École Centrale Lyon

Report

Controlling the false discovery rate via knockoffs

Authors:
Rina Foygel Barber
Emmanuel J. Candès

Students:
Julien Gaubil
Jeremie Marlhens

October 28, 2022

Contents

1 Introduction

This study is based on the first article on knockoffs [2]. The idea is to contextualize and enlighten the reader on the theory and the proofs on which this theory is based as well as to understand the measures of its application in particular on the identification of dynamic system.

1.1 General context

1.1.1 High dimensional statistics

One of the problems encountered in high dimensional statistics is variable selection. In particular, a well-studied problem is when we look for the explanatory variables of a linear model. A well known solution is then the linear regression which is equivalent to its least squares formula:

$$\hat{x} = \min_{x \in \mathbb{R}^p} \|y - Ax\|_2^2$$

We have a matrix A of size $p \times n$, but in high dimension statistics we have a flat matrix, that is to say $p \gg n$. Then the problem can be ill-posed, the system is underdetermined and so we have an infinity of solutions. A solution to this problem is to add constraints, in our case we are interested in a sparsity constraint. In order to implement this sparsity constraint we can add a condition on the norm ℓ_0 of the regression vector, the Lagrangian version of the problem is then:

$$\hat{x} = \min_{x \in \mathbb{R}^p} \|y - Ax\|_2^2 + \lambda \|x\|_0$$

This problem being NP-hard, we can relax the constraints, it is the lasso approach (Least Absolute Shrinkage and Selection Operator) developed by Robert Tibshirani by replacing the norm ℓ_0 by the norm ℓ_1.

$$\hat{x} = \min_{x \in \mathbb{R}^p} \|y - Ax\|_2^2 + \lambda \|x\|_1$$

This problem has no closed form solution, but fortunately it is a convex problem an can be solved by gradient descent algorithms. As the ℓ_1 norm is not derivable on zero, another way of solving this convex problem is through proximal algorithms, as the proximal of the ℓ_1 norm can be computed and it corresponds to a soft thresholding. The iterative shrinkage-thresholding algorithm (ISTA) is the forward-backward iterative scheme for the lasso. A more recent version improves the rate of convergence of ISTA: FISTA (Fast ISTA). Orthogonal Matching Pursuit is an other algorithm that solves the ℓ_1 regularized problem, however, here, one provides the number of non-null hypothesis.

The lasso is a way to select variables that have an effect on y, from null variables that are independant on y. It depends on both noise and correlation between variables but in actually, the lasso does not necessarily cover the real explanatory variables. In particular, there is an explicit trade-off between FDP and True Positive proportion (TPP) (being the fraction of true positive discovered) [13].

1.1.2 False Discovery Rate

In statistics, when we test a hypothesis, we work under the null hypothesis, often called H_0. The idea is then to calculate, knowing that the null hypothesis is true, the probability of observing the observed data. If this probability is below a certain threshold value, we reject it. The p-value corresponds to the smallest threshold value for which we would reject the null hypothesis. Then we could define two types of error: the type I error, that is to say a false positive, that is to say that we reject the null hypothesis when it is true. Then a type II error corresponds to a false positive, that is to say that we accept the null hypothesis when it is false. One of the objectives of statistical testing is to control these errors. The threshold to reject the test is determined by $1-\alpha$, then the type I error can be controlled by α, this latter corresponds to the probability of type I error. The probability of type II error is β, it is often harder to control because we the alternative hypothesis can be anything.
The idea of False Discovery Rate (FDR) is to generalize the type I error when testing multiple hypothesis: it is the proportion of false discoveries among the discoveries (rejections of the null hypothesis). It can be written as follow:

$$FDR = \mathbb{E}\left[\frac{V}{V+S}\right]$$

where V is the number of false discoveries and S is the number of true discoveries. FDR and its importance in different applications (example genes, alleles and disease). The control of the FDR is particularly important in certain applications such as biology, where we are trying to understand whether certain genes are linked to certain diseases. We therefore want to avoid targeting genes that have nothing to do with the disease and we try to control the FDR to do so.

A common procedure to control the FDR is the Benjamini Hochberg (BHq) procedure [3]. The procedure is the following: given Z-scores $Z_1, \cdots, Z_p$ corresponding to p hypotheses such that $Z_j \sim \mathcal{N}(0,1)$ if the j^{th} hypothesis is null. We then define a data dependent threshold T for a desired level q of FDR control:

$$T = \min\left\{t : \frac{p \cdot \mathbb{P}\left\{|\mathcal{N}(0,1)| \geq t\right\}}{\#\left\{j : |Z_j| \geq t\right\}}\right\}$$

Then the procedure rejects p_j for $Z_j \geq T$.

One constraint is that the Z-scores have to follow a normal law, the knockoffs offer a more general framework. This won't be developed in the present study but one can prove that the knockoffs procedure and the BHq are equivalent in the orthogonal design setting [2].

1.1.3 Dynamic systems

The mathematical approach to the study of dynamical systems goes back to the beginning of physics, notably with Galileo and Newton. For simple systems, we can find the laws governing the dynamical system by applying the laws of physics, for example the laws of gravitation to an object in free fall is sufficient to predict its evolution knowing the initial

conditions. On the other hand, at other scales, the equations governing a dynamical system cannot be deduced from a traditional physical approach. For example, we can think of the meteorological models developed by Lorenz in the 1950s and the whole theory of chaotic systems [12]. An approach that is gaining importance in recent years has been enabled by the developments of data science, it is to infer the laws of a system from data and with very little *a priori*.

An object in motion can then be described by a system of ordinary differential equations. We restrict ourselves to this framework, but note that other mathematical frameworks can describe dynamical systems.

$$\begin{aligned} \frac{dx}{dt} &= f(x(t), u(t)) \\ y &= g(x(t), u(t)) \end{aligned}$$

Recent developments in control engineering have made many efforts to apply data science principles to engineering problems, such as system identification and control [4]. Some of these include: Compressive sensing, Tailored sensing, sparse sensors placement, dynamic mode decomposition and Sparse identification of nonlinear dynamics (SINDy) that we describe here.

The field of system identification takes roots in control theory, where the purpose is to identify the system in order to better control it. It aims at establishing a model from time series observations of the system. This field has grown into a multidisciplinary field that concerns statistics, machine learning, and is applied in control engineering and econometrics [10].When the functions g and f are linear, the problem can be solved and has been extensively studied [10], however when they are nonlinear the problem is much more involved and still a challenge for most of the cases [11].

1.2 Presentation of the Knockoffs

The article focuses on False discovery in variable selection for a linear model [2]. In the case of variable selection, for a linear relation, we have

$$y = X\beta + z$$

In the article setup, we need $2p \leq n$, an thus we are not in the so called high dimension context, further developments of the knockoffs have been applied to high dimension statistics. The FDR in this case is the expected proportion of selected variables that correspond to null variable $\beta_j = 0$ among all the selected variables. Let the selection procedure return a subset $\hat{S} \subset \{1, \cdots, p\}$

$$FDR = \mathbb{E}\left[\frac{j : \beta_j = 0 \text{ and } j \in \hat{S}}{j : j \in \hat{S} \vee 1}\right]$$

The objective is to follow a selection procedure that upper bounds the FDR with a given level q. In this setup, the different variables have a correlation structure given by $\Sigma = X^T X$, the idea of knockoff is to create another set of variables $\tilde{X}$ that have a similar correlation structure but that the correlation between the original variable and its knockoff is weak.

1.2.1 A way to build knockoffs and define statistics

We need to build the knockoff so that they satisfy the previous statistics. In the case of variable selection for the linear regression, we can build it as follow:

Step 1: Building the knockoffs We build the knockoffs $\tilde{X}$ so that we have the following correlation matrices:

$$\tilde{X}^T\tilde{X} = \Sigma \text{ and } X^T\tilde{X} = \Sigma - diag(s)$$

with $s \in \mathbb{R}_+^p$. We are looking for the largest s to ensure good statistical power, in other words we are looking for knockoffs that are the most uncorrelated from the original variables while respecting the given correlation structure. To construct $\tilde{X}$, choose $s \in \mathbb{R}_+^p$ satisfying $diag(s) \preceq 2\Sigma$. Let $\tilde{U} \in \mathbb{R}^{n\times p}$ the orthonormal matrix orthogonal to the span of the features X, and $C \in \mathbb{R}^{p\times p}$ such that $C^TC = 2diag(s) - diag(s)\Sigma^{-1}diag(s)$. Then we can define the knockoffs of X to be:

$$\tilde{X} = X(I - \Sigma^{-1}diag(s)) + \tilde{U}C$$

Step 2: Compute a statistics The statistics $W_j, \forall j \in \{1, \cdots, p\}$. The statistic j allows to determine if a variable is considered as non null, in particular a large W_j value is in favor of an alternative hypothesis of the null hypothesis $\beta_j = 0$. We consider the lasso model defined previously and the vector $\hat{\beta}(\lambda)$

$$\hat{\beta}(\lambda) = \underset{b\in\mathbb{R}^p}{\arg\min}\left\{\frac{1}{2}\|y - \left[X\tilde{X}\right]b\|_2^2 + \lambda\|b\|_1\right\}$$

We define a first statistics $Z_j = sup\left\{\lambda : \hat{\beta}_j(\lambda) \neq 0\right\}$. This yields a vector $(Z_1, \cdots, Z_p, \hat{Z}_1, \cdots, \hat{Z}_p)$. Then for all j in $\{1, \cdots, p\}$ we define the statistics W

$$W_j = max(Z_j, \tilde{Z}_j) \cdot \begin{cases} +1 \text{ if } Z_j > \tilde{Z}_j \\ -1 \text{ if } Z_j < \tilde{Z}_j \\ 0 \text{ if } Z_j = \tilde{Z}_j \end{cases}$$

Step 3: Choose model (in other words, the level of FDR to control) The procedure is as follow: we select W_j such that $W_j \geq T$ for T a data dependent threshold defined as follow:

$$T = min\left\{|W_k| : \frac{1 + |\ \{j/W_j \leq -|W_k|\}\ |}{max(1,\ |\ \{j/W_j \geq -|W_k|\}\ |)} \leq q\right\}$$

Then, the FDR is controlled at a level q.

1.2.2 Limits of the 2015 paper

The original paper paved the theory for further development of the knockoffs, however some issues may be to solve. First we need $p \leq n$ (presented framework $2p \leq n$ and extension for $p \leq n \leq 2p$), one would need to extend the problem to actual high dimension statistics when $p \geq n$ to apply in particular in gene/disease problems.
Then the procedure to build knockoffs is restricted to linear models (we need the sufficiency property). This could be extended to any other models as long as the knockoff properties are verified.

1.2.3 State of the art

Extension were realized to build knockoffs for a linear model in high dimensions $p \geq n$ [6].

2 Theory and details of selected proofs

The proof will treat the case where the number of observation n and the number of variable p verify $n \geq 2p$.

2.1 Construction of the Knockoff

We consider in this section a design matrix $X \in M_{n,p}(\mathbb{R})$ whose columns are $l2$-normalized. Let's note $\Sigma = X^T X$ the covariance-variance matrix of X.

We want to create a knockoff version $\tilde{X}$ of the design matrix X that verifies $X^T X = \tilde{X}^T \tilde{X}$ and $X^T \tilde{X} = X^T X - diag(s)$ where $s \in \mathbb{R}^p$. It means that X and $\tilde{X}$ share the same correlation structure except for the diagonal terms where $X_j^T \tilde{X}_j = 1 - s_j$

Definition 2.1 (equi-correlated Knockoffs)**.** The s_j are chosen to be all equals to $min(2\lambda_{min}, 1)$ where λ_{min} is the minimal eigenvalue of Σ. It minimizes $| < X_j, \tilde{X}_j > |$.

Definition 2.2 (Semi-definite Program Knockoffs)**.** The knockoffs are selected so that the mean correlation between X and $\tilde{X}$ is minimal, which is equivalent to the SDP problem :

$$\begin{cases} \min\limits_{s} \sum_j |1 - s_j| \\ \text{s.t..}\ s_j \geq 0,\ diag(s) \leq 2\Sigma \end{cases}$$

Proposition. *For both Knockoff constructions presented, $diag(s) \leq 2\Sigma$ and the covariance matrix of the augmented design, $G = \left[X\tilde{X}\right]^T \left[X\tilde{X}\right]$, is symetric semi-definite positive. Then, for a matix C given by a Cholesky decomposition of $\Sigma - (\Sigma - diag(s))\Sigma^{-1}(\Sigma - diag(s))$, for an orthnormal matrix $\tilde{U} \in M_{n,p}(\mathbb{R})$ s.t. $X^T\tilde{U} = 0$, $\tilde{X} = X(I - \Sigma^{-1}diag(s)) + \tilde{U}C$ satisfies the desired covariance structure.*

Proof.

$$G = \begin{pmatrix} \Sigma & \Sigma - 2diag(s) \\ \Sigma - diag(s) & \Sigma \end{pmatrix}$$

By Schur complement property,

$$G = \begin{pmatrix} \Sigma & \Sigma - 2diag(s) \\ \Sigma - diag(s) & \Sigma \end{pmatrix}$$

$$\text{G}\succeq 0 \iff A = \Sigma - (\Sigma - diag(s))\Sigma^{-1}(\Sigma - diag(s)) \succeq 0 \iff \begin{pmatrix} \Sigma & diag(s) \\ diag(s) & \end{pmatrix} \succeq 0$$

$$\text{Then } G \succeq 0 \iff \begin{cases} diag(s) \geq 0 \\ 2\Sigma \geq diag(s) \end{cases}$$

$diag(s) \leq 2\Sigma$ clearly by definition of equi-correlated and SDP Knockoffs, and in both cases $diag(s) \geq 0$, therefore $G \succeq 0$ for equi-correlated and SDP Knockoffs.

$A \succeq 0$, let $C \in M_p(\mathbb{R})$ so that the Cholesky decomposition of A gives $A = C^T C$
Since $n \geq 2p$, it is possible to find $\tilde{U} \in M_{n,p}(\mathbb{R})$ orthonormal s.t. $\tilde{U}^T X = 0$.
Then, with $\tilde{X} = X(I - \Sigma^{-1}) + \tilde{U}C$ satisfies the covariance structure:

$$X^T\tilde{X} = \Sigma - \Sigma\Sigma^{-1}diag(s) + X^T\tilde{U}C = \Sigma - diag(s)$$

and

$$\begin{aligned}\tilde{X}^T X &= \Sigma - diag(s)\Sigma^{-1}\Sigma - \Sigma\Sigma^{-1}diag(s) + diag(s)\Sigma^{-1}\Sigma\Sigma^{-1}diag(s) + C^T\tilde{U}^T\tilde{U}C \\ &= \Sigma - diag(s) - diag(s) + diag(s)\Sigma^{-1}diag(s) + 2diag(s) - diag(s)\Sigma^{-1}diag(s) = \Sigma\end{aligned}$$

□

2.2 The statistic W

Definition 2.3 (sufficiency property)**.** The statistic W is said to be sufficient if it is a function only of the Gram matrix of the augmented design, $G = \left[X\tilde{X}\right]^T \left[X\tilde{X}\right]$ and of the scalar products between features and response $X^T y$.

Definition 2.4 (antisymmetry property)**.** $\forall S \subset [p]$, swapping X_j and $\tilde{X}_j$ in the augmented design matrix $\left[X\tilde{X}\right]$ changes the sign of W_j, the j-th component of the statistic W constructed.

Definition 2.5 (apparition of a variable in the Lasso)**.** For the Lasso problem where the estimator is given by : $\hat{\beta}(\lambda) = \underset{b}{argmin}\{\frac{1}{2}||y - Xb||^2 + \lambda||b||_1\}$, we introduce as Z_j the λ for which the variable X_j first enters the model. $Z_j = sup\{\lambda / \hat{\beta}_j(\lambda) \neq 0\}$

Proposition. *The statistic* $(Z_1, ..., Z_p, \tilde{Z}_1, ..., \tilde{Z}_p)$ *is computed on the augmented design* $\left[X\tilde{X}\right]$. $\forall\ j \in [p]$, *let* $W_j = max(Z_j, \tilde{Z}_j) \times (-1_{(Z_j<\tilde{Z}_j)} + 1_{(Z_j>\tilde{Z}_j)}$. *Then* $W = (W_1, ..., W_p)$ *satisfies the sufficiency and anti-symmetry properties.*

Proof. The lasso problem is equivalent to $\min_b \frac{1}{2}b^T X^T X b - b^T X^T y + \lambda\|b\|_1$, which is a function of $X^T X$ and $X^T y$. W therefore satisfies the sufficiency property.

It is clear from the construction of W that it satisfies the antisymmetry property. □

Lemma 2.1. *Let* $\epsilon \in \{-1, 1\}^p$ *a sequence of signs independent of the components* W_j *of the statistic* W, *so that* $\epsilon_j = 1$ *for all* $j \in [p]$ *s.t.* $\beta_j \neq 0$ *and* $\epsilon_p \to \{-1, 1\}$ $\forall j$ *s.t.* $\beta_j = 0$. *Then the equality in law* $(W_1, ..., W_p) \overset{d}{=} (W_1\epsilon_1, ..., W_p\epsilon_p)$ *holds.*

The proof is available in the supplementary information.

Proposition. $\forall j$ *s.t.* $\beta_j = 0$, *the signs of the* W_j *are i.i.d. and* $sign(W_j) \sim \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$. *They are also independent of the signs* $sign(W_i)$ $\forall i$ *s.t.* $\beta_i \neq 0$, *and independent of the* $|W_j|$ $\forall j \in [p]$.

2.3 p-values for the Knockoffs

In this part we consider $m = |\ \{j/W_j \neq 0\}\ |$. The method indeed never selects a variable of index j so that $W_j = 0$, they can therefore be neglected. We can, without loss of generality, consider $|W_1| \geq |W_2| \geq ... \geq |W_m|$.

Definition 2.6 (p-values)**.** $\forall j \in [m]$ (the indexes s.t. $W_j \neq 0$), we define the p-value

$$p_j = \begin{cases} \frac{1}{2} \text{ if } W_j > 0 \\ 1 \text{ if } W_j < 0 \end{cases}$$

Proposition. *The p-values* $\{p_j/j \in [m], \beta_j = 0\}$ *are i.i.d. and follow the law* $\frac{1}{2}\delta_{\frac{1}{2}} + \frac{1}{2}\delta_1$. *They are also independent of the* $sign(W_i)$ $\forall i$ *s.t.* $\beta_i = 0$.

Proof. $\forall j$ s.t. $\beta_j = 0$, $sign(W_j)$ are i.i.d. and follow the law $\frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, then those p_j clearly follow the law $\frac{1}{2}\delta_{\frac{1}{2}} + \frac{1}{2}\delta_1$ by definition.

The $sign(W_j)$ s.t. $\beta_j = 0$ are independent from the $sign(W_i)$ s.t. $\beta_i \neq 0$. It follows that p_j s.t. $\beta_j = 0$ are independent from the p_i s.t. $\beta_i \neq 0$.

□

Proposition (stochastic dominance of the p-values)**.** $\forall j \in [m]$ *s.t.* $\beta_j = 0$, $\forall u \in [0,1]$, $P(p_j \leq u) \leq u$, *which we denote* $p_j \overset{d}{\geq} U([0,1])$.

Proof. Clear because $\forall j \in [m]$, $\forall u \in [0, \frac{1}{2}[$, $\mathbb{P}(p_j \leq u) = 0 \leq u$ and $\forall u \in [\frac{1}{2}, 1[$, $\mathbb{P}(p_j \leq u) = \frac{1}{2} \leq u$, and $\mathbb{P}(p_j \leq 1) = 1$.

□

2.4 Sequential procedure

The following definition is provided in a general setting:

Definition 2.7 (sequential procedure for p-values)**.** We consider the test hypothesis $H_1, ..., H_m$ where $\forall i \in [m]$, $(H_i) : \beta_i = 0$. The associated p-values are denoted p_j. If $p_j > c$, we accept (H_i), otherwise we reject (H_i). Let $K \subset [m]$ and q a control rate.

1. we first calculate

$$\hat{k_1} = max\{k \in K / \frac{1 + \mid \{j \leq k / p_j > c\} \mid}{max(1, \mid \{j \leq k / p_j \leq c\} \mid)} \times \frac{c}{1-c} \leq q\}$$

2. $\forall j \in K$ s.t. $j \leq \hat{k_1}$, $p_j \leq c$, we reject $(H_j) \implies \beta_j \neq 0$.

Proposition (Link with Knockoffs and T)**.** *With* $T = min\{|W_k| / \frac{1+\mid \{j/W_j \leq -|W_k|\} \mid}{max(1, \mid \{j/W_j \geq -|W_k|\} \mid)} \leq q\}$, $m = \mid \{j / W_j \neq 0\} \mid$ *and if the* W_i *are ordered by decreasing magnitude* $|W_1| \geq |W_2| \geq ... \geq |W_m|$, *then* $\forall j \in [m]$, $W_j \geq T \Leftrightarrow j \leq \hat{k_1}$ *and* $p_j \leq \frac{1}{2}$

Proof. We apply the procedure provided by the previous definition with $K = \{k \in [m] / |W_k| > |W_{k+1}|\} \cup \{m\}$ and a threshold $c = \frac{1}{2}$:

Then we have by definition:

$$\hat{k_1} = max\{k \in K / \frac{1 + \mid \{j \leq k / p_j > \frac{1}{2}\} \mid}{max(1, \mid \{j \leq k / p_j \leq \frac{1}{2}\} \mid)} \leq q\}$$

$$= max\{k \in K / \frac{1 + |\{j / W_j \leq -[W_k]\}|}{max(1, |\{j / W_j \geq -|W_k|\}|)} \leq q\}$$

by definition of K, and the fact that the $(W_k)_{k \in K}$ are ordered in non-decreasing order. It comes from the fact that the $(W_k)_{k \in K}$ are ordered in non-decreasing order :

$$\hat{k_1} = \underset{k}{argmin}\{|W_k| / \frac{1 + \mid \{j / W_j \leq -[W_k]\} \mid}{max(1, \mid \{j / W_j \geq -|W_k|\} \mid)} \leq q\}$$

If we pose $T = min\{|W_k|/\frac{1+|\ \{j/W_j \leq -[W_k|\}\ |}{max(1,\ |\ \{j/W_j \geq -|W_k|\}\ |)} \leq q\}$, we have by the decreasing order of magnitude of the W_i:

$$\forall j \in [m],\ W_j \geq T \ \Leftrightarrow\ j \leq \underset{k}{argmin}\{|W_k|/\frac{1+|\ \{j/W_j \leq -[W_k|\}\ |}{max(1,\ |\ \{j/W_j \geq -|W_k|\}\ |)} \times \leq q\} \text{ and } W_j \geq 0$$
$$\Leftrightarrow\ j \leq \hat{k_1} \text{ and } W_j \geq 0 \ \Leftrightarrow\ j \leq \hat{k_1} \text{ and } p_j \leq \frac{1}{2}$$

which concludes the proof.

□

2.5 Control of the FDR for the Knockoffs

We still consider $m = |\{j/W_j \neq 0\}|$ and $|W_1| \geq |W_2| \geq ... \geq |W_m|$. We consider $K = \{k \in [m]/|W_k| > |W_{k+1}|\} \cup \{m\}$.

Lemma 2.2. *$\forall k \in [m]$, let $V^+(k) = |\ \{1 \leq j \leq k/\beta_j = 0,\ p_j \leq c\}\ |$ and $V^-(k) = |\ \{1 \leq j \leq k/\beta_j = 0,\ p_j > c\}\ |$ with the convention $V^{\pm}(0) = 0$. Let $F_k = \sigma(\bigcup_{j\in[m]} \sigma(p_j) \cup \bigcup_{k' \geq k} \sigma(V^{\pm}(k')))$ which defines the filtration in reversed time $(F_k)_{k=m,...,1}$. Let $M(k) = \frac{V^+(k)}{1+V^-(k)}$. Then $(M(k))_{0\leq k\leq m}$ is a super-martingale (in reversed time) w.r.t. the filtration $(F_k)_{0\leq k\leq m}$.*

For the procedure previously described with $\hat{k_1}$, we also have $\mathbb{E}[M(\hat{k_1})] \leq \frac{c}{1-c}$.

Proof. $(F_k)_{0\leq k\leq m}$ clearly is a filtration in inverse time.

$\forall k \in [m],\ M(k) = \frac{V^+(k)}{1+V^-(k)}$ clearly is F_k-measurable by definition of F_k.

Let $k \in [m]$.

If $\beta_k \neq 0$, we clearly have $V^+(k) = V^+(k-1)$ and $V^-(k) = V^-(k-1)$.

Else, if $\beta_k = 0$,

$$M(k-1) = \frac{V^+(k) - 1_{p_k \leq c}}{1 + V^-(k) - 1_{p_k > c}} = \frac{V^+(k) - 1_{p_k \leq c}}{1 + V^-(k) - (1 - 1_{p_k \leq c})} = \frac{V^+(k) - 1_{p_k \leq c}}{V^-(k) + 1_{p_k \leq c})}$$

The authors state that $P(1_{p_k \leq c}) = \frac{V^+(k)}{V^+(k)+V^-(k)}$ and deduce that

$$\mathbb{E}[M(k-1)|F_k] = \begin{cases} \frac{V^+(k)}{1+V^-(k)} = M(k) \text{ if } V^-(k) > 0 \\ V^+(k) - 1 = M(k) - 1 \text{ if } V^-(k) = 0 \end{cases}$$

and it comes

$$\mathbb{E}[M(k-1)|F_k] = \begin{cases} M(k) \text{ if } \beta_k \neq 0 \\ M(k) \text{ if } \beta_k = 0,\ V^-(k) > 0 \\ M(k) - 1 \text{ if } \beta_k = 0,\ V^-(k) = 0 \end{cases}$$

We therefore have $\mathbb{E}[M(k-1)|F_k] \leq M(k)$, which proves that $(M(k))_{0\leq k\leq m}$ is a super-martingale (in reversed time) w.r.t. the filtration $(F_k)_{0\leq k\leq m}$.

$\hat{k_1}$ clearly is a (F_k)-stopping time :

$$\forall k \in [m], (\hat{k_1} \leq k) = \bigcap_{k \leq j \leq m} (\frac{1 + \mid \{j \leq k / p_j > c\} \mid}{max(1, \mid \{j \leq k / p_j \leq c\} \mid)} \times \frac{c}{1-c} > q) \ \in F_k$$

clearly by definition of F_k.

By the optional stopping theorem (since $\hat{k_1}$ is clearly bounded), it comes that $\mathbb{E}[M(\hat{k_1})] \leq \mathbb{E}[M(m)]$ since $(M(k))_{k \in [m]}$ is a super-martingale in inversed time.

Let Y be a random variable that follows the binomial law of parameters $N = |\{j/\beta_j = 0\}|$ and c, $B(N, c)$. It follows from the stochastic dominance $\forall j$ t.q. $\beta_j = 0,\ p_j \overset{d}{\geq} U([0,1])$ that $V^+(m) \overset{d}{\leq} B(N,c)$. Then, $\forall k \in [m]$, $P(V + (m)\ k) \leq P(Y > k)$. What's more, $V^-(m) = N - V^+(m)$ and since $x \to \frac{x}{1+N-x}$ is a non-decreasing function, it comes :

$$\mathbb{E}[M(\hat{k_1})] \leq \mathbb{E}[M(m)] = \mathbb{E}[\frac{V^+(m)}{1 + N - V^+(m)}] \leq \mathbb{E}[\frac{Y}{1+N-Y}] = \sum_{1 \leq i \leq N} P(Y = i) \frac{i}{1+N-i}$$

$$= \sum_{1 \leq i \leq N} c^i \ (1-c)^{N-i} \ \frac{N!}{i!(N-i)!} \frac{i}{1+N-i} = \frac{c}{1-c} \sum_{1 \leq i \leq N} c^{i-1} \ (1-c)^{N-i+1} \ \frac{N!}{(i-1)!(N-i+1)!}$$

$$= \frac{c}{1-c} \sum_{1 \leq i \leq N} P(Y = i-1) \leq \frac{c}{1-c}$$

which concludes the proof.

□

Theorem 2.3. *If, $\forall j \in [m]$ (s.t. $\beta_j = 0$), the p-values p_j are i.i.d., verify the stochastic dominance $p_j \overset{d}{\geq} U([0,1])$, and are independent from the sign(W_i) $\forall\ \beta_i$ s.t. $\beta_i \neq 0$. Then the selection procedure described with $\hat{k_1}$ that selects a model of support $\hat{S}$ verifies $\mathbb{E}[\frac{V}{max(1,R)}] \leq q$ where $R = |\hat{S}|$ is the number of variables selected and $V = \mid \{j \in \hat{S} / \beta_j = 0\} \mid$ is the number of false discoveries.*

Proof.

$$\mathbb{E}[\frac{V}{max(1,R)}] = \mathbb{E}[\frac{\mid \{j \leq \hat{k_1} / \beta_j = 0,\ p_j \leq c\} \mid}{1 + \mid \{j \leq \hat{k_1} / \beta_j = 0,\ p_j > c\} \mid} \times \frac{1 + \mid \{j \leq \hat{k_1} / \beta_j = 0,\ p_j > c\} \mid}{max(1, \mid \{j \leq \hat{k_1} /\ p_j \leq c\} \mid)}]$$

By definition of $\hat{k_1}$,

$$\frac{1 + \mid \{j \leq \hat{k_1} / \beta_j = 0,\ p_j > c\} \mid}{max(1, \mid \{j \leq \hat{k_1} /\ p_j \leq c\} \mid)} \ \leq \ \frac{1-c}{c} \times q$$

And by the previous lemma :

$$\mathbb{E}[\frac{|\ \{j \leq \hat{k_1}/\beta_j = 0,\ p_j \leq c\}\ |}{1 + |\ \{j \leq \hat{k_1}/\beta_j = 0,\ p_j > c\}\ |}] = \mathbb{E}[M(\hat{k_1})] \leq \frac{c}{1-c}$$

Hence the result

$$\mathbb{E}[\frac{V}{max(1,R)}] \leq \frac{c}{1-c}\frac{1-c}{c} \times q = q$$

□

Proposition (Control of the FDR for knockoffs)**.** *We suppose the W_i ordered by decreasing magnitude $|W_1| \geq |W_2| \geq ... \geq |W_m|$. The procedure of selection of Knockoffs defined by $\hat{S} = \{j/W_j \geq T\}$ where T is a data-depending threshold defined by $T = min\{|W_k|/\frac{1+|\ \{j/W_j \leq -|W_k|\}\ |}{max(1,\ |\ \{j/W_j \geq -|W_k|\}\ |)} \leq q\}$ enables a control of the FDR at the level q : $\mathbb{E}[FDR] \leq q$*

Proof. $\hat{S} = \{j/W_j \geq T\} = \{j/j \leq \hat{k_1},\ p_j \leq \frac{1}{2}\}$ by a result previously proved. The selection of the model $\hat{S}$ therefore corresponds to the selection procedure described for the p-values with the threshold $c = \frac{1}{2}$.

Then

$$FDR = \mathbb{E}[\frac{|\ \{j \leq \hat{k_1}/\beta_j = 0,\ p_j \leq \frac{1}{2}\}\ |}{max(1,\ |\ \{j \leq \hat{k_1}/\ p_j \leq \frac{1}{2}\}\ |\ )}] = \mathbb{E}[\frac{V}{max(1,R)}]$$

The p-values $p_j\ \forall j$ t.q. $\beta_j = 0$ are i.i.d. and verify the stochastic dominance $p_j \overset{d}{\geq} U([0,1])$. They are also independent from the $sign(W_i)\ \forall i$ t.q. $\beta_i \neq 0$. The previous theorem therefore provides :

$$FDR \leq q$$

□

3 Knockoffs applied to System identification

3.1 Approach by *Brunton et al.*[5]

3.1.1 Problem

The system identification problem consists in building mathematical models of dynamical systems based on observed data from the system. More precisely, by looking at how the system evolves over time, we can try to infer a model with little *a priori* information. It is thus one type of inverse problem. One recent approach is inspired by the developments

and different applications of the lasso [5]. We will show that it is a linear regression problem with a parsimonious overlapping vector where it can be solved with the lasso. We will then focus on controlling the FDR through knockoffs.

We first consider a fully observed nonlinear dynamic system with unkown dynamics f.

$$\frac{dx}{dt} = f(x(t))$$

3.1.2 Solution

We have discrete measurements of this system that we store in the following matrix:

$$X = \begin{bmatrix} x^T(t_1) \\ . \\ . \\ . \\ x^T(t_n) \end{bmatrix}$$

with $x^T(t_k) = [x_1(t_k),\ \dots\ , x_m(t_k)]$.

And we calculate the derivatives of the measurements. Different approaches can be used such as the finite differences. If the data is noisy, it can be smoothed, different methods exist [7]

$$\dot{X} = \begin{bmatrix} \dot{x}^T(t_1) \\ . \\ . \\ . \\ \dot{x}^T(t_n) \end{bmatrix}$$

$\dot{x}^T(t_k) = [\dot{x}_1(t_k), \cdots, \dot{x}_m(t_k)]$

We define a dictionnary of function $\Theta(X)$:

$$\Theta(X) = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & X^{P_2} & X^{P_3} & \cdots & cos(X) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

with $\Xi = [\xi_1, \xi_2, \cdots, \xi_m]$ and $\xi_k = [\xi_{k,1}, \xi_{k,2}, \cdots, \xi_{k,p}]$. The dictionary base depends on the system studied and requires the knowledge of the expert. One condition is that the system equations are expected to be sparse in the whole dictionary, allowing thus to use sparse promoting regressions as the lasso, with ξ_k being sparse vectors.

We can see in this case that we have the following linear system to solve:

$$\dot{X} = \Theta(X)\Xi$$

Separating for each dimension of the dynamical system we obtain the following equations for its dynamics:

$$\dot{x_k} = f_k(x) = \Theta(x^T)\xi_k$$

And for the whole system:

$$\dot{x} = f(x) = \Xi^T \Theta(X^T)^T$$

For this study we will use the chaotic nonlinear Lorenz system described with the following nonlinear equations:

$$\begin{cases} \dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z \end{cases}$$

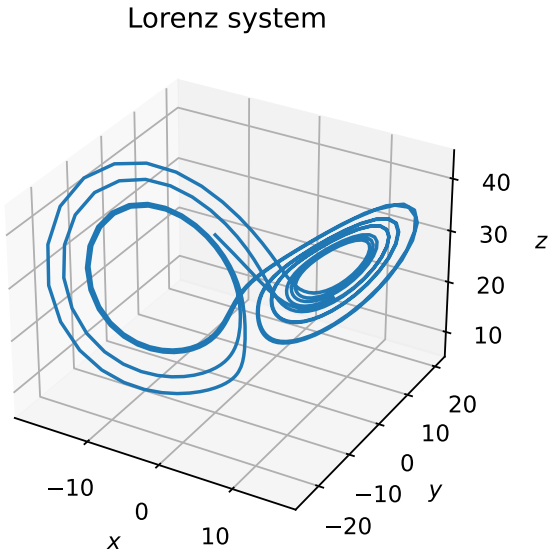


Figure 1: Simulation of a chaotic system: the Lorenz system. This simulation correspond to the training set in the following section

We will use the Python package PySindy to simulate and store the dynamic models used in [8].

3.1.3 Selection of the model and cross validation

$$\arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 + \lambda^2 \|x\|_0$$

As noted in the introduction, the common practice of parameter selection in data science is by cross-validation. It is a reliable procedure whose principle is to separate the data to have a training set and a testing set. The training set allows to fit the parameters of the model, for example to find the explanatory variables of the linear model here, and the testing set allows to check that the model can generalize to new data and thus avoid over fitting. An other way that is common in applied mathematics and statistics is information criteria that we won't use here.

3.1.4 Improvements over the original method

The original using a λ threshold scan, we implement a lasso path using the LARS from scikit learn that makes the simulations faster. We can compare the performance from the original method is similar to ours using the lasso path and the fitting the 3 derivatives corresponding to the dimensions independently in the supplementary results.

3.2 Implementations

We simulate the Lorenz system for two different starting point. Each set consists of $n = 500$ time points with $dt = 0.01$. For the function library we took the polynomial library up to order 4, meaning all the polynomial from order 0 to order 4: $1, x, y, z, x^2, xy, xz, y^2, \cdots, z^4$.

We realize a lasso path on each dimension x, y and z of the problem independently (figure 2). We then realize a cross-validation independently for each dimension of the system on a test (figure 3). By choosing the best λ (which may be different for each dimension) we obtain the following model:

$$\begin{cases} \dot{x} &= -4.298x + 7.582y - 0.175xz + 0.001y^2 + 0.035yz + 0.011x^3 \\ \dot{y} &= 0.126 + 12.969x + 6.456y + 0.072xy - 0.101xz - 0.040y^2 - 0.326yz - 0.016x^3 \\ & \quad -0.014xz^2 + 0.001x^3z \\ \dot{z} &= -2.699z + 0.898xy + 0.111y^2 \end{cases}$$

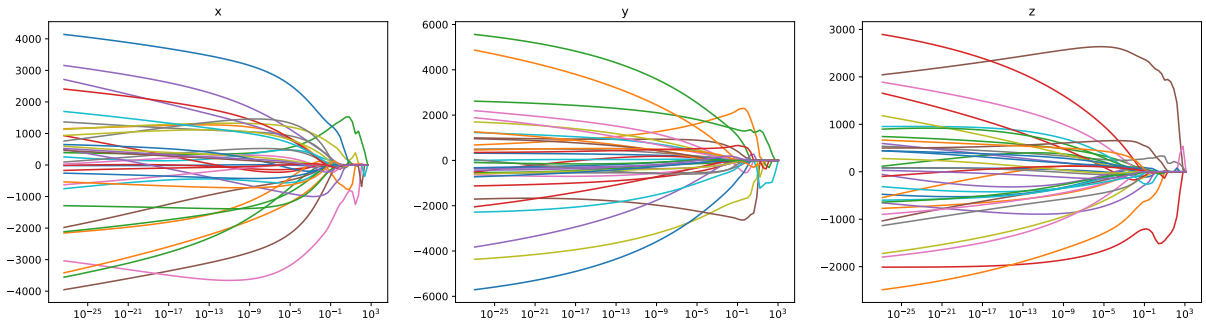


Figure 2: Lasso path for the three dimensions

We can compare the recovered model and the system on the training set for example (similar results on the testing set) (figure 4 and 5). We can see that in spite of the errors in the recovery of the support the predictions are rather good, in spite of deviations from the real system it seems that the model is found the right attractor, that it escapes sometimes and joins the real system.

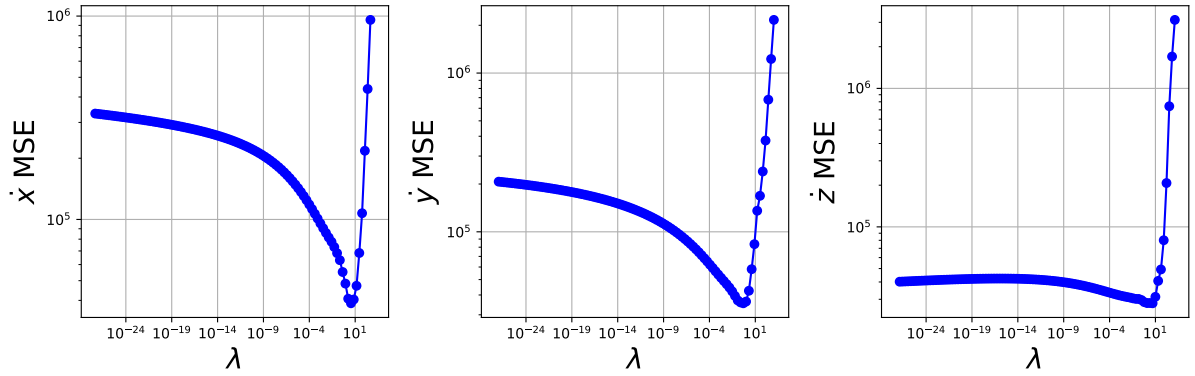


Figure 3: Cross-validation on the computed derivative testing set. The cross-validation is independent for the three dimensions and can lead to different optimal λ

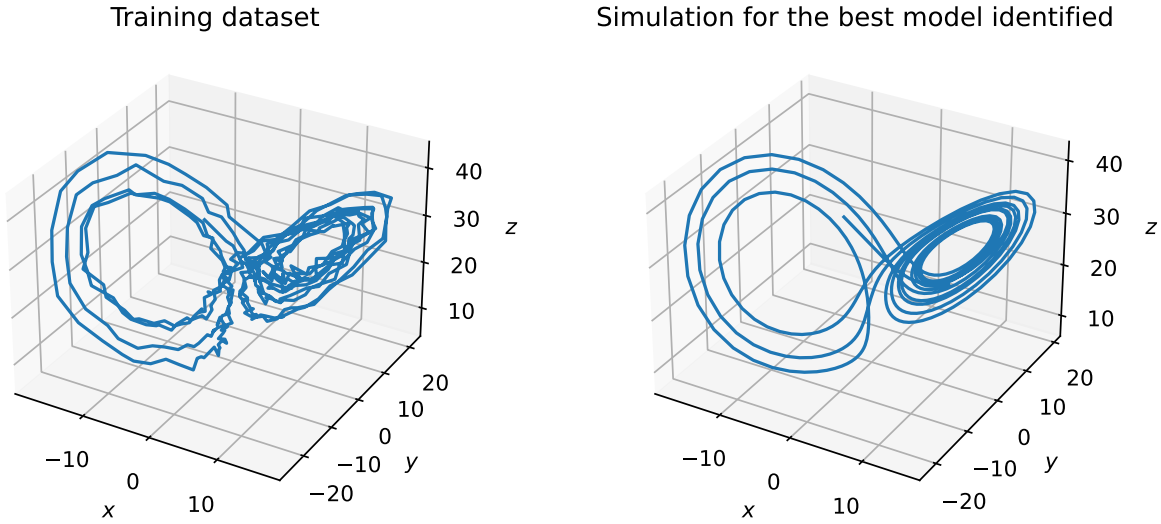
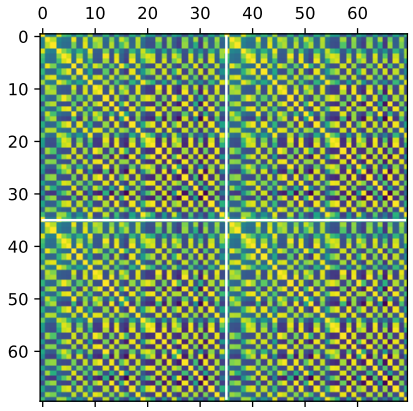
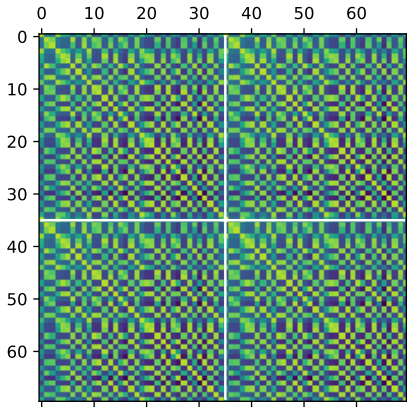


Figure 4: Simulation of the identified model, and comparisons with the training dataset in three dimensions. The simulation is given the same starting point as the training dataset and simulated for the same duration and time step.



(a) Correlation matrix for equicorrelated knockoffs



(b) Correlation matrix for SDP knockoffs

Figure 6: Correlation matrices $[X\tilde{X}]^T[X\tilde{X}]$ for a polynomial library of order 4 on the Lorenz system for SDP and equicorrelated knockoffs

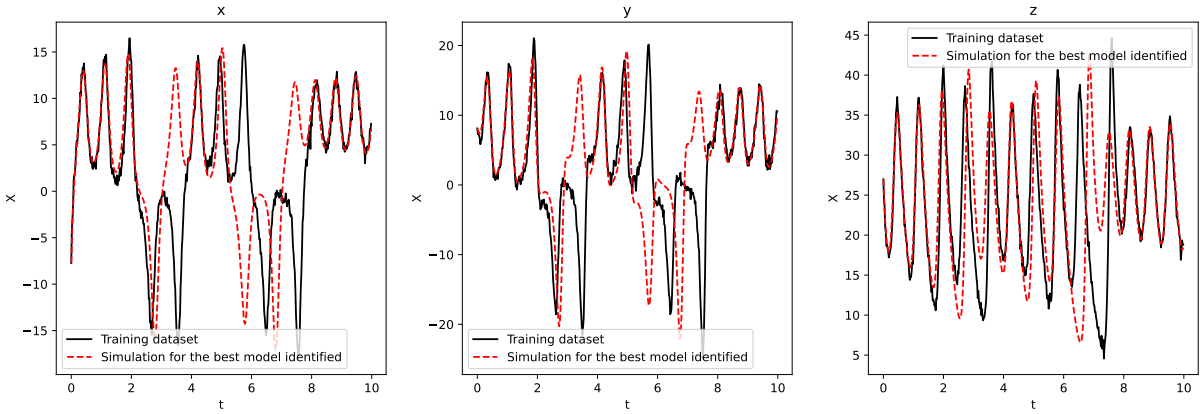


Figure 5: Simulation of the identified model, and comparisons with the training dataset. The simulation is given the same starting point as the training dataset and simulated for the same duration and time step.

3.2.1 Building knockoffs and challenges

Once we have our dictionary matrix, we can build the knockoff using the equicorrelated or the SDP procedure. We then plot the pairs $(Z_j, \tilde{Z}_j)$ (figure 7 and 8). As we can see many points stand on the $x = y$ line, meaning that the knockoffs enter the model at the samte time as their original counterpart on the lasso path. In our case it is because the knockoffs and the original variable are the same vector. Obviously this will prevent the knockoff procedure to give us reliable results for variable selection as you can see in the next part.
One explanation comes from the property of the original covariance matrix and the knockoff design. Remember that we have:

$$[X\tilde{X}]^T[X\tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - diag(s) \\ \Sigma - diag(s) & \Sigma \end{bmatrix}$$

The knockoff variable should be uncorrelated from its original variable but not from the other variable. In particular, if X_j is highly correlated with X_i, then $\tilde{X}_j$ should be highly correlated with X_i, and we can see that it will be harder for $\tilde{X}_j$ to be uncorrelated with its original variable X_j. We plot to covariance matrix of the augmented matrix $[X\tilde{X}]^T[X\tilde{X}]$ for the equicorrelated and SDP knockoffs (figure 6). One can compare these results with the random gaussian matrix one in the supplementary results to see why it is a drawback.

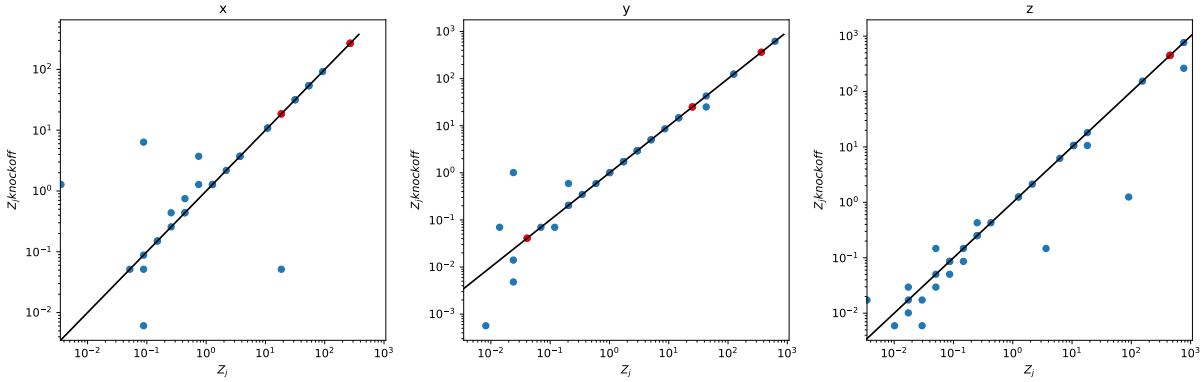


Figure 7: Equicorrelated knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for the polynomial 4 dictionnary matrix

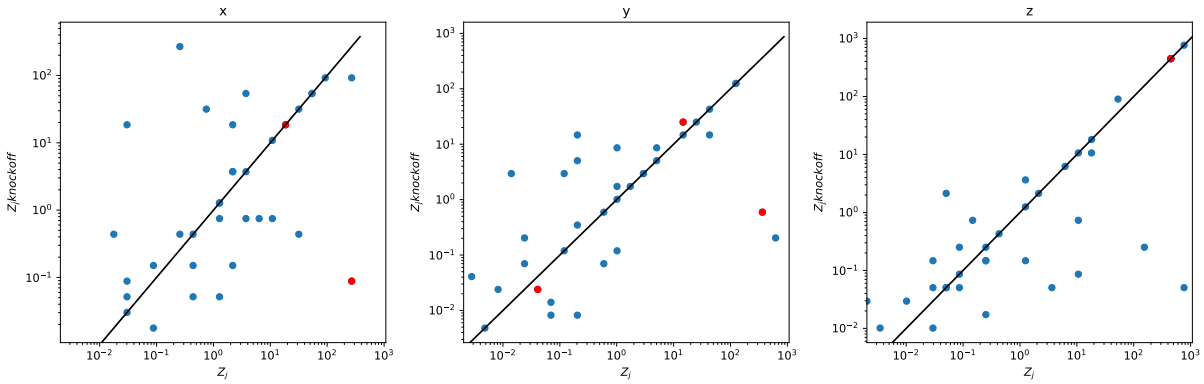


Figure 8: SDP knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for the polynomial 4 dictionnary matrix

From our experience with the application and for a future work, it would be interesting to think about a way to measure the power of the knockoffs related to the coherence of the matrix given a procedure to build knockoffs. We know for example that the coherence is linked to the eigenvalues using the Gershgorin's disk theorem [9], this would help to bound the equicorrelated knockoffs vector s, and then one may be able to measure the power of the knockoffs.

3.2.2 Type I and type II error consequences

Since the the knockoffs were build very poorly, the procedure to control the FDR spans very uninteresting results, this shows a drawback of the method as it can be seen in figure 9 and 10.

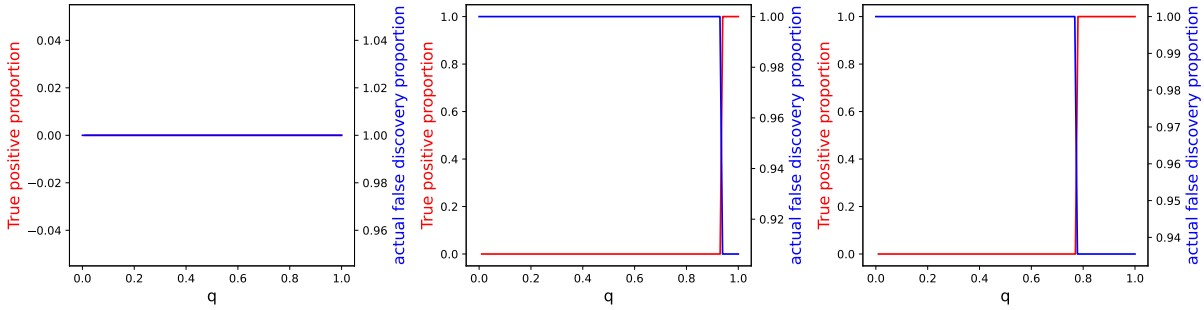


Figure 9: Evolution of the True Positive Proportion and the actual False Discovery Proportion using the equicorrelated knockoff procedure for the three different dimensions (x on the left, y on the middle and z on the right)

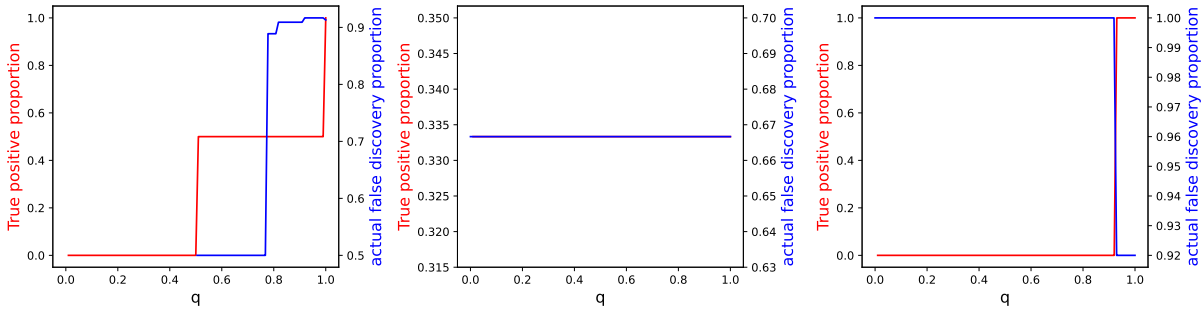


Figure 10: Evolution of the True Positive Proportion and the actual False Discovery Proportion using the SDP knockoff procedure for the three different dimensions (x on the left, y on the middle and z on the right)

3.3 Limits for its use in dynamic systems identification

The applications of knockoffs have been tested in particular to discover the genes involved in certain diseases [1]. This can be justified because it seems important not to target genes that do not cause the disease in order to avoid the toxicity of treatments, so it seems that this is the most important value to control. In the case of the identification of dynamic systems, and more particularly of non-linear functions, if we want to make good predictions, we need to recover the support exactly. We would then need to study tools to guarantee the recovery support of the vector. Results show that using only the lasso will result in a compromise between the TPP and the FDR [13], so it may not be the ideal framework. Further investigations will be conducted to study more closely the guarantee of these two conditions, but also to what extent false negatives or false positives differentiate the model from the exact system in the case of polynomial functions.

4 Conclusion

In this study, we have unrolled the theory for someone who would like to understand the ins and outs. Then we illustrate with two applications, the identification of dynamical systems and random gaussian matrices.
The knockoffs framework is a clever and innovative statistical one that has been and will

be further developed in other papers. It can be used in many framework as this area of statistical learning expands in many area of science (biology, engineering...). In particular further developments include different ways to build knockoff that may changed on the context (type of model, high dimensions...), assessing quality of the original matrix to build consistent knockoffs, controling the True Positive Proportion.

References

[1] J. -M. Azaïs and Y. de Castro. Multiple Testing and Variable Selection along Least Angle Regression's path, 2019. arXiv: 1906.12072. URL: http://arxiv.org/abs/1906.12072.

[2] Rina Foygel Barber and Emmanuel J. Candés. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015. ISSN: 21688966. DOI: 10.1214/15-AOS1337. arXiv: 1404.5609.

[3] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[4] Steven L. Brunton and J. Nathan Kutz. Data-Driven Science and Engineering. *Data-Driven Science and Engineering*, 2019. DOI: 10.1017/9781108380690.

[5] Steven L. Brunton, Joshua L. Proctor, J. Nathan Kutz, and William Bialek. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 113(15):3932–3937, 2016. ISSN: 10916490. DOI: 10.1073/pnas.1517384113. arXiv: 1509.03580.

[6] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(3):551–577, 2018. ISSN: 14679868. DOI: 10.1111/rssb.12265. arXiv: 1610.02351.

[7] Rick Chartrand. Numerical Differentiation of Noisy, Nonsmooth Data. *ISRN Applied Mathematics*, 2011:1–11, 2011. ISSN: 2090-5564. DOI: 10.5402/2011/164564.

[8] Brian de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Kutz, and Steven Brunton. PySINDy: A Python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, 5(49):2104, 2020. DOI: 10.21105/joss.02104. arXiv: 2004.08424.

[9] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, number 9780817649470. 2013, pages 1–615. ISBN: 9780817649470.

[10] Lennart Ljung. Perspectives on system identification. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 17(1 PART 1), 2008. ISSN: 14746670. DOI: 10.3182/20080706-5-KR-1001.4277.

[11] Johan Schoukens and Lennart Ljung. Nonlinear System Identification: A User-Oriented Road Map. *IEEE Control Systems*, 39(6):28–99, 2019. ISSN: 1941000X. DOI: 10.1109/MCS.2019.2938121. arXiv: 1902.00683.

[12] Steven H. Strogatz. *Nonlinear Dynamics and Chaos with Student Solutions Manual*. 2018. DOI: 10.1201/9780429399640.

[13] Weijie Su, Malgorzata Bogdan, and Emmanuel Candès. False discoveries occur early on the lasso path. *Annals of Statistics*, 45(5):2133–2150, 2017. ISSN: 00905364. DOI: 10.1214/16-AOS1521. arXiv: 1511.01957.

[14] Linan Zhang and Hayden Schaeffer. On the convergence of the sindy algorithm. *Multiscale Modeling and Simulation*, 17(3):948–972, 2019. ISSN: 15403467. DOI: 10.1137/18M1189828. arXiv: 1805.06445.

5 Supplementary information and simulations

5.1 Original method using ℓ_0 regularization

The lasso is not used on the original paper, instead a ℓ_0 regularized problem is used:

$$\hat{x} = \underset{x \in \mathbb{R}^n}{\arg\min} \|Ax - b\|_2^2 + \lambda^2 \|x\|_0$$

To solve the problem they use the STLSQ (Sequentially thresholded least squares algorithm) that is guaranteed to converge toward a local minimum [14]. The steps are computed as follow:

$$\begin{aligned} x^0 &= A^\dagger b \\ S^k &= \left\{ j \in [n] : |x_j^k| \geq \lambda \right\}, k \geq 0 \\ x^{k+1} &= \underset{x \in \mathbb{R}^n, supp(x) \subseteq S^k}{\arg\min} \|Ax - b\|_2 k \geq 0 \end{aligned}$$

Instead of computing a lasso path using the LARS method, we do a λ threshold scanning, then we do a cross correlation as before. There are a few differences: first the sparsity level and cross validation is made on the three dimensions at the same time while they were done on each dimension independently in when we used the lasso, then we can also cross validates by computing the whole trajectory of X since we evaluate the three functions corresponding to the three dimensions at the same time. We obtain similar results than with the lasso in the main text.

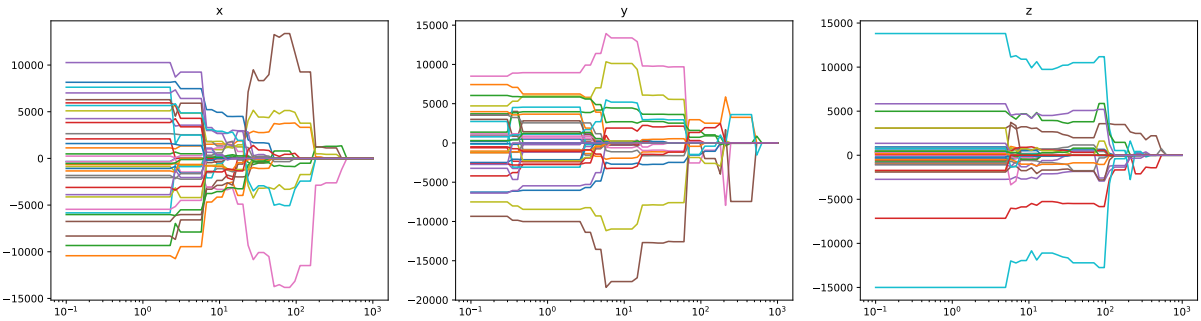


Figure 11: Pseudo path for the three dimensions using a λ threshold scan for $\lambda \in [10^{-1}, 10^3]$

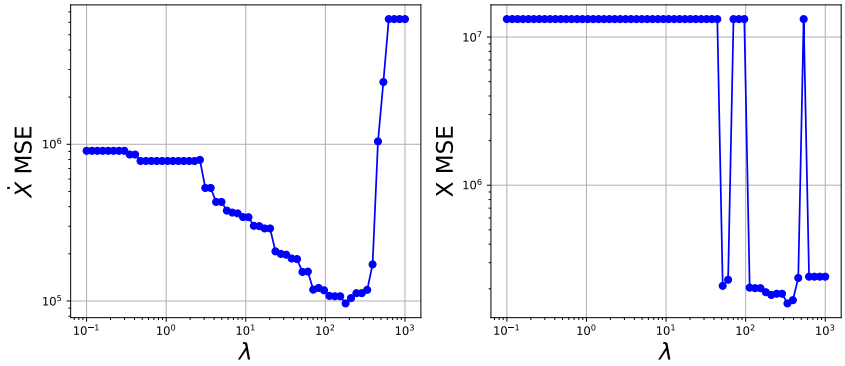


Figure 12: Cross-validation on the computed derivative testing set. The cross-validation is independent for the three dimensions and can lead to different optimal λ

The advantage to use the full simulation is from the fact that the model slightly deviates it may be a problem because the system is chaotic, when taking the full simulation we avoid this by comparing the whole simulation. We obtain the following equations:

$$\begin{cases} \dot{x} = 6.424y - 0.485xz + 0.083yz + 0.006xz^2 \\ \dot{y} = 18.610x - 0.253y + -0.040xz^2 + 0.001xz^3 \\ \dot{z} = -2.418z + 0.181x^2 + 0.769xy + 0.055y^2 - 0.011z^2 \end{cases}$$

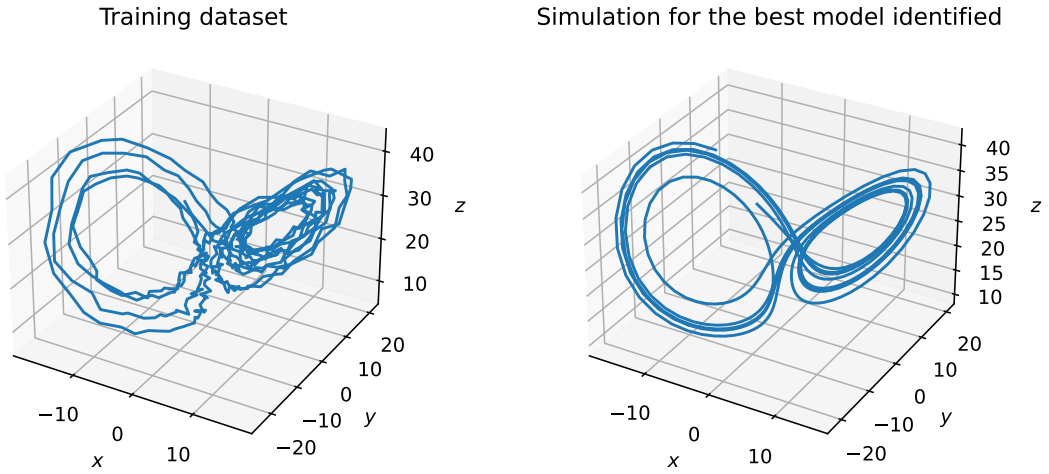


Figure 13: Simulation of the identified model, and comparisons with the training dataset in three dimensions. The simulation is given the same starting point as the training dataset and simulated for the same duration and time step.

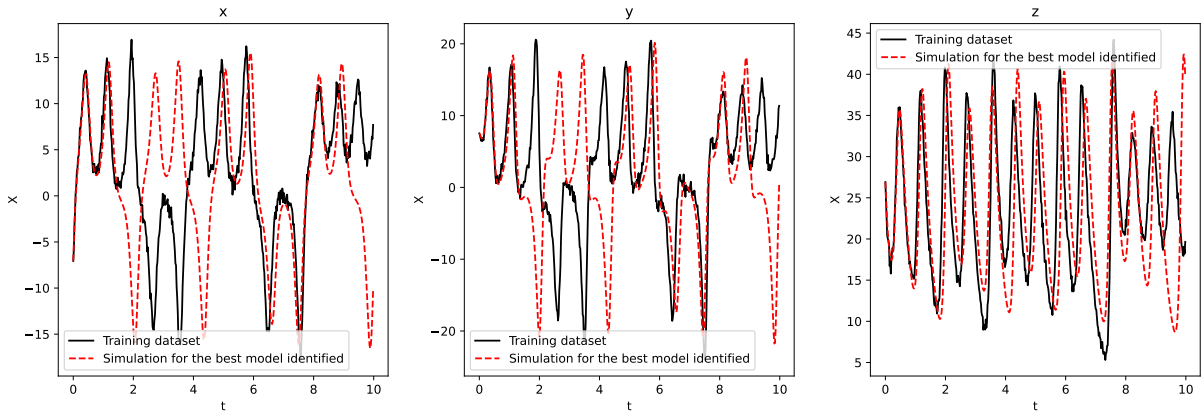


Figure 14: Simulation of the identified model, and comparisons with the training dataset. The simulation is given the same starting point as the training dataset and simulated for the same duration and time step.

5.2 Unpublished simulations

Other numerical changes as scaling the system by dividing the range by 10, 30 and 40 which didn't change much qualitatively the results. Simulations were ran without any noise, recovery of the true equations but not perfect, running the knockoffs was similar.

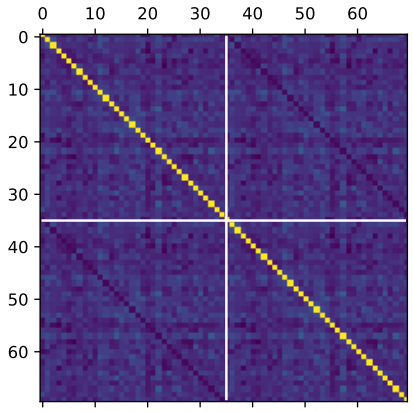
5.3 Knockoffs on random matrices

To compare the application to system identification using a dictionary matrix, whose columns are not independant since they depend on the same variables, we test here on random Gaussian matrices. We take the same dimensions as in the study: $n = 500$, $p = 35$ with each entry of the matrix independently following a normal distribution $\mathcal{N}(0, 1)$. We then pick a vector β with 3 nonnull entries at random positions. These elements follow independently a uniform law of parameters $[-5, 5]$. We then compute $y = X\beta + \eta$ with η a random gaussian vector of i.i.d. entries following a normal law of standard deviation 0.1 $\mathcal{N}(0, 0.1^2)$

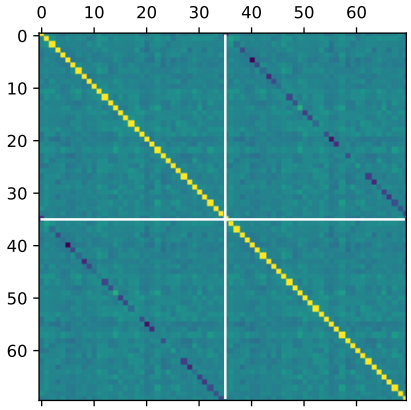
We can observe that the augmented matrix covariance looks way better than in our dictionary of polynomial (figure 15), in particular, variables are not highly correlated, which allows to build decent knockoffs, we can see that SDP knockoffs are slightly more uncorrelated compared to the equucorrelated knockoffs. This allows to have very few knockoff that are too correlated from their original variable (figure 17). And plotting the evolution of the FDR and the TPP we have a better consistence of the knockoff procedure (figure 16)

5.4 Removing correlated variables

We saw that one of the flaw of using the knockoffs framework on the system identification problem in the case of the Lorenz system was the matrix condition. In particular, we can compare the coherence of the two matrices. For the Gaussian matrix, we have a coherence $\mu = 0.797$ against a coherence of $\mu = 0.989$ for the polynomial dictionary matrix. A simple first solution would be to remove the very correlated variables. For pairs of variables whose correlation was above 0.94, the variable corresponding to the

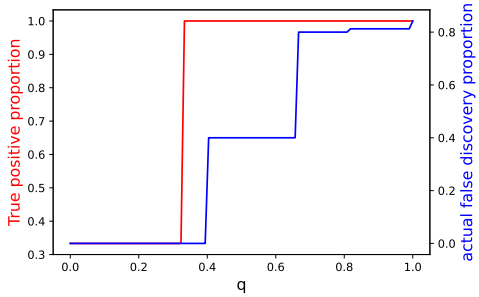


(a) Correlation matrix for equicorrelated knockoffs

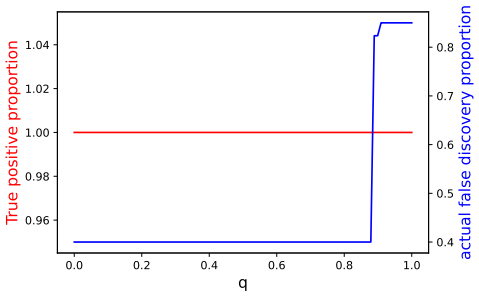


(b) Correlation matrix for SDP knockoffs

Figure 15: Correlation matrices $[X\tilde{X}]^T[X\tilde{X}]$ for a gaussian random matrices for SDP and equicorrelated knockoffs

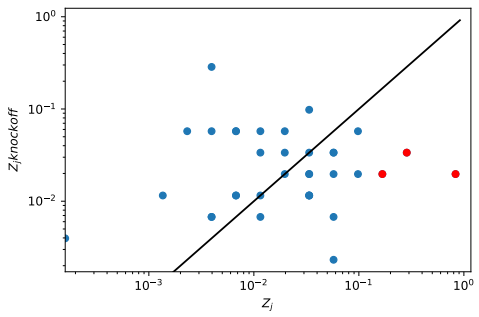


(a) Evolution of the True Positive Proportion and the actual False Discovery Proportion using the equicorrelated knockoff procedure

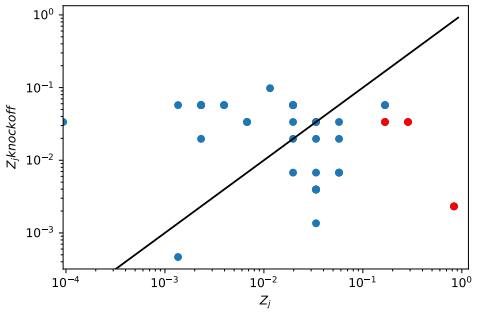


(b) Evolution of the True Positive Proportion and the actual False Discovery Proportion using the SDP knockoff procedure

Figure 16: Evolution of the True Positive Proportion and the actual False Discovery Proportion using the knockoff procedure with a varying parameter q

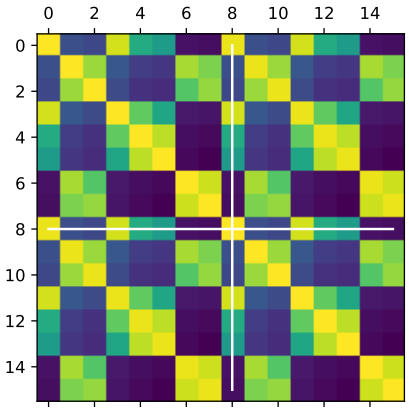


(a) equicorrelated knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for

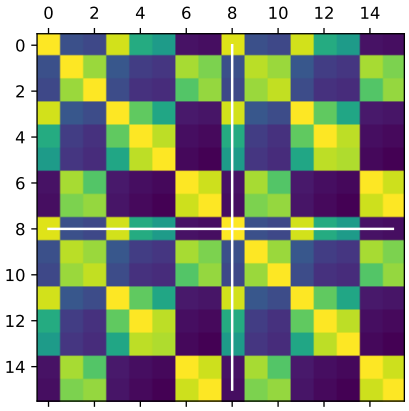


(b) SDP knockoff plotting pairs $(Z_j, \tilde{Z}_j)$

Figure 17: Knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for the gaussian random matrix case



(a) Correlation matrix for equicorrelated knockoffs



(b) Correlation matrix for SDP knockoffs

Figure 18: Correlation matrices $[X\tilde{X}]^T[X\tilde{X}]$ for a gaussian random matrices for SDP and equicorrelated knockoffs

higher polynomial order was removed. For thresholds higher than 0.94, nothing changed drastically, it changes when we remove at 0.94 or lower, however we lose a lot of variables, including one of the true variable (in particular, the variable x is very correlated with xz). The results are the following (figures 18, 21, 22, 19, 20). Even though the results look better thant the first attempts, the number of variables being reduced drastically, the method looses its interest since we want to look in a potentially big dictionary when little is known of the system.

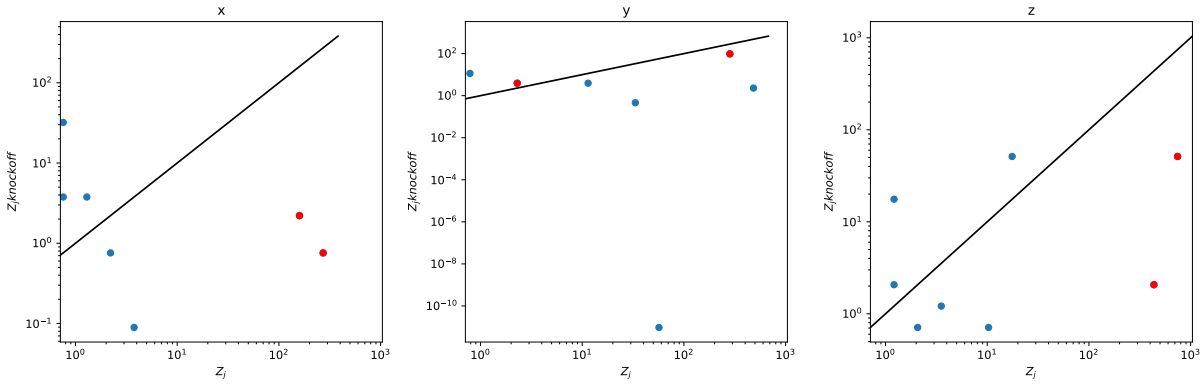


Figure 19: Equicorrelated knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for the filtered dictionnary matrix

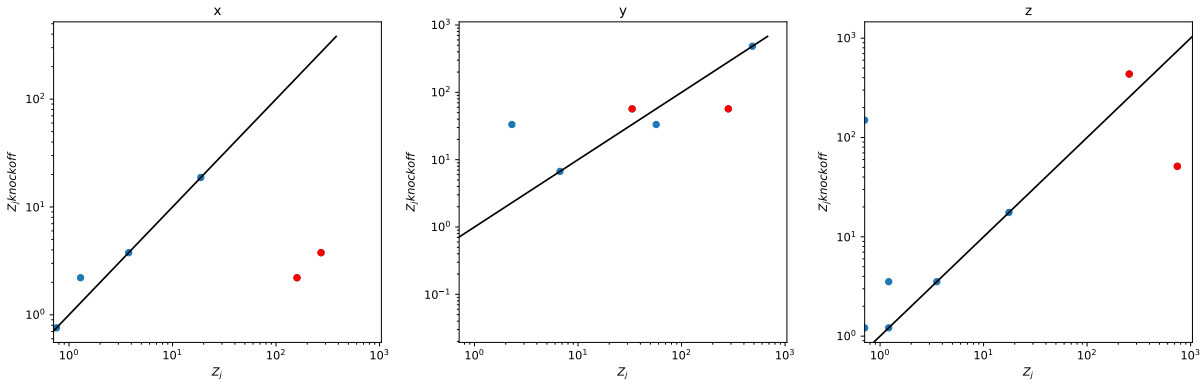


Figure 20: SDP knockoff plotting pairs $(Z_j, \tilde{Z}_j)$ for the filtered dictionnary matrix

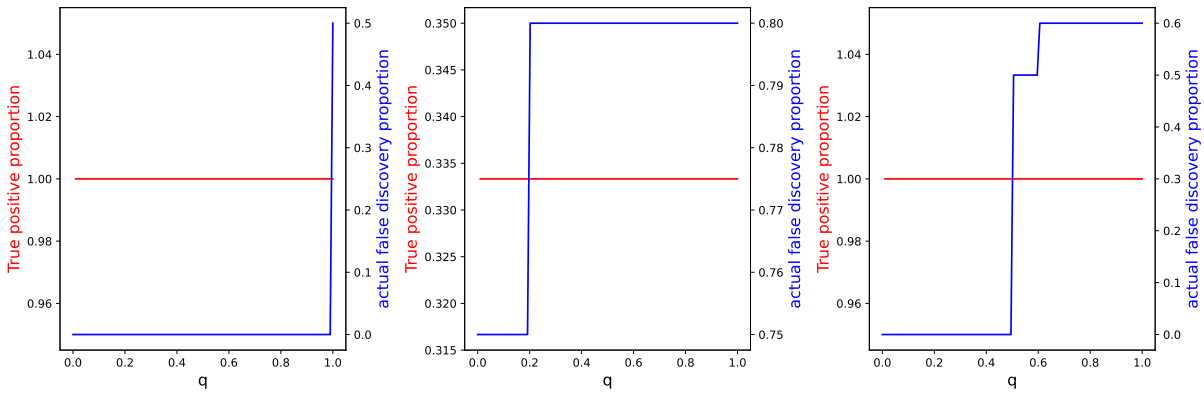


Figure 21: Evolution of the True Positive Proportion and the actual False Discovery Proportion using the equicorrelated knockoff procedure for the three different dimensions (x on the left, y on the middle and z on the right)

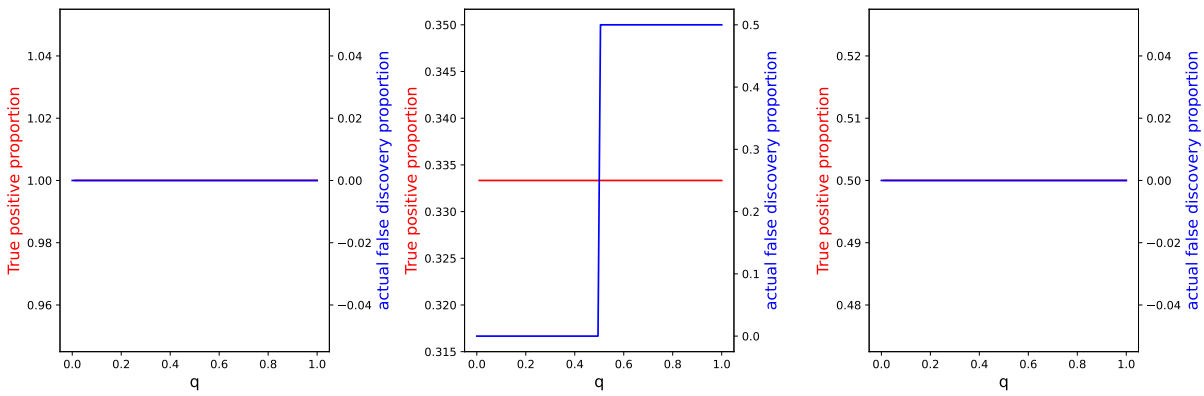


Figure 22: Evolution of the True Positive Proportion and the actual False Discovery Proportion using the SDP knockoff procedure for the three different dimensions (x on the left, y on the middle and z on the right)

5.5 Proof of *Lemma 2.1*

Proof. Let $S \subset [p]$, and let's pose $\left[X\tilde{X}\right]_{swap(S)}$ the matrix $\left[X\tilde{X}\right]$ so that $\forall j \in S,\ X_j$ and $\tilde{X}_j$ are swapped. We first note that $\left[X\tilde{X}\right]^T_{swap(S)}\left[X\tilde{X}\right]_{swap(S)} = \left[X\tilde{X}\right]^T\left[X\tilde{X}\right]$: it directly comes from the covariance structure properties respected by $\tilde{X}$.

We first prove that if $\forall j \in S,\ \beta_j = 0$, then the equality in law $\left[X\tilde{X}\right]^T_{swap(S)} y \overset{d}{=} \left[X\tilde{X}\right] y$ holds.

Since y follows a Normal law $N(\ X\beta, \sigma^2 I)$ $\left[X\tilde{X}\right]^T_{swap(S)} y$ follows a Normal law $N(\ \left[X\tilde{X}\right]^T_{swap(S)} X\beta, \sigma^2 \left[X\tilde{X}\right]^T_{swap(S)}\left[X\tilde{X}\right]_{swap(S)})$.

We have $\left[X\tilde{X}\right]^T_{swap(S)}\left[X\tilde{X}\right]_{swap(S)} = \left[X\tilde{X}\right]^T\left[X\tilde{X}\right]$, hence the equality of the variances.

Since $\tilde{X}X = \Sigma - diag(s),\ \forall i \neq j,\ X\tilde{X}_j^T X_i = X_j^T X_i$, and since $supp(\beta) \cap S =,\ \forall j \in S,\ \tilde{X}_j^T X\beta = X_j^T X\beta$.

Then $\left[X\tilde{X}\right]^T_{swap(S)} X\beta = \left[X\tilde{X}\right]^T X\beta$, and the equality in law follows.

$$\text{Let } \epsilon_j = \begin{cases} -1 \text{ if } j \in S \\ 1 \text{ else} \end{cases} \quad \text{then}$$

Then, by antisymmetry property $W_{swap(S)} = (W_1\epsilon_1, ..., W_p\epsilon_p)$.

We now consider $\epsilon \in \{\pm 1\}^p$ s.t. $\epsilon_j = 1$ if $\beta_j = 0$, $\epsilon_j \in \{\pm 1\}$ else. With $S = \{j/\epsilon_j = -1\}$, we have $S \cap supp(\beta) =$ and it comes, by sufficiency property of W:

$$\begin{aligned} W_{swap(S)} &\overset{d}{=} f(\ \left[X\tilde{X}\right]^T_{swap(S)}\left[X\tilde{X}\right]_{swap(S)},\ \left[X\tilde{X}\right]^T_{swap(S)} y) \\ &\overset{d}{=} f(\ \left[X\tilde{X}\right]^T\left[X\tilde{X}\right],\ \left[X\tilde{X}\right]^T_{swap(S)} y) \\ &\overset{d}{=} f(\ \left[X\tilde{X}\right]^T\left[X\tilde{X}\right],\ \left[X\tilde{X}\right]^T y) \overset{d}{=} W \end{aligned}$$

□