less effects and leads to sparse solutions and more interpretable models, making Lasso one of the most popular penalization [9, 28, 16]. However, obtaining such models sometimes involves heavy optimization issues. As far as we know, Lasso estimation for cumulative logit regression coefficients has not been performed yet. In this case, we solve the optimization thanks to the Frank-Wolfe algorithm [8].

Using Lasso also induces the critical choice of the penalty parameter which controls the number of selected variables. 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 emphasizes prediction, the validation step involving computing the prediction error and aiming at minimizing this. Furthermore, cross validation is often quite greedy and tends to overfit the data [25]. Other techniques, like StARS [11], can be adapted to a regression framework and aim at 'overselecting', that is selecting a larger set of variables which contains the relevant ones, allowing false positives. Some frameworks such as gene regulatory networks require this choice: indeed, false positives can then be eliminated by further biological experiments whereas omitted interactions cannot be recovered after that. On the contrary, we can prefer selecting a set of variables included in the set of true variables to avoid false positives ('underselection'). This constraint comes from the fact that after selection, the relevant variables have to be studied by scientists through new experiments. But new experiments are expensive or time-consuming and it would be a waste to involve false predictors. In this paper, we concentrate on the second option. Compared with our intentions, we dwell on two methods: stability selection [15] introduced by Meinshausen and Bühlmann and we develop a new intuitive and general method for variable selection, inspired from the knockoffs idea of Barber and Candès [5, 6]. The principle of the former is to estimate the probability for a variable to be in the model using bootstraps. The latter uses a matrix of knockoffs of the covariates to determine if a variable belongs to the model. Moreover, it can be performed in any regression framework. Note that none of these two methods lead to a choice of the

penalty parameter. Nevertheless, they provide an order of importance on the covariates allowing to select variables according to its target.

In this paper, we address the problem of covariate selection when the response is ordinal. Our goal is to determine which predictors are relevant and which are noise and we achieve it by using Lasso and by proposing a new method of type knockoffs to select covariates. The rest of the paper is organized as follows. In section 2, we describe the ordinal logistic regression using cumulative logits, which is a generalization of the logistic regression for an ordinal response variable. Section 3 is about Lasso estimation and inference using the Frank-Wolfe algorithm. In particular, we deal with the choice of the penalty parameter by introducing our new method for variable selection. In section 5, we present an application of our method to zero-inflated variables network inference. We also illustrate our method on both simulated (section 4) and real (section 6) data.

2 Cumulative logit model

This model has already been introduced and studied by McCullagh [13], Williams and Grizzle [26], Simon [20] or Agresti [1, 2]. Like logistic regression, it can be explained by the existence of a continuous latent variable [3] whose distribution is logistic. This regression is based on cumulative probabilities ratios.

2.1 Generalities

As written previously, the cumulative logit model is a generalization of the logistic regression for a response variable Y which takes $K > 2$ ordered categories. Let us now introduce the model. Consider we have p explanatory variables $(X_1, ..., X_p) =: X$. In the following, we denote $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{K-1}) \in \mathbb{R}^{K-1}$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p) \in \mathbb{R}^p$ and $\boldsymbol{\beta^*} = (\alpha_1, ..., \alpha_{K-1}, \beta_1, ..., \beta_p) = (\boldsymbol{\alpha}, \boldsymbol{\beta}) \in \mathbb{R}^{K-1+p}$.

We model $p^j_{\boldsymbol{\beta^*}}(x) := \mathbb{P}_{\boldsymbol{\beta^*}}(Y \leq j | X = x)$ for $j = 1, ..., K-1$ and $x \in \mathbb{R}^p$, by:

$$\text{logit } p^j_{\boldsymbol{\beta^*}}(x) = \alpha_j + \beta_1 x_1 + ... + \beta_p x_p, \tag{1}$$

i.e.,

$$p^j_{\boldsymbol{\beta^*}}(x) = \frac{\exp(\alpha_j + \beta_1 x_1 + ... + \beta_p x_p)}{1 + \exp(\alpha_j + \beta_1 x_1 + ... + \beta_p x_p)}.$$

Notice that the vector $\boldsymbol{\beta}$ of coefficients does not depend on the level j, assuming an identical effect of the predictors for each cumulative probability. Actually, assume we allow separate effects, that is replacing $\boldsymbol{\beta}$ by $\boldsymbol{\beta^j}$. This would imply nonparallel curves for different logits if x takes spread enough values and would digress the proper order among the cumulative probabilities [12, 2]. Although that model of separate effects can hold over a narrow range of explanatory variables values, it is more careful to avoid the model of separate effects, especially without much informations on the explanatory variables. That is why we focus on the simpler model of similar effects (1).

As $p^K_{\boldsymbol{\beta^*}}(x) = 1$, this model includes $K - 1 + p$ coefficients to be estimated ($K - 1$ coefficients for the vector $\boldsymbol{\alpha}$ and p coefficients for the vector $\boldsymbol{\beta}$).

Suppose that $(Y^i, X^i_1, \dots X^i_p)_{1 \leq i \leq p}$ are n independent and identically distributed vectors. Denote $X^{*i}_j = (0, ..., 1, ..., 0, X^i_1, ..., X^i_p)$ where 1 is at the j^{th} position, it is now possible to define:

- the log-likelihood:

$$L(\boldsymbol{\beta}^*) = \sum_{j=1}^{K} \sum_{i/Y^i=j} \log\Big[\frac{\exp(-\boldsymbol{\beta}^* X_{j-1}^{*i}) - \exp(-\boldsymbol{\beta}^* X_j^{*i})}{(1+\exp(-\boldsymbol{\beta}^* X_j^{*i}))(1+\exp(-\boldsymbol{\beta}^* X_{j-1}^{*i}))}\Big], \tag{2}$$

- the gradient of the log-likelihood:

$$\begin{aligned}\nabla L(\boldsymbol{\beta}^*) = &\sum_{j=1}^{K} \sum_{i/Y^i=j} \Big[\frac{X_j^{*i}\exp(\boldsymbol{\beta}^* X_j^{*i}) - X_{j-1}^{*i}\exp(\boldsymbol{\beta}^* X_{j-1}^{*i})}{\exp(\boldsymbol{\beta}^* X_j^{*i}) - \exp(\boldsymbol{\beta}^* X_{j-1}^{*i})} - \\ &\frac{X_j^{*i}}{1+\exp(-\boldsymbol{\beta}^* X_j^{*i})} - \frac{X_{j-1}^{*i}}{1+\exp(-\boldsymbol{\beta}^* X_{j-1}^{*i})}\Big].\end{aligned} \tag{3}$$

The gradient (3) will be useful at the optimization step (see subsection 3.1).

2.2 Coefficients interpretation

In the same way as for the logistic regression, we can consider odds and odds ratios for the cumulative logit model.

For $X = x$ fixed, for all $j \in \{1, ..., K\}$, the odds are defined by:

$$\text{odds}_j(x) = \exp(\text{logit}(p^j_{\boldsymbol{\beta}^*}(x))) = \frac{\mathbb{P}(Y \leq j | X = x)}{1 - \mathbb{P}(Y \leq j | X = x)} = \exp(\alpha_j + \beta_1 x_1 + ... + \beta_p x_p).$$

This ratio measures the tendancy for Y to be greater or smaller than j given $X = x$.

We consider also cumulative odds ratios that is odds ratios of cumulative probabilities which are defined by: $\dfrac{\text{odds}_j(x)}{\text{odds}_j(\tilde{x})}$ for all $x, \tilde{x} \in \mathbb{R}^p$. In the case where $\tilde{x} = x_i^{+z} := (x_1, ..., x_i + z, ..., x_p)$, then for all $j \in \{1, ..., K\}$, $\dfrac{\text{odds}_j(x)}{\text{odds}_j(x_i^{+z})} = \exp(-\beta_i z)$.

Observe that these cumulative odds ratios are the same for any level j. This comes from the assumption of identical effects of the covariates and accordingly, this model is often called proportional odds model [13].

Some references prefer the parametrization logit $p^j_{\boldsymbol{\beta}^*}(x) = \alpha_j - \beta_1 x_1 - ... - \beta_p x_p$ instead of the one used in (1). This parametrization only affects the interpretation of odds ratios, the minus sign corresponding to the usual interpretation [12].

Our purpose is to select the relevant variables, which means the variables X_i such that the regression coefficient β_i is non-zero. Notice that β_i also measures the conditional dependence between Y and X_i given $(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$. That is why we are especially interested in the nullity of the coefficients $\boldsymbol{\beta}$. Moreover, we make sparsity assumption, that is only a few β_i are non-null and a relatively small number of predictors play an important role. This sparsity assumption is convenient for scientists to restrict their studies to a smaller subset or predictors, namely in high dimensional settings. Instead of testing the nullity of each coefficient β_i, we add a L_1-penalization on the coefficients $\boldsymbol{\beta}$ in the log-likelihood to estimate the coefficients.

3 Estimation, inference

Without sparsity assumption, coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are usually estimated by maximization of the log-likelihood L using Fisher scoring algorithms [13, 24].

3.1 Lasso estimation of the coefficients $\boldsymbol{\beta}$

To ensure sparsity assumption, we penalize the log-likelihood L on the coefficients vector $\boldsymbol{\beta}$. For that, we need to solve the following optimization problem:

$$\underset{\substack{\boldsymbol{\alpha}\in A \\ \boldsymbol{\beta}\in\mathbb{R}^p}}{\text{argmax}} \left\{ L(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \lambda ||\boldsymbol{\beta}||_1 \right\}, \tag{4}$$

where A is the convex set of $\mathbb{R}^{K-1}$ defined by: $A := \{(\alpha_1, ..., \alpha_{K-1}) \in \mathbb{R}^{K-1} \,/\, \alpha_1 < \ \ldots \ < \alpha_{K-1}\}$, , $||.||_1$ denote the L_1 norm, that is $||\boldsymbol{\beta}||_1 = \sum_{i=1}^{p} |\beta_i|$ and $\lambda > 0$ is the penalty parameter. Solving (4) is equivalent, by lagrangian duality, to solve:

$$\underset{\substack{\boldsymbol{\alpha}\in A \\ \boldsymbol{\beta}\in B_\tau}}{\text{argmax}} \ L(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{5}$$

where B_τ is the following convex set of $\mathbb{R}^p$: $B_\tau := \{(\beta_1, ..., \beta_p) \in \mathbb{R}^p \,/\, ||\boldsymbol{\beta}||_1 \leq \tau\}$. There is a one-to-one correspondance between $\lambda > 0$ and $\tau > 0$.

We solve this optimization thanks to the Frank-Wolfe algorithm [8] which is described in further detail below. The idea is to replace the target function to be minimized by a linear approximation. In our case, we use the gradient provided in (3) to approximate the log-likelihood L given in (2).

Frank-Wolfe algorithm

– Step 1: Start with an initial value $\boldsymbol{\beta_0^*} = (\boldsymbol{\alpha_0}, \boldsymbol{\beta_0})$.

– Step 2: At each iteration k,

- Solve $s_k = (s_{\boldsymbol{\alpha}}, s_{\boldsymbol{\beta}}) \in \underset{\substack{s_{\boldsymbol{\alpha}}\in A \\ s_{\boldsymbol{\beta}}\in B_\tau}}{\text{argmin}} \ (-\nabla L(\boldsymbol{\beta_k^*}))' \begin{pmatrix} s_{\boldsymbol{\alpha}} \\ s_{\boldsymbol{\beta}} \end{pmatrix}$.
- The new value is $\boldsymbol{\beta_{k+1}^*} = (1 - \gamma_k)\boldsymbol{\beta_k^*} + \gamma_k s_k$, where $\gamma_k = \frac{2}{k+1}$.

– Step 3: Iterate Step 2 until convergence.

Notice that we use the convexity of the sets A and B_τ in the step 3. We can split the optimization problem (step 2) into two different optimization problems. The first one is relative to the vector $\boldsymbol{\alpha}$:

$$s_{\boldsymbol{\alpha}} \in \underset{s\in A}{\text{argmin}} \ (-\nabla L(\boldsymbol{\beta_k^*}))'_{|\boldsymbol{\alpha}} s,$$

and turns out to be a linear optimization under constraints. The second one concerns the vector $\boldsymbol{\beta}$:

$$s_{\boldsymbol{\beta}} \in \underset{s\in B_\tau}{\text{argmin}} \ (-\nabla L(\boldsymbol{\beta_k^*}))'_{|\boldsymbol{\beta}} s. \tag{6}$$

The optimization part on the $\boldsymbol{\beta}$ is equivalent to solve $-\tau \underset{s\in B_1}{\text{argmax}} (-\nabla L(\boldsymbol{\beta_k^*}))'_{|\boldsymbol{\beta}} s$. This yields to choose the coordinate $i_k \in \{1, \dots, p\}$ such that $|(-\nabla L(\boldsymbol{\beta_k^*})_{|\boldsymbol{\beta}})_{i_k}|$ maximizes the absolute value of the gradient (restricted to $\boldsymbol{\beta}$). Note that i_k is not necessarily unique ; in this case, we choose the first coordinate which verifies this. Then, we obtain the solution: $s_{\boldsymbol{\beta}} = -\tau \text{ sign}(-\nabla L(\boldsymbol{\beta_k^*})_{|\boldsymbol{\beta}})_{i_k} \, e_{i_k}$ for the optimization problem (6) where e_i denotes the i^{th} vector of the canonical basis.

3.2 Penalty parameter and variable selection

Unfortunately, all penalization methods require the choice of the (positive) penalty parameter, also referred as tuning or regularization parameter. We then need to tune the penalty parameter τ (involved in the constraints of the optimization problem (5)) which controls the number of selected variables: the closer to 0 τ is, the fewer the selected predictors are. Remind that our goal is to select only relevant variables and as a consequence, we would like to avoid false positives. Despite this purpose, our algorithm allows to choose its own threshold.

Two methods suit with regard to our problems and goals: the first one is the stability selection proposed by Meinshausen and Buhlmann [15] and we propose a new one, inspired from the knockoffs process used by Barber and Candès [5]. Actually, these two methods do not lead to a choice of the penalty parameter τ but they put the explanatory variables in order from the most relevant to the less, allowing the user to make its own choice. Furthermore, they both suit any regression.

3.2.1 Stability selection

The principle is to estimate the probability for a variable to be in the model in order to determine which variables are the most relevant.

Let us consider a set T of values for the penalty parameter τ. For each $\tau \in T$, the idea is to estimate the probability $p_i(\tau)$ for each variable i to be in the model. For that, we perform the penalized regression on B sets of n observations obtained by bootstrap. The estimated probability $\hat{p}_i(\tau)$ is the proportion of selection of the variable X_i among the B bootstrapped regressions for this fixed value of τ.

Usually, variable selection involves choosing an estimated model among $\{\hat{S}^\tau, \ \tau \in T\}$ where $\hat{S}^\tau$ is the estimated model relative to the fixed parameter τ, that is: $\hat{S}^\tau = \{i \in \{1, ..., p\} \ : \ \hat{\beta}_i(\tau) \neq 0\}$ where $\hat{\boldsymbol{\beta}}(\tau)$ denotes the estimated coefficients of the τ-penalized regression of Y on X. Instead of choosing one of these models, we choose here the model $\hat{S} := \{i \in \{1, ..., p\} \ : \ \max_{\tau \in T} \hat{p}_i(\tau) \geq p_{thr}\}$ for a fixed threshold p_{thr} (see [15] for more details). Notice that covariates are ordered according to $\max_{\tau \in T} \hat{p}_i(\tau)$.

The threshold value p_{thr} has a very small influence, which is very convenient. Indeed, results tend to be similar for a wide range of values of p_{thr} unlike the penalty parameter τ. However, Meinshausen and Bühlmann [15] still provide a procedure to choose the cut-off p_{thr} and the regularization region T. Under some simplifying assumptions, this procedure yields a bound for the expected number of false positives (selected by stability selection). Nonetheless, this bound depends on an unknown quantity (T-dependent) and

the assumptions can be a bit too strong. That is why we prefer using an arbitrary threshold (see section 4).

3.2.2 Revisited knockoffs

With a little abuse of notation, let X denote the $n \times p$ matrix of the n observations of the vector $(X_1, \ldots, X_p)$, called the design matrix. The principle, given in [5], is to use a matrix $\tilde{X}$ of knockoffs (of the variables X_i) whose structure is similar to X but independent from Y. The goal is to determine if a variable X_i is relevant by studying if it enters the model before its knockoff $\tilde{X}_i$. Indeed, as the knockoff is independent from Y, if a variable enters the model after its knockoff, we can rightfully suspect that this variable does not belong to the model.

We mainly differ from the method proposed by [5] in the construction of the knockoffs. We construct our knockoff matrix $\tilde{X}$ by swapping the n rows of the design matrix X. This way, the correlations between the knockoffs remain the same as the original variables but the knockoffs are not linked to the response Y. Let us note $\hat{\boldsymbol{\beta}}(\tau)$ the estimated coefficients of the τ-penalized regression of Y on the augmented matrix $[X, \tilde{X}]$. For each variable $i \in \{1, \ldots, p, p+1, \ldots, 2p\}$ of the augmented design, we consider $T_i := \inf\{\tau > 0,\ \hat{\beta}_i(\tau) \neq 0\}$. At this stage, we hope that T_i is small for the relevant variable, that is for $i \in \{1, \ldots, p\}$ such that $\beta_i \neq 0$ and large for the variables $i \in \{p+1, \ldots, 2p\}$ or for the variables $i \in \{1, \ldots, p\}$ 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 $i \in \{1, ..., p\}$, $W_i := T_i \wedge \tilde{T}_i \times \begin{cases} (+1) & \text{if } T_i < \tilde{T}_i \\ (-1) & \text{if } T_i \geq \tilde{T}_i \end{cases}$.

A negative value for W_i means that the variable X_i enters the model after its knockoff and we eliminate it. On the contrary, a positive value for W_i means that the variable X_i enters the model before its knockoff and is more likely to belong to the model. But variables X_i whose statistic W_i is positive are not necessarily relevant: we hope that W_i is small for most of relevant variables and large for the others. Thus, we are interested in the smallest positive values of the p-vector of statistics W which moreover indicates that the variable enters the model early. This suppose to define a threshold s for W_i under which we will keep the corresponding variables in the model. On the whole, we will choose the model $\hat{S}$ such that:

$$\hat{S} := \{X_i : 0 < W_i \leq s\}.$$

About the threshold s, Barber and Candès [5] provide 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 our case. We make the assumption that there is a breakdown in the distributions between the W_i corresponding to the X_i in 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 $p = 20$ covariates, only the four first ones were linked to Y. In the knockoffs procedure, variables $1, 2, 3, 4, 5, 7, 9, 10, 11, 12, 14$ had positive values for W_i. We can clearly observe a breakdown between the values of the four first ones and the others.

Consequently, we chose to use two change detection methods: the method proposed by Auger and Lawrence [4, 17, 18] and the CUSUM method for mean change detection. Let $W_{(i)},\ i = 1, ..., w$ 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 $j = 1, \ldots w-1$. Remark that w, the number of positive statistics W_i, is random ($w = 11$ on figure 1). Those

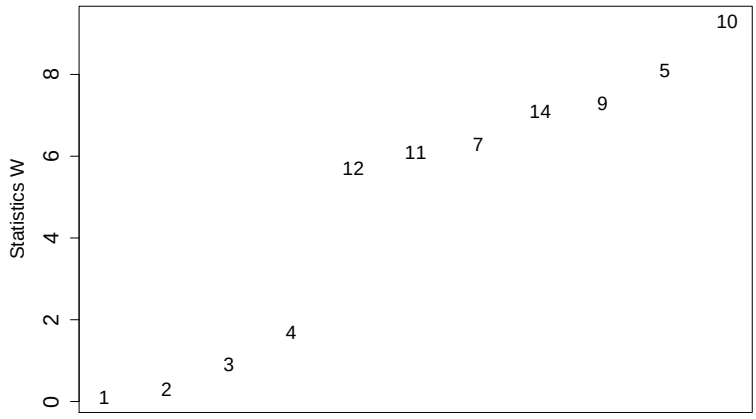

Figure 1: Example of sorted positive statistics W_i. Only variables X_1, X_2, X_3 and X_4 belong to the model (in this case, $\beta = (8, 6, 4, 2, 0, \ldots, 0)$).

two methods applied to the sorted gaps $e_{(j)}$ provide us two thresholds and we choose the minimum of these two thresholds. Let us name this threshold 'min threshold' for the sake of simplicity. It is though possible to choose its own threshold through displaying the sorted positive statistics W_i.

4 Simulation studies

We now describe experimental results in order to study the efficiency of our methods.

For both stability selection and revisited knockoffs, we center and reduce the explanatory variables so that the variances do not have influence on the estimated regression coefficients $\boldsymbol{\beta}$. We have performed different simulations with various distribution for the covariates. In the following, we present the results for $p = 50$ covariates, $K = 3$ ordered modalities for the response Y, $n = 100$ and 200 samples, $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$ and $\boldsymbol{\alpha}$ properly chosen (so that the response Y takes enough values in each of its 3 modalities). Covariates X are simulated as gaussian such that X_i and X_j are independent conditionally on the other X_k with probability $1 - \rho$, $\rho = 0.6$. The vector X of covariates is simulated with the R function `huge.generator`, for a random graph structure. Most of non-null correlations are between -0.3 and 0.3 and non-null partial correlations are about -0.13.

We have also used other kind of distribution for the vector of covariates. Results are given in Appendix A.2.

4.1 Stability selection

Settings. About stability selection, we tuned $B = 100$ bootstraps and the set of values for τ is $T = \{0.1, 0.4, 0.7, ..., 3.7\}$. The tuning of the set T suits to many problems since the covariates are centered and reduced. The set T can be changed but we would like to

point out that too large values for the penalty parameter τ lead to the full model (for all i, $\hat{p}_i(\tau) = 1$). We recommend therefore to be careful. However, the threshold p_{thr} can be modified as needed.

We represent the results by boxplots (figure 2) and ROC curves (figure 3). Boxplots are obtained on 50 repetitions of $n = 100$ and 200 samples of $p = 50$ variables and correspond to the estimated probabilities, that is $\max_{\tau \in T} \hat{p}_i(\tau)$, for each variable X_i. ROC curves exhibit the average false positive rates (FPR) and true positive rates (TPR) on 50 repetitions after thresholding the estimated probabilities for $p_{thr} \in \{0.1, 0.15, 0.2, \ldots, 1\}$.

Results and comments. Figures 2 and 3 show that the method is efficient. ROC curves (figure 3) illustrate that the procedure is sensible and specific and that a wide range of values for p_{thr} leads to similar results. Notice that the scale for the TPR starts at 0.8 and the scale for the FPR ends at 0.25. This means that the average TPR is around 0.8 for $p_{thr} = 1$ and the average FPR is around 0.14 ($n = 200$) and 0.27 ($n = 100$) for $p_{thr} = 0.1$. Boxplots of figure 2 show the distribution of the probabilities for each covariate to be in the model for 50 independent simulations. The difference of the distribution between the relevant covariates (from 1 to 4) and the other ones is very clear. The first three covariates are almost always detected for $p_{thr} = 1$. The fourth one (corresponding to a lower regression coefficient) is detected with a rate of 75 % if $p_{thr} = 0.75$ if $n = 200$ (resp. $p_{thr} = 0.55$ if $n = 100$).

The different simulations we have performed showed that the efficiency of these methods does not depend on the distribution of the predictors (see Appendix A.2 for more results).

4.2 Revisited knockoffs

Settings. We calculate the statistics T_i, $i = 1, \ldots, p, p+1, \ldots, 2p$ ($p = 50$) as $T_i := \min\big(\inf\{\tau \in \mathcal{T}, \hat{\beta}_i(\tau) \neq 0\} \cup \{1000\}\big)$ where $\mathcal{T} := \{0.1, 0.3, 0.5, \ldots, 10.1\}$. We choose 1000 as an arbitrary value which means that if $T_i = 1000$, X_i (or $\tilde{X}_{i-50}$ if $i > 50$) did not enter the model before $\tau = 10.1$. In the same way as for stability selection, the tuning of $\mathcal{T}$ does not matter a lot since the covariates are centered and reduced.

We represent the results by boxplots (figure 4) and detection rates (figure 5). Both are obtained on 100 repetitions of $n = 100$ and 200 samples of $p = 50$ variables. The boxplots display 'appearance' ranks of each variable which is a kind of relevance rank for each variable. This rank is calculated using the statistics W_i. These statistics sort in fact the covariates: covariates with the smallest positive value of W get the first rank, those with the smallest negative value get the last rank. So, the variable with the rank 1 is the most relevant for the model and so on. Note that covariates can have the same rank, thus the last rank does not have to be $p = 50$. These ranks provide in fact a sorting of covariates. The figure of covariates detection rate (figure 5) present the detection rates of each covariate after thresholding the statistics W_i, $i = 1, \ldots, 50$. The threshold is the min threshold introduced in subsection 3.2.2.

Results and comments. Again Figure 4 show the good efficiency of the procedure. As expected, the boxplots indicate that X_1, X_2, X_3 and X_4 enter the model in this order. The

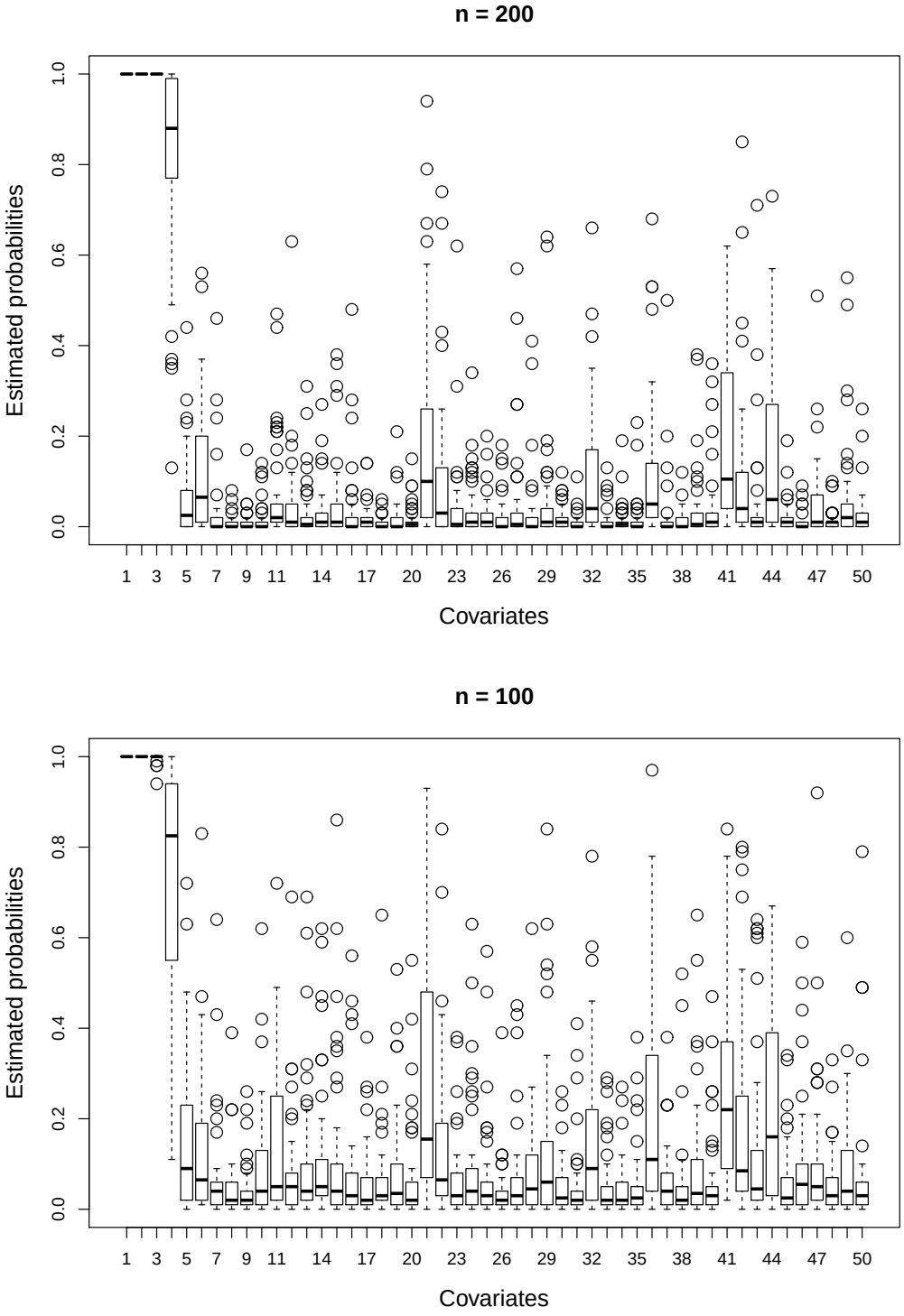

Figure 2: Boxplots of estimated probabilities $\max_{\tau \in T} \hat{p}_i(\tau)$ for each covariate. Regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 50 repetitions constituted by $n = 100$ and 200 samples of $p = 50$ variables with stability selection method.

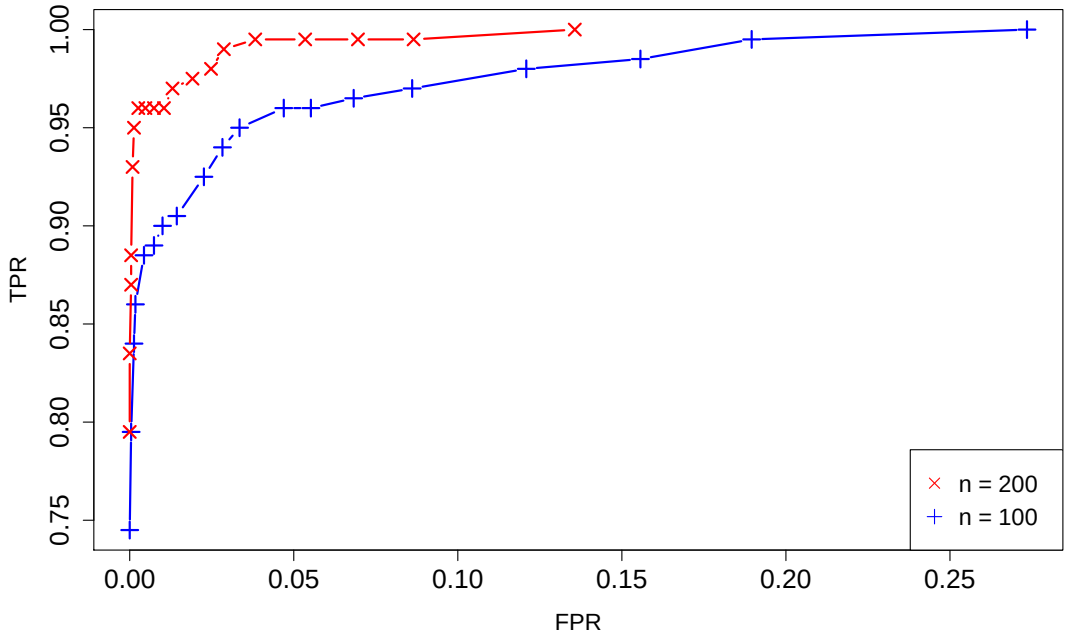

Figure 3: ROC curves obtained by thresholding for $p_{thr} \in \{0.1, 0.15, 0.2, \ldots, 1\}$ with stability selection method. TPR and FPR refer to average TPR and FPR obtained on 50 repetitions of $n = 100$ and 200 samples of $p = 50$ variables.

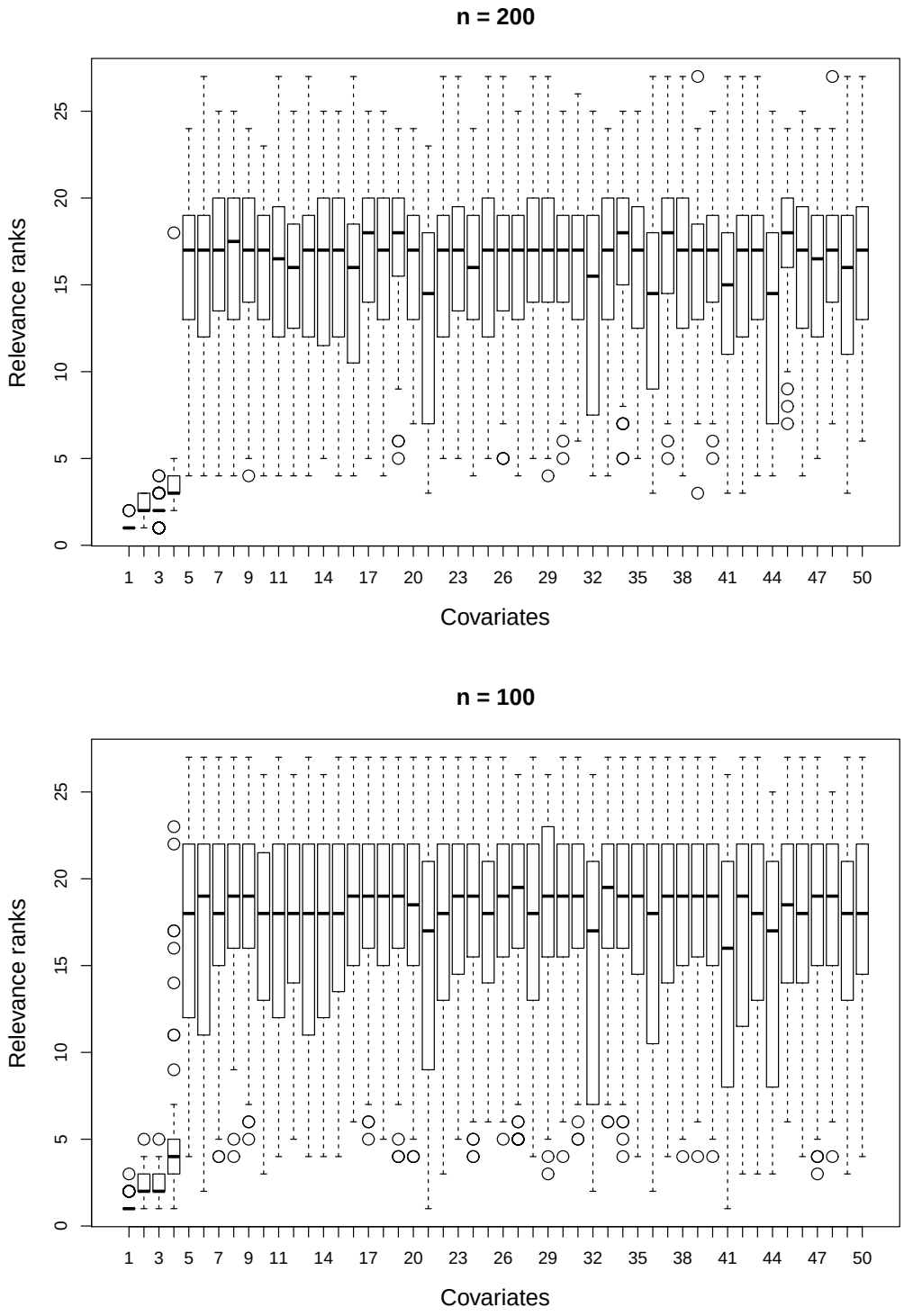

Figure 4: Boxplots of 'appearance'/relevance ranks for each variable. Regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 100 repetitions constituted by $n = 100$ and 200 samples of $p = 50$ variables with revisited knockoffs method.

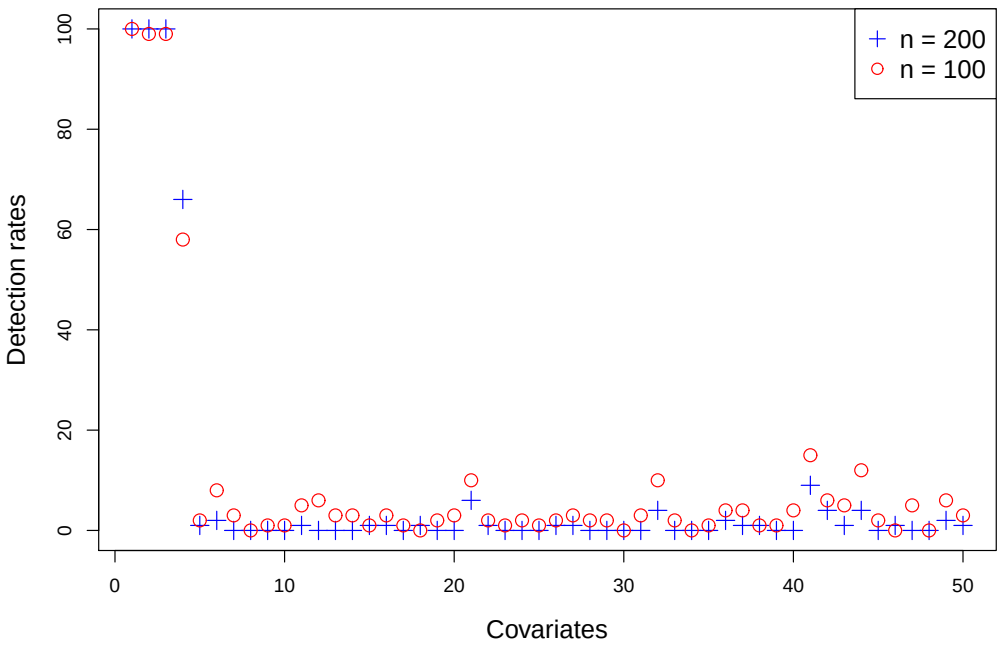

Figure 5: Detection rates on 100 repetitions after applying revisited knockoffs method with the min threshold (see subsection 3.2.2). Regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$.

difference between the relevant covariates and the others is clear. The detection rates after thresholding given in Figure 5 illustrate also this phenomenon; the third first variable are almost always detected, the second is detected in about 60 % of the simulations whatever $n = 100$ or 200. The rate of detection of the irrelevant covariates is very low. For both stability selection and revisited knockoffs, we observe that some covariates, namely X_{21}, X_{32}, X_{41} and X_{44}, are more often detected than others. It is probably due to the dependance structure of X. In particular, these covariates are dependent to X_1, X_2, X_3 and X_4 conditionally on the others.

We also performed simulations with independent covariates and show results in the Appendix A.1.

5 Application to zero-inflated networks inference

Many authors used regressions to address dependency graphs inference issues. In Gaussian graphical models (GGM), the nullity of a coefficient of the regression is equivalent to the nullity of the corresponding partial correlation and thus to the corresponding conditional independence. Meinshausen and Bühlmann [14] used linear regressions in this gaussian framework. Even if the equivalence is not clear in another context, the Ising model provides similar properties in that the nullity of a coefficient of the logistic regression is equivalent of the conditional independence of the corresponding variables. [10] and [19] inferred binary graphs with logistic regressions using the Ising model. That is why, we would like to apply

our model of regression to networks inference. Unfortunately, the partial distribution (1) is not consistent with any joint distribution [21]. Out of curiosity, we tried to apply this regression in a network framework though. Besides, simulating different kind of data is also interesting with respect to robustness. In this section, we focus on networks of zero-inflated variables.

5.1 Data simulation

We aim at simulating data sets being like abundance data (naturally zero-inflated). For that, we first simulate a gaussian p-vector X whose graph structure is a chain, that is X_j and X_k are dependent conditionally to the rest of the variables iff $|j-k| \leq 1$. As previously, we use the R function `huge.generator` for a band graph structure. By default, $\mathbb{E}(X_i) = 0$ and $\text{var}(X_i) = 1$ for all $i \in \{1, \dots, p\}$.

After that, we change the means and variances so that $\mathbb{P}(X_i \geq 0)$ is close to 1 for $i = 1, \dots, p$. Means $\mu_i,\ i = 1, \dots, p$ are chosen between 1 and 100: 50% are chosen uniformly between 1 and 5, 25% between 6 and 10, 15% between 11 and 50 and 10% between 51 and 100. Variances σ_i^2 depend on the means in this manner: $\sigma_i = 1.1\frac{\mu_i}{2}\mathbb{1}_{\mu_i \leq 5} + 0.9\frac{\mu_i}{2}\mathbb{1}_{5<\mu_i\leq 10} + 0.5\frac{\mu_i}{2}\mathbb{1}_{10<\mu_i\leq 50} + 0.3\frac{\mu_i}{2}\mathbb{1}_{50<\mu_i}$.

At last, we add a zero-inflation by multiplying X by a p-vector Ber of Bernoullis, depending on vector X. We simulate Ber as following: for all $i \in \{1, \dots, p\}$, $Ber_i \sim \mathcal{B}(\pi(X_i))$ where $\pi : \mathbb{R}^+ \to [0,1],\ x \mapsto \dfrac{\exp(a+bx)}{1+\exp(a+bx)}$ where $a = \log(10^{-2})$ and $b = 3$. Thus, the closer to 0 x is, the closer to 0 $\pi(x)$ is and then, the associated Bernoulli is more likely to be 0. The observations are then $Z = Ber \cdot X$. This way, Z_i is more likely to be 0 if X_i is small. Note that we work with n observations of the variable Z to infer the network of the latent variable X.

5.2 Application of polytomous regression to the simulated data

The goal now is to retrieve the links between the variables $X_i,\ i = 1, ..., p$, given theoretically by the precision matrix Σ^{-1}, with the observed variables $Z_i,\ i = 1, ..., p$. In this case, the underlying graph structure is a chain, noted by $X_1 - X_2 - \ldots - X_p$ where $X_i - X_j$ represents an edge between X_i and X_j.

Our approach is quite conventional and involves performing penalized polytomous regression (see (1)) of each variable (converted to classes) on the remaining unaltered variables, and then using the sparsity pattern of the regression vector to infer the underlying neighborhood structure. This procedure leads to two graphs: the graph 'or' and the graph 'and'. Let us denote $\hat{\beta}_i^j,\ i \neq j$ the estimated coefficients of the regression of X_j (considered as the response variable) on the covariate X_i. The graph 'and' contains an edge between X_j and X_k if $\hat{\beta}_k^j \neq 0$ and $\hat{\beta}_j^k \neq 0$ whereas the graph 'or' contains an edge between these two variables only if one of the estimated coefficients is non-null. We only build the graph 'and' since it contains less false positive edges, making specificity higher.

But before performing regressions, we need to transform the response variable relative to each regression so that it is ordinal. For convenience, let us denote R the response. R

comprises some zeros and the rest is continuous. Consequently, we break down the non-null values of R into classes determined by quantiles of R, so that classes are balanced. The number of modalities that we choose depends on the number of non-null values of R. We choose the number of modalities K as: $K = \left[\frac{\#\{i \in \{1,\dots,n\} \ / \ R^i \neq 0\}}{20}\right] + 1$, where [.] denotes the floor function. In our simulations, K goes from 2 to 11 depending on the zero-inflation. When $K = 2$, Y only discriminates if the original variables equals 0 or not.

Finally, we perform regressions with revisited knockoffs methods for variable selection.

5.3 Results

We choose $p = 50$ and $n = 200$. Correlations between linked variables are set about -0.45 and partial correlations about -0.38. The zero-inflation represents about 12%, varying from 0 to 75% according to the variables.

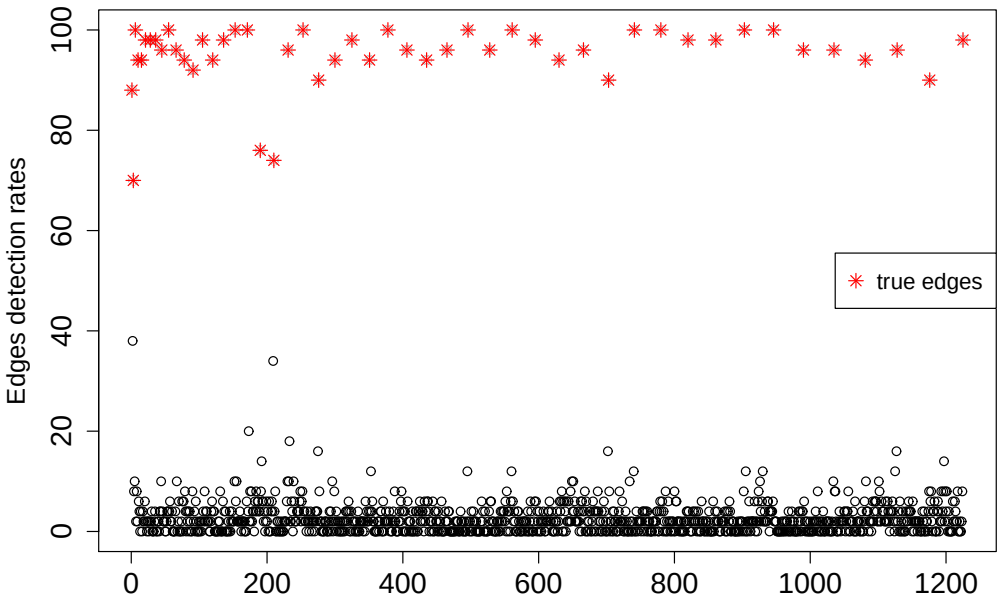

Figure 6: Edges detection rates on 50 repetitions after applying revisited knockoffs method with the min threshold (see subsection 3.2.2). Circles and stars represent respectively false and true edges.

Figure 6 shows edges detection rates with revisited knockoffs and the min threshold. True edges are the most detected. Almost all of them are detected more than 90% whereas almost false edges are detected less than 10% of the simulations. Moreover, there is a distinct gap between true and false edges.

6 Real data

We conclude numerical experiments by applying our revisited knockoffs method on real data. The data set has been produced by a team of researchers from INRA (Institut National de la Recherche en Agronomie) and Frankurt University and consists of the bacterial communities of 82 samples of truffles analyzed by high throughput sequencing (unpublished data Splivallo, Vahdatzadeh & Deveau). A total of 242 operational taxonomic units (OTUs) with varying relative abundance were obtained across the different samples. Each OTU is made of groups of microorganisms that have very high DNA sequence similarities - i.e. genetically closely related microorganisms.

The goal is to detect potential interactions between these OTUs in order to identify potential networks of interactions occurring in situ between microorganisms among the thousands of potential interactions that could exist but that cannot be experimentally measured. Functional interactions can then be experimentally analyzed on small sets of microorganisms [7]. This data set is particular in that it contains a lot of zeros: many OTUs are indeed present only in a few samples and the data set requires a prior sorting. For that, we eliminated OTUs present in less than $\frac{82}{3} \approx 27$ samples. The final data set contains 82 samples of 62 OTUs. However, it remains a lot of zeros so that GGM are not appropriate to analyze the link between the variables. We then apply our penalyzed regression and revisited knockoffs method.

Notice that our knockoffs procedure is random since the matrix of knockoffs is obtained by swapping randomly the rows of the design matrix X. This randomness allows us to weight edges and to sort them. In this work, we choose to repeat 80 times the revisited knockoffs procedure as we did in the previous section on simulated network data and select edges which are detected more than 57 times. We then weight each of these edges according to the number of times it has been detected.

Results and comments. Our knockoffs procedure produced a network containing a total of 50 edges between 44 OTUs, whose the main cluster contains 33 OTUs and 42 edges. Figure 7 displays this main cluster. The network is organized in two clusters linked together by 3 OTUs. Cluster A is made of OTUs that tend to co-variate and regroups OTUs corresponding to closely related bacteria in terms of functional abilities, thus being very likely to naturally co-occur and interact within truffles. OTUs from cluster B tend also to co-occur while OTU 1 and OTU 2 that connect the two clusters show a tendency for exclusion patterns. A similar negative link between OTU1 and 2 is also observed in other data sets (Splivallo et al. in prep), supporting the validity of the predictions made by the knockoffs procedure. Interestingly, a third of the OTUs highlighted in the network are available for culture in the laboratory at INRA and experimental tests could be done in the future to validate the predictions of the model.

7 Discussion

In this paper, we proposed new methods to infer the logistic regression with cumulative logits, a regression for ordinal response variable. We gave an algorithm to select covariates and estimate the regression coefficients by maximizing a penalized version of the likeli-

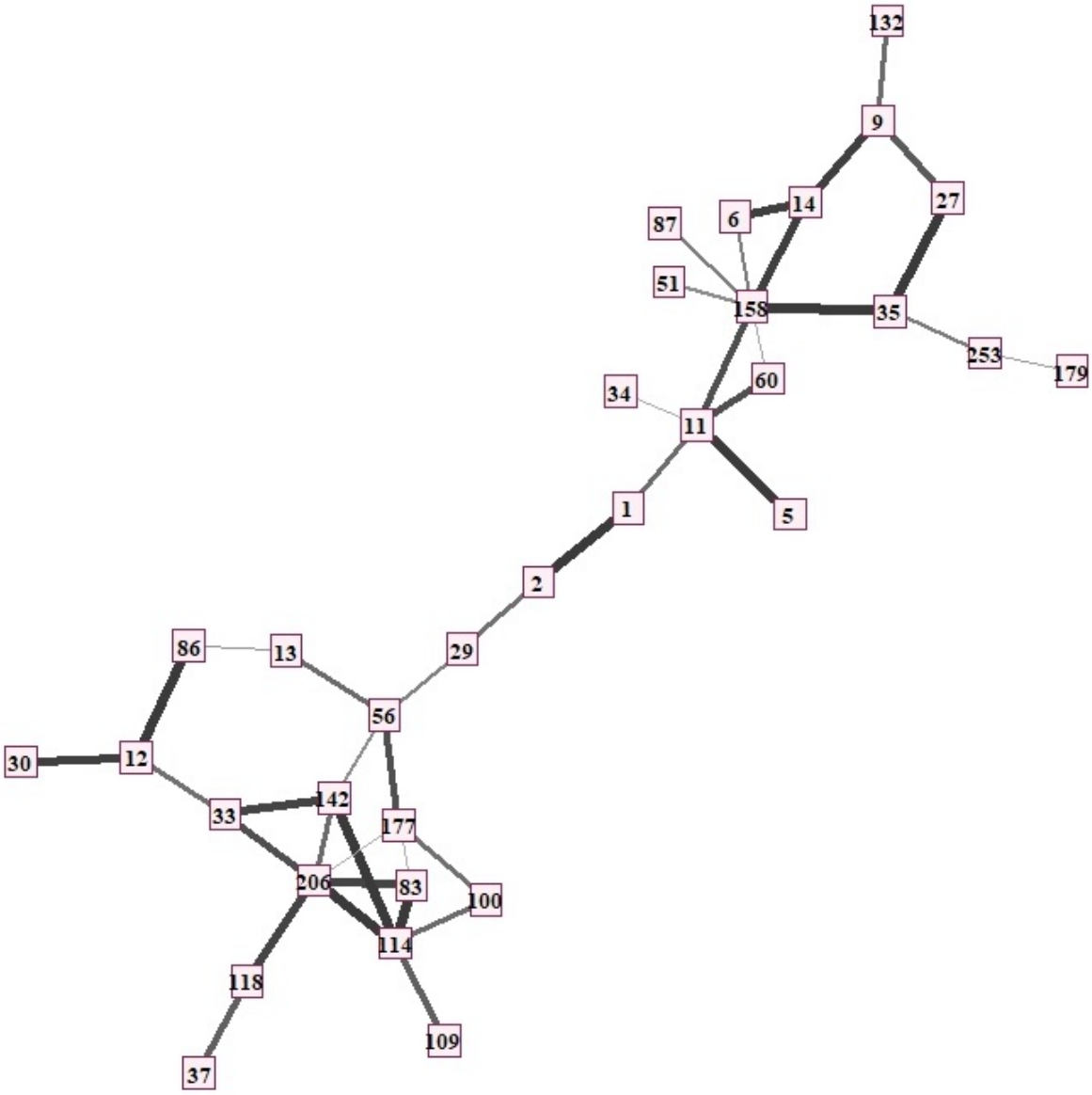

Figure 7: Main cluster of truffle OTUs network obtained after applying 80 times the knockoffs method. Edges are represented if they are detected more than 57 times and they are weighted according to their number of detection.

hood which did not already exist to our knowledge. For this, we also developed a new method based on the knockoffs idea from [5] to perform variable selection, that is really intuitive in the manner to build the knockoffs and to select the covariates. Moreover this method of selection is not specific to the polytomic regression and is also suitable for other kind of regressions and can be used in other contexts. We have seen that these penalized regression and knockoffs procedure turn out to be very pertinent and efficient as the many and diverse simulations above and in the appendix exemplify. We choose not to try to establish theoretical guarantees because of the cumulative logit regression model strong complexity. However, we are aware of the importance to do it and this could be subject to a future substantial theoretical work.

The developed regression method and knockoffs procedure allow us to use penalized polytomic regression to infer network in a context where Gaussian graphical models are not adapted because of the presence of zeros. We used it to infer a network of zeros-inflated covariates first in a simulation case and next on a set of abundance variables in an ecological context of interactions between microorganisms (truffles in our case). Practitioners were very satisfied with the results. Here again, even if our method works well to infer networks, the theoretical questions remain important because the cumulative logit regression model is not consistent with any joint distribution. That opens up prospects for graph inference when data are zero-inflated.

Acknowledgements and Funding

We wish to particularly thank Stéphane Chrétien for his precious help and advices and Aurélie Muller-Gueudin for her suggestions and comments which greatly improved this paper. We also acknowledge Richard Splivallo and Maryam Vahdatzadeh (Institute of Molecular Biosciences, Goethe University Frankfurt) for having provided us with truffles data. Their work was supported by PEPS-Mirabelle Trufinet project and INRA metaprogramm MEM POPART project.

References

[1] A. Agresti, *Analysis of ordinal categorical data*, 2nd ed., Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, 2010. MR 2742515

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

[3] J. Anderson and P. Philips, *Regression, discrimination and measurement models for ordered categorical variables*, Applied statistics (1981), pp. 22–31.

[4] I.E. Auger and C.E. Lawrence, *Algorithms for the optimal identification of segment neighborhoods*, Bull. Math. Biol. 51 (1989), pp. 39–54. MR 978902

[5] R.F. Barber and E.J. Candès, *Controlling the false discovery rate via knockoffs*, Ann. Statist. 43 (2015), pp. 2055–2085. MR 3375876

[6] R.F. Barber and E.J. Candes, *A knockoff filter for high-dimensional selective inference*, arXiv preprint arXiv:1602.03574 (2016).

[7] K. Faust and J. Raes, *Microbial interactions: from networks to models*, Nature Reviews Microbiology 10 (2012), p. 538.

[8] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Res. Logist. Quart. 3 (1956), pp. 95–110. MR 0089102

[9] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, CRC press, 2015.

[10] K. Koh, S.J. Kim, and S. Boyd, *An interior-point method for large-scale l_1-regularized logistic regression*, J. Mach. Learn. Res. 8 (2007), pp. 1519–1555. MR 2332440

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

[12] I. Liu and A. Agresti, *The analysis of ordered categorical data: an overview and a survey of recent developments*, Test 14 (2005), pp. 1–73. With discussion and a rejoinder by the authors. MR 2203424

[13] P. McCullagh, *Regression models for ordinal data*, J. Roy. Statist. Soc. Ser. B 42 (1980), pp. 109–142. MR 583347

[14] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the lasso*, Ann. Statist. 34 (2006), pp. 1436–1462. MR 2278363

[15] N. Meinshausen and P. Bühlmann, *Stability selection*, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (2010), pp. 417–473. MR 2758523

[16] M.Y. Park and T. Hastie, *L1-regularization path algorithm for generalized linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69 (2007), pp. 659–677.

[17] F. Picard, S. Robin, M. Lavielle, C. Vaisse, G. Celeux, and J.J. Daudin, *A statistical approach for cgh microarray data analysis*, Ph.D. diss., INRIA, 2004.

[18] F. Picard, S. Robin, E. Lebarbier, and J.J. Daudin, *A segmentation/clustering model for the analysis of array cgh data*, Biometrics 63 (2007), pp. 758–766.

[19] P. Ravikumar, M.J. Wainwright, and J.D. Lafferty, *High-dimensional Ising model selection using ℓ_1-regularized logistic regression*, Ann. Statist. 38 (2010), pp. 1287–1319. MR 2662343

[20] G. Simon, *Alternative analyses for the singly-ordered contingency table*, Journal of the American Statistical Association 69 (1974), pp. 971–976.

[21] A.S. Suggala, E. Yang, and P. Ravikumar, *Ordinal Graphical Models: A Tale of Two Approaches*, in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y.W. Teh, eds., Proceedings of Machine Learning Research Vol. 70, 06–11 Aug, International Convention Centre, Sydney, Australia. PMLR, 2017, pp. 3260–3269.

[22] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B 58 (1996), pp. 267–288. MR 1379242

[23] J.F. Trevor Hastie Robert Tibshirani, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.*, 2nd ed., Springer Series in Statistics, Springer, 2013.

[24] S.H. Walker and D.B. Duncan, *Estimation of the probability of an event as a function of several independent variables*, Biometrika 54 (1967), pp. 167–179. MR 0217928

[25] L. Wasserman and K. Roeder, *High dimensional variable selection*, Annals of statistics 37 (2009), p. 2178.

[26] O.D. Williams and J.E. Grizzle, *Analysis of contingency tables having ordered response categories*, Journal of the American Statistical Association 67 (1972), pp. 55–63.

[27] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel, and K. Lange, *Genome-wide association analysis by lasso penalized logistic regression*, Bioinformatics 25 (2009), pp. 714–721.

[28] P. Zhao and B. Yu, *On model selection consistency of Lasso*, J. Mach. Learn. Res. 7 (2006), pp. 2541–2563. MR 2274449

[29] J. Zhu and T. Hastie, *Classification of gene microarrays by penalized logistic regression*, Biostatistics 5 (2004), pp. 427–443.

A Experiments

In section 4, we described experimental results for dependent gaussian covariates. Let us now present additional experimental results relative to independent gaussian covariates and other distributions. In the same way as for section 4, we maintain some simulations settings: $p = 50$ covariates, $K = 3$, $n = 100$ samples, and vector of regression coefficients is $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$.

A.1 Independent (gaussian) covariates

In this case, covariates X are simulated as $X \sim \mathcal{N}_p(0, I_p)$.

A.2 Other distributions

In this case, covariates X_i are independent and distributions are the following:

- if $i \equiv 1 \pmod 3$, $X_i \sim \mathcal{N}(0, 1)$
- if $i \equiv 2 \pmod 3$, $X_i = \dfrac{Z_i - \mu}{\mu}$ where $Z_i \sim \mathcal{P}(\mu)$ and μ is chosen uniformly on $\{1, \ldots, 40\}$.
- if $i \equiv 0 \pmod 3$, $X_i \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$.

Parameters of these distributions are chosen so that each covariate is centered and reduced.

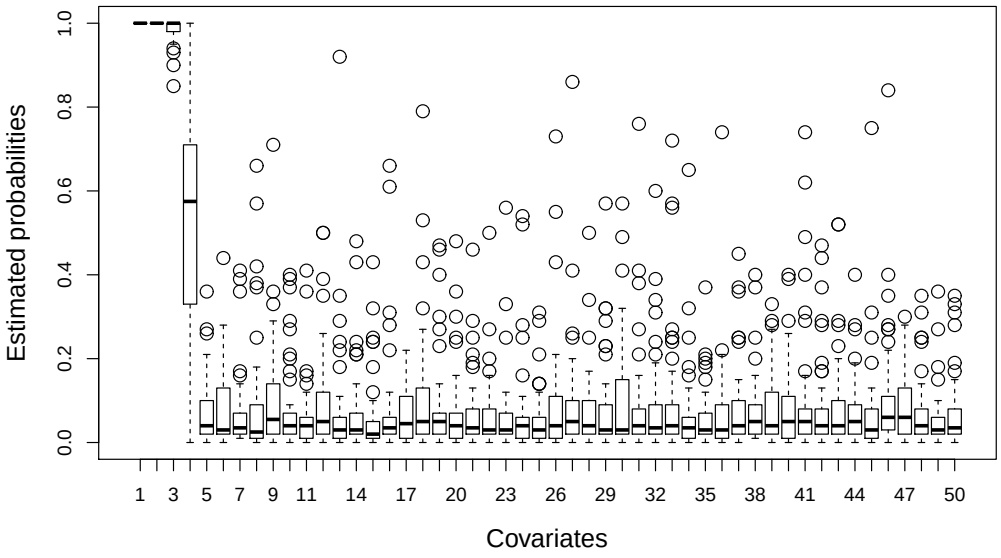

Figure 8: Boxplots of estimated probabilities $\max_{\tau \in T} \hat{p}_i(\tau)$ for each covariate. Covariates are independent and gaussian and regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 50 repetitions constituted by $n = 100$ samples of $p = 50$ variables with stability selection method.

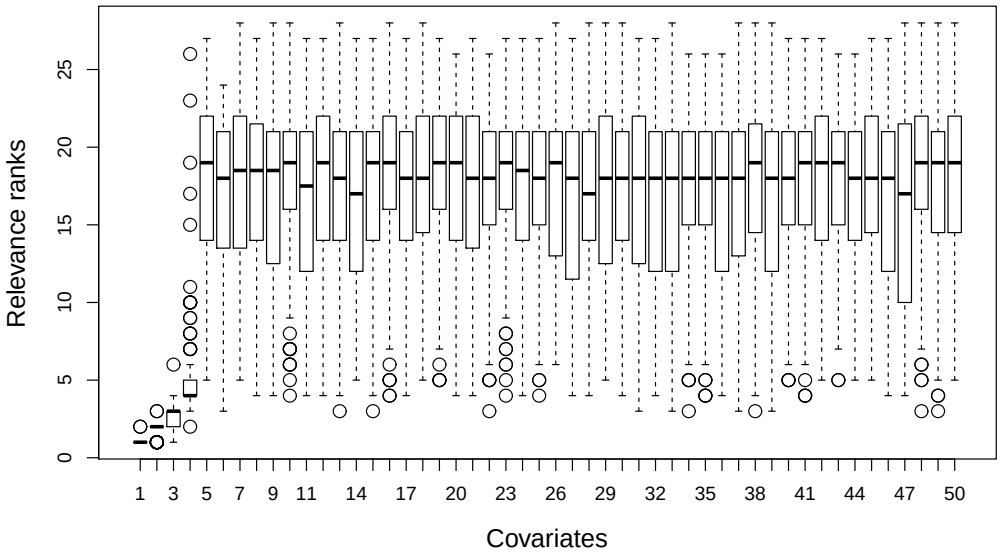

Figure 9: Boxplots of 'appearance'/relevance ranks for each variable. Covariates are independent and gaussian and regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 100 repetitions constituted by $n = 100$ samples of $p = 50$ variables with revisited knockoffs method.

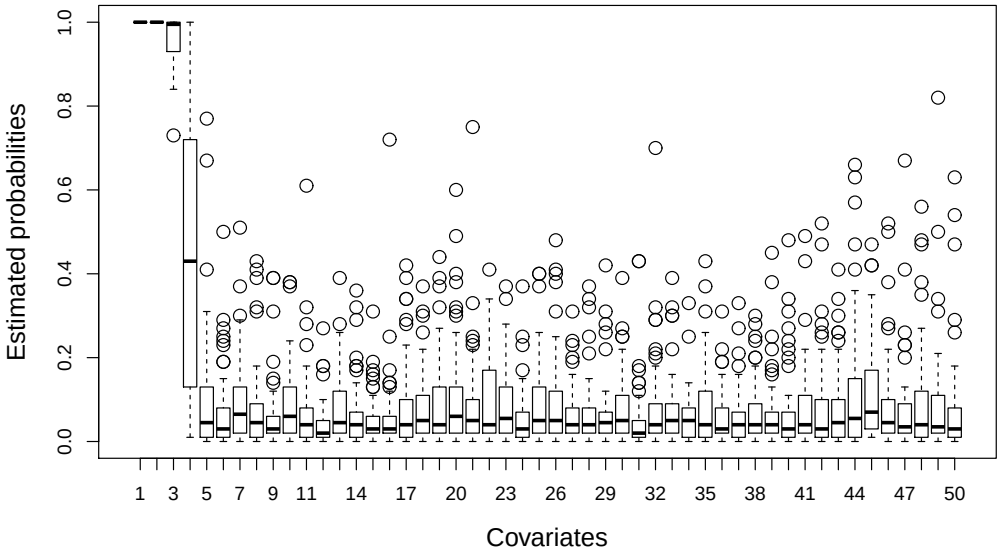

Figure 10: Boxplots of estimated probabilities $\max_{\tau \in T} \hat{p}_i(\tau)$ for each covariate. Covariates are independent and distributions are described in appendix A.2. Regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 50 repetitions constituted by $n = 100$ samples of $p = 50$ variables with stability selection method.

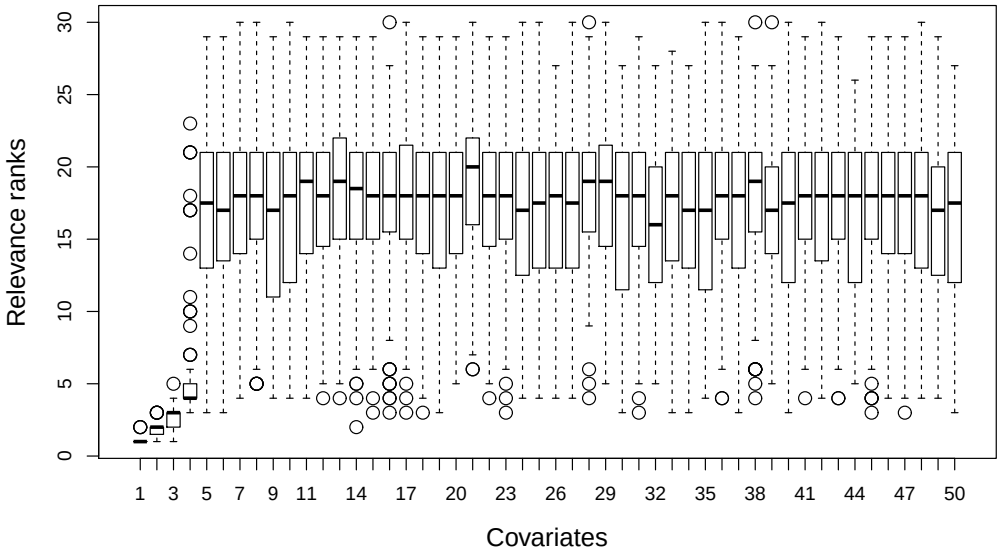

Figure 11: Boxplots of 'appearance'/relevance ranks for each variable. Covariates are independent and distributed as in appendix A.2. Regression coefficients $\boldsymbol{\beta} = (8, 6, 4, 2, 0, \ldots, 0)$. Boxplots are obtained on 100 repetitions constituted by $n = 100$ samples of $p = 50$ variables with revisited knockoffs method.