Social Sciences & Humanities Open 11 (2025) 101380

Contents lists available at ScienceDirect

Social Sciences & Humanities Open

journal homepage: www.sciencedirect.com/journal/social-sciences-and-humanities-open

Regular Article

# Model-X knockoffs in the replication crisis era: Reducing false discoveries and researcher bias in social science research

Jing Zhou [a], Sebastian Scherr [b,*]

[a] School of Engineering, Mathematics and Physics, University of East Anglia, United Kingdom
[b] Center for Interdisciplinary Health Research & Department of Media, Knowledge, and Communication, University of Augsburg, Germany

ARTICLE INFO

*Keywords:*
Multivariate linear regression
Variable selection
Model-X knockoffs
Simulation study
Big data
High dimensional statistics
True model

ABSTRACT

The present study addresses problems faced by data-driven social science caused by having too much or not enough data. In particular, an abundance of data or a (sudden) lack thereof makes it challenging to identify the most important predictors in a sea of noise using the most parsimonious and reproducible model possible. In this article, we present the model-X knockoff method, which was introduced by Candès et al. (2018) for reducing the false identification of significant effects due to flexibility-ambiguity issues, to a broader audience, particularly within the social sciences and humanities. Our goal is to provide an accessible starting point and ideally spark interest among researchers in these fields to explore how model-X knockoffs can enhance their work. The findings from a performance contrast simulation indicate that model-X knockoffs select fewer relevant variables than other statistical methods to automatically identify variables, resulting in fewer mistakes. The simulation findings also demonstrate that model-X knockoffs are stable and less sensitive to even small changes in the dataset than other procedures, making them a viable way to reduce researcher degrees of freedom and increase the reproducibility of scientific findings. An additional real data example demonstrates the operational utility of the simulation.

Data-driven science simultaneously suffers from two problems: having too much data and not having enough. Big data, digital trace data, ecological momentary assessment data, sensory user data, and data from social media application programming interfaces (APIs) are either available in abundancy or access is limited or suddenly shut off, creating a research *APIcalypse* (Bruns, 2019) by not having access to data for research. However, what about the abundancy side of the problem? What if there are more predictors ($p$) than observations ($n$) for a given social phenomenon? Such high dimensional data scenarios (i.e., $p > n$ settings) provide researchers with much more room for making decisions when analyzing data, which is more likely than not going to challenge the analysis of large-scale datasets in the future. Importantly, the number of available predictors might dramatically outgrow the discovery of new social science phenomena, bringing attention to the challenge of identifying "important predictors in a sea of noise" (Candès et al., 2018, p. 552). The aim of this paper is to introduce model-X knockoffs as a new and viable strategy for reducing researcher degrees of freedom and improving variable selection in high-dimensional datasets. Specifically, we will address the challenge of selecting variables in large datasets where the number of predictors exceeds the number of observations, while controlling for false discovery rates, especially when the available data is dynamic and constantly changing.

When it is easy and relatively cheap to access an abundance of predictors, a researcher will be even more likely to falsely find evidence supporting the existence of an effect than correctly find evidence that the effect does not exist. This tendency occurs due to "researcher degrees of freedom" (Simmons et al., 2011, p. 1359), which is a flexibility-ambiguity problem that involves important questions about whether more data should be collected, decisions about inclusion and exclusion criteria for observations, and the selection of relevant model covariates. Similarly, Ioannidis (2005) highlights the relevance of the flexibility of designs, definitions, outcomes, and analytical models, which leads to more false positive findings especially when a study is conducted in a smaller field, when expected effect sizes are smaller, and when more relationships between variables are tested. In a similar vein, The Open Science Foundation (2015) emphasized the importance of addressing researcher degrees of freedom in their large-scale replication project, highlighting the threats they pose to the reproducibility of scientific research. In fact, researchers should disclose some of their decisions before conducting research in light of an increasing awareness of

* Corresponding author.
*E-mail address:* sebastian.scherr@uni-a.de (S. Scherr).

https://doi.org/10.1016/j.ssaho.2025.101380
Received 13 December 2023; Received in revised form 3 February 2025; Accepted 24 February 2025
Available online 28 February 2025

open science practices (Dienlin et al., 2021) and disclosure-based, intersubjectively replicable science (The Open Science Foundation, 2015). Specifically, and in line with the journal's scope, model-X knockoffs can enhance the replicability of studies by reducing false positives, clarifying whether a variable's original significance was genuine or the result of chance or bias. Combined with other techniques, such as directed acyclic graphs (DAGs; see Pearl, 2009), they can help identify variables that initially appear to be significant predictors in one study but fail to replicate under different conditions or in new data samples. By providing rigorous control over false positives, model-X knockoffs contribute to the development of leaner, more reliable models when re-analyzing data from earlier experiments or applying the same analytical framework to new datasets.

However, there are at least three aspects of data-driven science in which open science best practice recommendations are hard or even impossible to accomplish (e.g., Simmons et al., 2011). First, researcher may not be able to list all available variables (e.g., when the data stems from a social media API). There might simply be too many available variables and data points. Moreover, at any given time, an entity can revoke access to its API or shut down access to certain types of content (e.g., Pfeffer et al., 2022). Second, social media companies are not only constantly running A/B testing for different interfaces and user experiences but also regularly running large-scale experiments, unbeknownst to users (e.g., Bond et al., 2012; Kramer et al., 2014; Rajkumar et al., 2022), which affect the data accessible through their APIs. It is virtually impossible for researchers to determine all of the conditions under which the data from an API was gathered. Third, it is impossible to know whether the owner granting access to the API eliminated any observations before the data became publicly accessible.

Nonetheless, there is good news. In this article, we shed light on a statistical approach to tackling the researcher degrees of freedom issue and the production of false-positive findings in large datasets. Using social media platforms such as X, formerly Twitter, as a popular example of an *academic API* (Pfeffer et al., 2022), we first introduce the concept of the statistical *true model* as a preface to our discussion of the statistical "knockoff method" (Candès et al., 2018). Then, we discuss the simulation we conducted in the present study and show how the knockoff method is suitable for building interpretable models under big data conditions when inference is drawn from finite samples with arbitrary or unknown response distributions. A real data application is also provided to showcase the practical value of the method.

## 1. The (unknown) true model

Modern statistical textbooks sometimes refer to a concept called the "true model" (Gelman & Hill, 2007, p. 47), which is an abstract underlying mathematical model that defines how observed data is originally generated. It describes a formula for the underlying data generating process. In this sense, the true model is the most parsimonious model because it only includes the variables necessary to explain and predict a set of observations at a given point in time. The true model is, therefore, the best of all possible models that try to explain and predict these observations. By definition, all alternative models that are not the true model include too many (overfitting) or too few (underfitting) predictor variables.

The true model focuses on data aggregation. The model changes based on the level of data aggregation under observation (e.g., hourly, daily, weekly, local, regional, cross-country). Researchers have even gone so far as to challenge the mere existence of a true model because all possible models are ultimately wrong (e.g., Perretti et al., 2013). Importantly, the true model can only be known beforehand if all theoretical premises can be assumed or if the data is simulated and follows specific, predefined rules. When working with aggregated, observed data, one cannot know the true model; however, all models that can be identified will only get close to the true model. Even though a model fits the data quite well or is almost as good as the true model, there will always be hidden unknown variables or variables that are simply not observed, complicating any inferences made from the observations.

Importantly, there are variable selection techniques that help to approximate the True Model by identifying relevant variables. Among the more popular techniques are the Lasso method and model-X knockoffs. While the Lasso method is particularly suitable for scenarios in which the number of variables exceeds the number of observations (Tibshirani, 1996), model-X knockoffs are a better choice for settings where controlling the false discovery rate is crucial and producing more parsimonious results, i.e., identifying *fewer* variables that are relevant to describe the true model (Candès et al., 2018).

Let us consider a hypothetical example involving a theory called "platform theory." This theory encompasses a comprehensive set of rules that explains every aspect of how a social media platform operates, including how content is presented to its users and the effects of this content on users around the world at any given moment. Through the platform's API, a researcher can observe the entire population of content on the platform but is usually restricted to drawing only a sample. This sample may be limited by factors such as the number of items, a specific time period, or geographical location. Additionally, even if the data request parameters remain constant, the sample can vary each time it is downloaded. In other words, while the population of all content remains constant, the samples themselves are dynamically changing constantly. Consequently, the researcher cannot know the true model of how the platform operates or effectively test the "platform theory." Furthermore, the variability in sample data not only means that observed samples will differ from the unknown true model but also increases the likelihood of obtaining false positive findings when testing specific hypotheses about the platform. In such a dynamic context, researchers are faced with an overwhelming number of decisions, which complicates the determination of whether an analysis is accurate and how close it is to the true model. To reduce the researcher degrees of freedom in selecting relevant predictors, the model-X knockoff method (Candès et al., 2018) is a valuable tool.

## 2. Method

### 2.1. The knockoff method: a primer

Technological advancements have not only drastically reduced the cost of data collection but also made data collection more versatile. Large-scale datasets tend to contain a massive collection of predictor variables $X = (X_1, \ldots, X_p)$ that often exceeds the number of observations $n$. However, only a small fraction of variables are relevant for predicting the variable of interest (i.e., the outcome variable $Y$). In this case, variable selection becomes crucial for further model interpretation and alleviating the over-/underfitting problem of having included too many or too few predictors.

Two measures are crucially tied to variable selection: control of the false discovery rate (FDR) and guaranteeing enough statistical power. The following formulas were used in the present study:

$$\mathrm{FDR} = \mathrm{E}\left[\frac{\text{the number of selected irrelevant variables}}{\text{the number of selected variables}}\right]$$

$$\mathrm{Power} = \mathrm{E}\left[\frac{\text{the number of selected relevant variables}}{\text{the number of relevant variables}}\right]$$

Variable selection should aim to control FDR under a prespecified level $q$ (i.e., the wrongfully selected irrelevant variables take no more than $(q \times 100)\%$ of the total number of selected variables). At the same time, a high selection power is desired; therefore, as many relevant variables as possible should be selected. However, predictors are often correlated with the outcome in a linear or nonlinear fashion, providing leeway for the development of machine learning procedures to identify relevant predictors (e.g., Scherr & Zhou, 2020). For the aforementioned example about a "platform theory," we used model-X knockoffs as a variable selection method that controls FDR.

First, we introduce *conditional independence*, which defines all null or irrelevant variables. In the regression setting, conditional independence holds for irrelevant variables that are redundant in explaining the outcome variable given the relevant ones. A variable $X_j$ is a null variable if and only if the response variable $Y$ is independent of $X_j$ and conditional on all other random variables $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$. For generalized linear models in which the response variable $Y$ depends on $X_j$ as a linear combination $\eta = \beta_1 X_1 + \ldots + \beta_p X_p$, a variable $X_j$ is irrelevant to $Y$ if and only if the coefficient $\beta_j = 0$. We then performed the model-X knockoff variable selection by constructing a new family of $X$ distributional knockoffs $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$ that satisfied the following properties as defined by Candès et al. (2018):

(1) (pairwise exchangeability) the joint distribution of the original variables $X$ and the corresponding knockoff variables $\widetilde{X}$ denoted by $(X, \widetilde{X})$ remains the same if a subset of variables is swapped with the corresponding knockoffs, e.g., $(X_1, X_2, X_3, \widetilde{X}_1, \widetilde{X}_2, \widetilde{X}_3)$ follows the same distribution as $(\widetilde{X}_1, \widetilde{X}_2, X_3, X_1, X_2, \widetilde{X}_3)$. In other words, the correlation between *knockoff variable $X_1$* and *knockoff variable $X_2$* is the same as the correlation between the *original variable $X_1$ and the original variable $X_2$.*

(2) (independence) $\widetilde{X}$ is independent of $Y$ given $X$, i.e., the knockoffs are irrelevant to the response of interest. That means that knockoffs are developed without information about the outcome variable(s).

Property (1) suggests that "knockoffs" mimic the original variables in their distribution and can partially explain the variability of the outcome variable. Property (2) guarantees that the knockoffs are irrelevant to $Y$. These two properties collectively provide the underlying intuition of the approach: since the knockoffs are irrelevant to *Y*, they would have null coefficients in the true model. Consequently, the relevant variables are expected to exhibit significantly different coefficients compared to their knockoffs. Property (1) is particularly meaningful for null variables, where both the original variables and their knockoffs will have zero or near-zero estimated coefficients. This property is utilized to introduce the "flip-sign" property of the score statistic (Candès et al., 2018) and is further used to derive an estimator of the false discovery rate in Equation (2).

Constructing knockoffs satisfying the above two properties is critical for guaranteeing that the variable selection has enough statistical power. To extend constructing knockoffs for Gaussian distributed variables to more general cases, Candès et al. (2018) developed two strategies for knockoff construction: 1) exact and 2) approximate construction strategies for non-Gaussian distributed continuous variables.

Exact construction generates the knockoff variables using the conditional distribution of the corresponding original variables given all other variables, which simultaneously approximates the original variable in the distribution and minimizes the correlation with the other variables. A *Sequential Conditional Independent Pairs* algorithm for exact construction was also proposed in Algorithm 1 in Candès et al. (2018). Approximate construction requires the first two moments—the mean and variance of $(X, \widetilde{X})$—to remain unchanged if a set of variables in $X$ is swapped with its knockoffs in $\widetilde{X}$. Therefore, adequately constructed knockoffs show a pairwise exchangeability of nulls, suggesting that the conditional distribution of $(X, \widetilde{X})$ on $Y$ does not change if the nulls (irrelevant predictors) are swapped with their knockoffs. Both strategies have been implemented in the R package *knockoff* and have recently sparked broader interest in knockoff methods and derivates of this method (e.g., Barber et al., 2020).

Generating knockoffs is being further developed, but this falls outside the scope of the current paper. Innovations include the metropolized knockoff sampling (Bates et al., 2021) and deep knockoff (Romano et al., 2020), which are both contributions to adapting the method to general continuous distributions. Specifically, Romano et al. (2020) discuss the importance of property (1) and suggested several goodness-of-fit diagnostics for testing if property (1) holds in practice. The proposal tests are based on testing if the joint distribution of $(X, \widetilde{X})$ remain unchanged after swapping either all of them or randomly selected subset of them. The results have been implemented in Python and have been made available on their GitHub page [weblink].

Under the assumption of pairwise exchangeability of nulls, variable selection is performed by contrasting the importance of each pair $(X_j, \widetilde{X}_j)$ using a score statistic $W_j$ for $j = 1, \ldots, n$, which satisfies a "flip-sign" property. In other words, $W_j$ adds a negative sign when swapping $X_j$ and $\widetilde{X}_j$ for relevant variables but remains unchanged for irrelevant ones.

For example, we estimated the linear correlation between $Y$ and $(X, \widetilde{X})$ as follows:

$$Y = (X, \widetilde{X})\beta + \varepsilon$$

In addition, we used the following Lasso estimator:

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \quad \left\| Y - \left(X, \widetilde{X}\right)\beta \right\|_2^2 + \lambda \left\| \beta \right\|_1 .$$

A typical choice of score statistic follows $W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}|$, which compares the relevance of $X_j$ and $\widetilde{X}_j$ by contrasting the magnitude of the coefficient estimators for the original variable and the corresponding knockoff variable. A variable $X_j$ is selected if $W_j$ exceeds certain threshold $t$, i.e., the relative relevance of $X_j$ compared to that of its knockoff variable $\widetilde{X}_j$ that reaches a certain threshold $t$. The threshold $t$ is data dependent and can be chosen based on Equation (2).

Importantly, FDR is controlled when choosing the threshold $t$. Since $X_j$ and $\widetilde{X}_j$ are both irrelevant to $Y$, $\beta_j$ and $\beta_{j+p}$ are either both set to zero or one is randomly larger than the other as a result of estimation uncertainty. Consequently, for all irrelevant variables, the number of $W_j \geq t$ is approximately equal to that of $W_j \leq -t$. Thus, the adjusted FDR, which reflects the knockoff selection, is as follows:

$$\text{FDR} = \text{E}\left[\frac{\text{the number of irrelevant variables having } W_j \geq t}{\text{the number of } W_j \ \geq -t}\right] \tag{1}$$

Since the numerator is approximately equal to $W_j \leq -t$ for irrelevant variables, $W_j \leq -t$ happens when the knockoffs are estimated to be more important than their original counterparts. This can be expected to occur much more frequently in irrelevant variables. Therefore, we constructed an upward biased estimator of FDR as follows:

$$\widehat{\text{FDR}} = \text{E}\left[\frac{\text{the number of } W_j \leq -t}{\text{the number of } W_j \geq -t}\right] \tag{2}$$

This biased estimator $\widehat{\text{FDR}}$ enlarges the numerator of Equation (1) for the true FDR by searching through all variables in the numerator instead of only the irrelevant variables. This procedure is relatively accurate, especially for small FDR values (Weinstein et al., 2017; 2020; Zhou & Claeskens, 2022). Therefore, using the estimator $\widehat{\text{FDR}}$ as described above, we can choose a threshold $t$ when controlling the estimated $\widehat{\text{FDR}} \leq q$, with $q$ representing a preselected FDR level. The data-dependent value of $t$ is essentially determined by the prespecified value $q$, which is the main parameter of the model-X knockoff selection and controls the number of selected variables. Typically, the number of selected variables increases when we increase the level $q$, which lets more irrelevant variables be selected after the relevant ones are exhausted. Since one typically has no information on the relevance of each variable, the true FDR in Equation (1) cannot be estimated in practice. Therefore, we used an estimator $\widehat{\text{FDR}}$ for practical use. By estimating $\widehat{\text{FDR}}$ as in Equation (2), we can also estimate FDR as a function of the predetermined levels $q$ (i.e., the total number of selected variables) to realize FDR control.

### 2.2. Performance contrast simulation: the basics

The numerical performance of the model-X knockoff filter was determined in R using the package *knockoff*. To show the advantage of the knockoff filter in controlling FDR, we compared it to a widely used alternative variable selection method called the Lasso method (see Scherr & Zhou, 2020; Tibshirani, 1991). The Lasso is a regularization technique used to enhance the prediction accuracy and interpretability of statistical models by imposing a penalty on the absolute size of the coefficients. Specifically, the Lasso minimizes the residual sum of squares subject to a constraint on the sum of the absolute values of the coefficients. A key feature of the Lasso is the regularization parameter ($\lambda$), which critically affects model performance. A larger $\lambda$ increases the shrinkage level, resulting in sparser models with fewer non-zero coefficients but potentially more bias. Conversely, a smaller $\lambda$ yields a model with more variables and reduced bias but also increased variance. Consequently, the results of the Lasso model can vary significantly depending on the chosen value of $\lambda$.

The choice of the tuning parameter is a complex issue and is beyond the scope of this paper. In linear regression, the regularization parameter is typically chosen by minimizing the cross-validation prediction error, which ensures optimality in terms of prediction accuracy. We do not advocate adjusting the regularization parameter if it compromises the optimality of the predefined loss function, such as the commonly used cross-validation error. However, this optimality does not guarantee controlled false discoveries or optimal selection power. Controlled variable selection is not directly achievable through Lasso or similar regularized estimators. Model-X knockoffs, however, enhance the Lasso by providing controlled variable selection, which is particularly useful for finite samples. Thus, our focus is on comparing the controlled variable selection performance of the Lasso with the model-X knockoff, using the Lasso coefficient difference statistic.

We examined performance contrasts in two scenarios and simulated how variations in the number of variables and observations influence variable selection and model performance. In the first scenario, we simulated adding more variables to an existing dataset. In the second scenario, we simulated adding more observations to an existing dataset. We aimed to explore the extent to which even small changes to a dataset result in different or more irrelevant variables being incorrectly selected as relevant model predictors (i.e., yielding a higher FDR or a high fluctuation of selected variables), and we contrasted this with another established method of variable selection.

### 2.3. Performance contrast simulation: specific adjustments for its application

We used a simulation to answer the following questions: How can a researcher best perform these two scenarios? What is the best strategy for selecting the best predictors in a given situation? What is the FDR when choosing a specific method of data selection?

We simulate two scenarios, in both of which we assumed that the "platform theory" was linearly associated with $k = 10$ variables, that is,

$$Y = X\beta + \varepsilon, \tag{3}$$

where $Y$ is a continuous response variable of interest such as users' time browsing posts; $X = (X_1, \ldots, X_p)^\top$ is the vector of explanatory variables such as users' content preferences, number of likes on each category of posts, etc.; $\varepsilon$ is a latent random error. To reflect the fact that 10 variables are associated with $Y$, we randomly set 10 components of $\beta$ to be nonzero and set the signal strength to be either $-3.5$ or 3.5. The rest of the components of $\beta$ are set to be zero such that the corresponding variables are irrelevant to $Y$. We vary the number of observations $n$ and the number of variables $p$ in the two examples. To address multicollinearity between the variables, we sample $X$ from a multivariate Gaussian distribution with the correlation between any two variables $i, j$ being $0.5^{|i-j|}$. The random error $\varepsilon$ is sampled from a standard Gaussian distribution. The following flowchart (see Fig. 1) is intended to illustrate the simulation procedure.

The first example aims to simulate a real life scenario in which we have more participants signing up for a social media platform after the initial data collection period, thereby creating a dataset with more observations. We reflect this by fixing the number of variables $p = 300$ and gradually increasing the number of observations from $n = 100$ to $n = 300$ with increments of 50 individuals.

In the second example, we assumed that the "platform theory" was linearly associated with 10 variables among $p = 110$, and we collected a sample of $n = 100$ observations to test that assumption. Over the course of the simulation, we still believed that these 10 variables were important. However, we also considered additional variables that were offered by the social media platform through its API after the initial data collection period and had not previously been accessible to the public. These variables potentially provided additional information on the "platform theory."

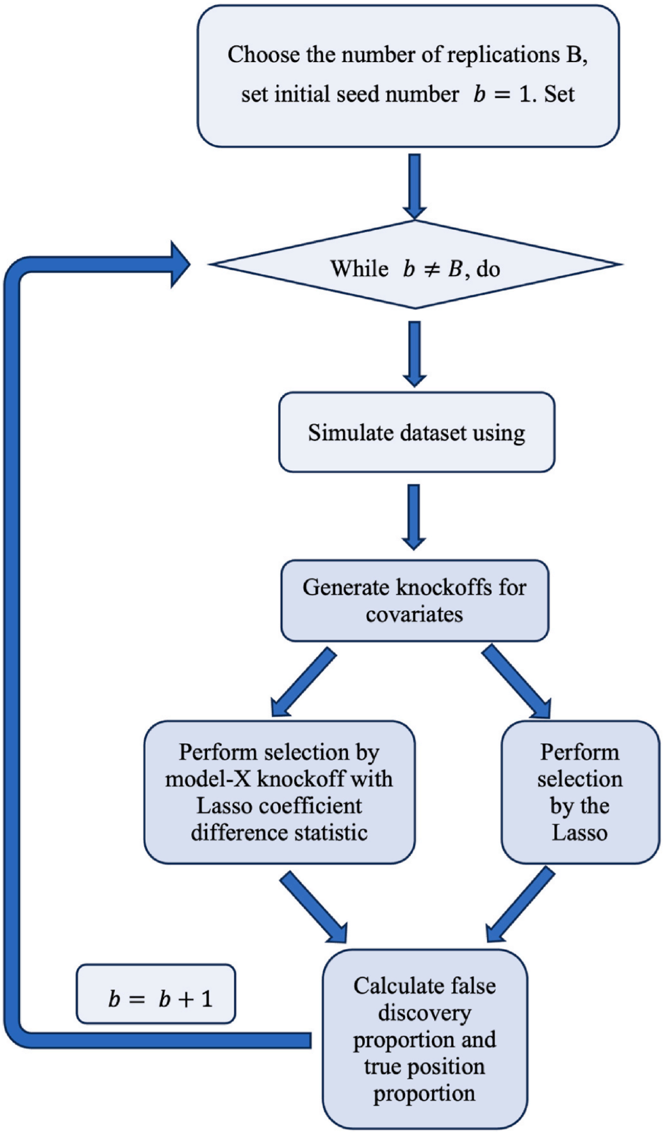
Choose the number of replications B, set initial seed number $b = 1$. Set
While $b \neq B$, do
Simulate dataset using
Generate knockoffs for covariates
Perform selection by model-X knockoff with Lasso coefficient difference statistic
Perform selection by the Lasso
Calculate false discovery proportion and true position proportion
$b = b + 1$

**Fig. 1.** Flowchart for Model-X knockoff procedure with lasso regression for simulating datasets and evaluating variable selection accuracy.

Optimal variable selection is important because using redundant variables typically causes overfitting problems and yields uncontrollable model variability. Nevertheless, theory-based variable selection is oftentimes not accurate, leaving the researcher with some wiggle room to decide which variables and information are used to test hypotheses that are frequently relatively vague. However, when choosing variables in large-scale datasets (e.g., X API) with thousands of possibly relevant predictors and observations, it is (and will be even more) crucial to support variable selection algorithmically based on data that simultaneously controls for falsely selecting irrelevant variables while increasing power and confidence in findings.

In the two examples in our simulation study, we demonstrated the ability to control a false discovery bias using the model-X knockoff method (Candès et al., 2018). Within each simulation, we validated our findings using identical replications in which we varied the core parameters of our underlying assumptions (i.e., the number of predictors and the number of observations available to test our theory) in order to contrast their impact.

The variable selection results were described in terms of estimated FDR defined as

$$eFDR = \frac{1}{R}\sum_{r=1}^{R}\frac{\textit{number of falsely selected variables in the } r'\textit{th replication}}{\textit{total number of selected in the } r'\textit{th replication}}, \tag{4}$$

and power defined as

$$ePower = \frac{1}{R}\sum_{r=1}^{R}\frac{\textit{number of correctly selected in the } r'\textit{th replication}}{k} \tag{5}$$

and the average number of selected variables defined as

$$eN = \frac{1}{R}\sum_{r=1}^{R}\textit{number of selected in the } r'\textit{th replication} \ . \tag{6}$$

The number of simulation replications is set to be $R = 500$. We set the nominal FDR level for the knockoff filter at 0.2. The nominal FDR level controlled the selection restriction; in some cases, none of the variables were selected when setting the nominal FDR level too low because (1) the restriction was too tight or (2) the number of true relevant variables was too small (for example, only one out of five selected variables was a false discovery, and FDR was still 1/5 = 0.2 in this case, meaning that 20% of the irrelevant variables were selected). To present the average performance over 500 replications, we generated a simulation dataset using seed numbers from 1 to 500. By fixing the pseudo seed number, we determined the following random numbers in our simulations. All materials including a manual to perform all analyses are accessible as online supplements on OSF [weblink].

## 3. Results

The performance contrast simulation results in Table 1 show that the model-X knockoff selected fewer variables than the Lasso method. When it comes to FDR control, the model-X knockoff method made many fewer mistakes, as evident from a combined look at the number of selected variables and FDR. For example, the model-X knockoff method only used 12 relevant variables and irrelevantly chose a wrong one at the odds of 0.14 (see top panel of Table 1).

Importantly, with respect to statistical power, both the Lasso and the model-X knockoff filter exhibit power close to 1. This indicates that, when the FDR threshold $q$ is chosen appropriately, the model-X knockoff can effectively filter out relevant variables while simultaneously controlling false discoveries.

Importantly, when more variables were added (see the bottom panel of Table 1), the Lasso method was more responsive to the dataset and the number of relevant variables. In contrast, the knockoff method was very stable throughout and much less sensitive to even small changes in the dataset (Fig. 2).

**Table 1**
Performance contrast simulation in two example scenarios.

| Example 1. Adding more observations | | | | | | |
|---|---|---|---|---|---|---|
| | | $p = 300$ | $p = 300$ | $p = 300$ | $p = 300$ | $p = 300$ |
| | | $n = 100$ | $n = 150$ | $n = 200$ | $n = 250$ | $n = 300$ |
| Lasso | *eFDR* | 0.56 | 0.47 | 0.39 | 0.40 | 0.39 |
| | *ePower* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | *Selected* | 24 | 20 | 17 | 18 | 18 |
| Knockoff | *eFDR* | 0.14 | 0.17 | 0.15 | 0.16 | 0.15 |
| | *ePower* | 0.83 | 1.00 | 1.00 | 1.00 | 1.00 |
| | *Selected* | 12 | 12 | 12 | 12 | 12 |
| Example 2. Adding more variables | | | | | | |
| | | $p = 110$ | $p = 130$ | $p = 150$ | $p = 170$ | $p = 190$ |
| | | $n = 100$ | $n = 100$ | $n = 100$ | $n = 100$ | $n = 100$ |
| Lasso | *eFDR* | 0.56 | 0.51 | 0.49 | 0.47 | 0.58 |
| | *ePower* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | *Selected* | 24 | 22 | 21 | 20 | 25 |
| Knockoff | *eFDR* | 0.21 | 0.17 | 0.22 | 0.15 | 0.21 |
| | *ePower* | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| | *Selected* | 13 | 13 | 13 | 12 | 13 |

*Note.* $p$ = number of variables, $n$ = number of observations. eFDR = empirical false discovery rate defined in Equation (4), i.e., the number of selected, irrelevant variables divided by the number of selected variables. ePower = the average number of correctly selected relevant variables divided by the number of true relevant ones; this is defined in Equation (5). Selected = the average number of selected variables; this is defined in Equation (6).

### 3.1. Real data application: online news popularity

This application employs a Mashable dataset (Fernandes et al., 2015), which includes 39,797 articles characterized by 58 predictive attributes (e.g., linguistic features, temporal metadata) and one outcome variable (number of shares). Model-X knockoffs, using a Poisson regression framework with a 0.15 false discovery rate (FDR) threshold and Lasso coefficient difference statistic, identified seven common significant predictors of article virality. The dataset was partitioned into

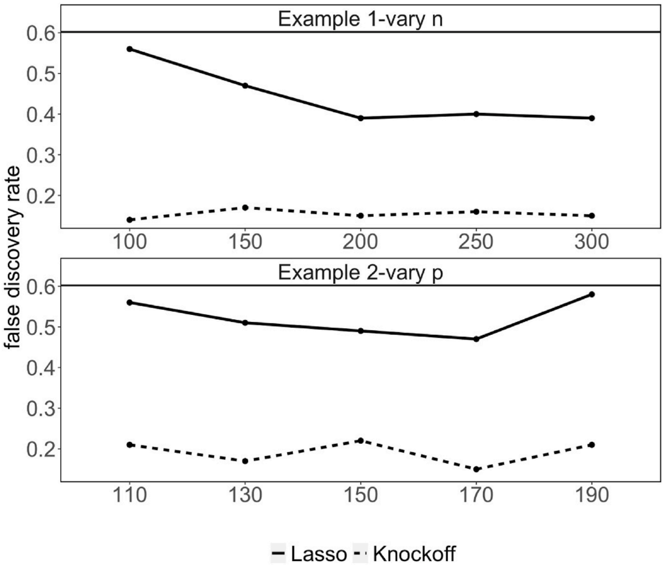

**Fig. 2.** Averaged False Discovery Rate in Two Simulation Examples
*Note.* In Example 1 (top panel), more observations $n$ are added (i.e., vary $n$), in example 2 (bottom panel), more variables $p$ are added (i.e., vary $p$).

three sequential 33.33% folds to simulate incremental data acquisition, revealing patterns such as higher shares for articles. Six predictors were consistently identified as predictors of the number of shares for news articles: *Average length of the words in the content* (predictor 11), *the data channel* [is] *'entertainment'* (predictor 14), [average shares of the] *best keyword* (predictor 26), [minimum number of] *shares of referenceapid articles in Mashable* (predictor 28), *closeness to LDA topic 2* (predictor 41),[1] and *text subjectivity* (predictor 44). These findings enable data-driven strategies, such as optimizing content length for specific categories, tailoring publication wording, and keyword choices to maximize engagement. The R code for this analysis is available on the Open Science Framework [weblink], and the dataset is publicly accessible via Kaggle [weblink]. This real-world data application demonstrates how model-X knockoffs can be applied to editorial decision-making, providing actionable insights to optimize audience reach.

## 4. Discussion

Having an abundance of available data can be a problem. In addition, using high dimensional scenarios in which there are more predictors ($p$) than observations ($n$) can be a challenge for the future analysis of large-scale applications (e.g., Brady, 2019). We have argued that access to large-scale data can facilitate bias from false-positive findings in large datasets by increasing researcher degrees of freedom (Simmons et al., 2011). Researchers can disclose their criteria for data collection, inclusion and exclusion criteria for observations, and offer a rationale for selecting relevant model variables and covariates. However, data abundancy exacerbates the flexibility-ambiguity issues involved in those decisions. In addition to open science practices (e.g., Dienlin et al., 2021), model-X knockoffs (Candès et al., 2018) have been introduced in the present study as a strategy to reduce researcher degrees of freedom and identify the true model (Gelman & Hill, 2007).

Specifically, the numerical performance of a model-X knockoff in R using the package *knockoff* can be helpful in two specific scenarios for scholars. In the first scenario, more variables were added to an existing dataset; in the second scenario, more observations were added to an existing dataset. The performance contrast simulation results indicated that model-X knockoffs select fewer variables than, for example, the common Lasso method (e.g., Scherr & Zhou, 2020). Overall, the model-X knockoffs performed better and made fewer mistakes, as indicated by the number of selected variables and the respective FDR results. The reason behind this is that the knockoff method chooses the difference as a contrast between used and selected variables. As such, compared to the Lasso method, the knockoff method is very stable and operates in a similar fashion as an extra filter. Importantly, although our simulation involved adding observations and variables, the findings can also be interpreted in reverse. They hold true when observations and/or variables are removed from a dataset.

In this article, we showcased model-X knockoffs (Candès et al., 2018) as a relatively novel method that can help researchers efficiently identify relevant variables, such as in high dimensional environments, where the available predictors outnumber the available observations. The biggest advantage of model-X knockoff variable selection methods is their ability to reduce researcher degrees of freedom (Simmons et al., 2011); therefore, they can help increase the transparency and reproducibility of scientific findings (e.g., Dienlin et al., 2021). The primary focus of this paper is the performance contrast simulation, which serves as a foundational tool for exploring theoretical applications. To illustrate the utility of this simulation, we used a hypothetical application around the "platform theory." Social media platforms, provide data that can be used to apply our suggested approach in empirical social science and humanities research.

In the first scenario (top panel of Table 1), we used the simulation to analyze data from $n$ participants and selected $p$ variables as one could do through e.g., X API. It is important to emphasize that the "platform theory" discussed is purely hypothetical and serves as an illustrative example rather than a validated explanation. Among the $p$ variables examined, only a subset was relevant to this hypothetical framework. Additionally, an API may offer more variables in the future (i.e., adding more variables $p$), and the platform could also introduce new data that might affect the theoretical model.

In the second scenario (bottom panel of Table 1), the researcher identified specific $p$ variables, e.g., based on the existing literature and conceptual models, as they were deemed relevant for testing the hypothetical "platform theory." If these variables are accurate predictors, new users joining the platform could necessitate adjustments to the theory over time which is simulated in our second scenario, in which model-X knockoffs help identifying a stable number of relevant predictors largely independent of the fact that new cases, i.e., new users signed up for the platform and their data is available through the API.

Importantly, FDR control methods are necessary when multiple hypotheses are simultaneously tested using large datasets. Originally a big issue in genome-wide studies (Storey & Tibshirani, 2003), a researcher can e.g., test an abundance of media use data points from social media environments to determine their association with topics such as mental well-being. However, when multiple hypotheses are tested simultaneously, the probability of observing a significant result by chance alone increases, leading to a high FDR. In such cases, methods such as the well-known Bonferroni method are much too conservative and lead to many missed findings. Instead, new FDR control methods are being used to adjust $p$-values to control the proportion of false discoveries among all significant results, but not equally across disciplines. Although $p$-value-based FDR control methods, such as the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), are usually stable and effective, they rely on $p$-values, which can be a controversial and challenging topic when $n$ and $p$ are large.

There are at least three reasons why model-X knockoffs should be considered a successful FDR control method. First, model-X knockoffs do not rely on $p$-values and therefore address some of the limitations of the widely-used Benjamini-Hochberg procedure. Second, the Benjamini-Hochberg procedure assumes that the hypothesis tests, as well as the $p$-values are independent. In the regression setting, this independence assumption is barely met. Since the model-X knockoff relies on parameter estimation rather than $p$-values, it naturally bypassed this independence requirement of Benjamini-Hochberg. Moreover, this is something that other procedures do not take into account. Third, model-X knockoffs can be effective when the number of variables is large relative to the sample size.

### 4.1. Limitations

As with any statistical method, model-X knockoffs have limitations that should be considered before using them. One important limitation is the model-X knockoff method's computational intensity. This method involves generating a large number of knockoffs to construct and test hypotheses about the importance of each feature. For very large datasets, the computational cost may become prohibitive, making it difficult to use the method in practice. Additionally, as noted by Ren et al. (2023), the method introduced in this paper relies on a single instance of randomly generating knockoff variables. To derandomize the approach and achieve more robust selection, one could use the de-randomization procedure outlined by Ren et al. (2023). This procedure involves combining results multiple knockoff generations, which inevitability increases computational complexity.

Moreover, model-X knockoffs have been shown to be effective for

[1] Fernandes et al. (2015) applied Latent Dirichlet Allocation (LDA) to the Mashable news articles, identifying five dominant topics and quantifying each article's alignment with them. Predictor 41 reflects the closeness to 'LDA topic 2,' an undefined construct in the original study.

real-world data in empirical applications (Pan, 2022; Sesia et al., 2018), however, they face specific challenges in the context of social science data. The additional challenges of social science data include biased sampling frames and social network effects (Hargittai, 2015), for which the use of model-X knockoffs can be particularly advantageous in order to control the false discovery rate.

Further, model-X knockoffs are designed for variable selection, not for variable transformation or model building. This means that the method may not be useful in situations where the goal is to identify complex (e.g., non-linear) relationships between variables. Finally, the threshold that controls the false discovery rate $q$ is predetermined by the users, and it follows a similar idea as the nominal significance level in hypothesis testing. A larger threshold $q$ will lead to both relevant and irrelevant variables being selected. There is no data-driven approach to determining the threshold to our knowledge. Users must assess and adjust the threshold case by case. For example, when users hope to select as many relevant variables as possible and do not worry about irrelevant ones, the threshold $q$ can be higher. However, when the primary goal is finding the most relevant variables, the threshold $q$ should be set lower to achieve a stricter control of false discoveries.

## 5. Conclusion

In this paper, we have introduced model-X knockoffs as a viable strategy for reducing researcher degrees of freedom and improving variable selection in high-dimensional datasets. The performance contrast simulation results demonstrate that integrating FDR-controlled variable selection with the commonly used Lasso estimator can substantially enhance the finite sample selection performance of regularized estimators. The model-X knockoff selection method shows relatively stable empirical false discovery rates compared to the Lasso, even with minor changes in dataset size, which improves the FDR control for variable selection. Additionally, the knockoff framework offers at least two significant advantages over traditional $p$-value-based FDR control methods, such as the Benjamini-Hochberg procedure: 1) $p$-values are not available for regularized estimators like the Lasso; $p$-value-based FDR control is feasible only with debiased estimators (Javanmard & Montanari, 2014; van de Geer et al., 2014). To our knowledge, practical implementations of $p$-values for debiased estimators, such as SCAD penalties, Multiple Comparison Procedures (MCP), or the group Lasso, are not yet available. In contrast, the model-X knockoff framework allows for FDR control directly with regularized estimators. 2) The knockoff framework does not rely on the assumption of independent $p$-values, which is often unrealistic to assume in practice. In conclusion, the study findings suggest that the model-X knockoff method can be a valuable tool for researchers seeking to identify the most parsimonious model that explains and predicts a set of observations, particularly in scenarios where there are more predictors than observations. This can ultimately help increase the transparency and reproducibility of scientific findings and mitigate the risk of false-positive findings in large datasets.

## CRediT authorship contribution statement

**Jing Zhou:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Sebastian Scherr:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

Barber, R. F., Candès, E. J., & Samworth, R. J. (2020). Robust inference with knockoffs. *Annals of Statistics, 48*(3), 1409–1431. https://doi.org/10.1214/19-AOS1852

Bates, S., Candès, E., Janson, L., & Wang, W. (2021). Metropolized knockoff sampling. *Journal of the American Statistical Association, 116*(535), 1413–1427. https://doi.org/10.1080/01621459.2020.1729163

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature, 489*(7415), 295–298. https://doi.org/10.1038/nature11421

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science, 22*(1), 297–323. https://doi.org/10.1146/annurev-polisci-090216-023229

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society, 22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B, 80*(3), 551–577. https://doi.org/10.1111/rssb.12265

Dienlin, T., Johannes, N., Bowman, N. D., … de Vreese, C. (2021). An agenda for open science in communication. *Journal of Communication, 71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

Fernandes, K., Vinagre, P., & Cortez, P. (2015). A proactive intelligent decision support system for predicting the popularity of online news. *Proceedings of the 17th EPIA 2015*.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The Annals of the American Academy of Political and Social Science, 659*(1), 63–76. https://doi.org/10.1177/0002716215570866

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Javanmard, A., & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research, 15*(1), 2869–2909.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences, 111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Pan, Y. (2022). Feature screening and FDR control with knockoff features for ultrahigh-dimensional right-censored data. *Computational Statistics & Data Analysis, 173*, Article 107504. https://doi.org/10.1016/j.csda.2022.107504

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2 ed.). Cambridge University Press.

Perretti, C. T., Munch, S. B., & Sugihara, G. (2013). Reply to Hartig and Dormann: The true model myth. *Proceedings of the National Academy of Sciences, 110*(42), E3976–E3977. https://doi.org/10.1073/pnas.1312461110

Pfeffer, J., Mooseder, A., Hammer, L., Stritzel, O., & Garcia, D. (2022). This sample seems to be good enough! Assessing coverage and temporal reliability of twitter's academic API. *arXiv preprint arXiv:2204.02290*. https://doi.org/10.48550/arXiv.2204.02290

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science, 377*(6612), 1304–1310. https://doi.org/10.1126/science.abl4476

Ren, Z., Wei, Y., & Candès, E. (2023). Derandomizing knockoffs. *Journal of the American Statistical Association, 118*(542), 948–958. https://doi.org/10.1080/01621459.2021.1962720

Romano, Y., Sesia, M., & Candès, E. (2020). Deep knockoffs. *Journal of the American Statistical Association, 115*(532), 1861–1872. https://doi.org/10.1080/01621459.2019.1660174

Scherr, S., & Zhou, J. (2020). Automatically identifying relevant variables for linear regression with the lasso method: A methodological primer for its application with R and a performance contrast simulation with alternative selection strategies. *Communication Methods and Measures, 14*(3), 204–211. https://doi.org/10.1080/19312458.2019.1677882

Sesia, M., Sabatti, C., & Candès, E. J. (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika, 106*(1), 1–18. https://doi.org/10.1093/biomet/asy033

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences, 100*(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

The Open Science Foundation. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B, 58*(1), 267–288. http://www.jstor.org/stable/2346178.

van de Geer, S., Bühlmann, P., Ritov, Y.a., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics, 42*(3), 1166–1202. https://doi.org/10.1214/14-AOS1221