Title

Controlling costs: Feature selection on a budget.

Journal
Stat, 11(1)

ISSN
0038-9986

Authors
Yu, Guo
Witten, Daniela
Bien, Jacob

Publication Date
2022-12-01

DOI
10.1002/sta4.427

Peer reviewed

Published in final edited form as:

Stat. 2022 December ; 11(1): . doi:10.1002/sta4.427.

Controlling costs: Feature selection on a budget

Guo Yu[1], **Daniela Witten**[2], **Jacob Bien**[3]

[1]Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, California, USA

[2]Department of Statistics and Biostatistics, University of Washington, Seattle, Washington, USA

[3]Department of Data Sciences and Operations, Marshall School of Business, University of Southern California, Los Angeles, California, USA

Abstract

The traditional framework for feature selection treats all features as costing the same amount. However, in reality, a scientist often has considerable discretion regarding which variables to measure, and the decision involves a tradeoff between model accuracy and cost (where cost can refer to money, time, difficulty or intrusiveness). In particular, unnecessarily including an expensive feature in a model is worse than unnecessarily including a cheap feature. We propose a procedure, which we call cheap knockoffs, for performing feature selection in a cost-conscious manner. The key idea behind our method is to force higher cost features to compete with more knockoffs than cheaper features. We derive an upper bound on the weighted false discovery proportion associated with this procedure, which corresponds to the fraction of the feature cost that is wasted on unimportant features. We prove that this bound holds simultaneously with high probability over a path of selected variable sets of increasing size. A user may thus select a set of features based, for example, on the overall budget, while knowing that no more than a particular fraction of feature cost is wasted. We investigate, through simulation and a biomedical application, the practical importance of incorporating cost considerations into the feature selection process.

Keywords

feature cost; feature selection; multiple knockoffs; weighted false discovery proportion

1. | INTRODUCTION

The traditional framework for feature selection ignores the fact that, in practice, different features may have different costs. In reality, practitioners must balance the opposing demands of model accuracy and budget considerations. For example, as we will see in Section 4, in medical diagnosis, doctors often have a wide range of options for what features to measure: A laboratory result may provide highly relevant information yet is expensive in terms of money, time, and the burden on patients; a simple questionnaire

Correspondence: Guo Yu, Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA, 93110, USA. guoyu@ucsb.edu.
Present address: Guo Yu, 5519 South Hall, University of California Santa Barbara, Santa Barbara, CA, 93110, USA

or even demographic information may be less informative but incurs lower costs. When a questionnaire would suffice for forming an accurate diagnosis, performing a laboratory examination would be practically misguided. Likewise, how should we decide whether to sequence a patient's entire genome or simply to conduct some cheap lab tests? This same challenge appears in other domains. For example, to determine the veracity of an online news article, do we require high-quality features based on an expert's reading, or do features derived from natural language processing suffice?

Consider the response of interest Y and a set of features $X_1, \ldots, X_p$, where for each feature X_j, there is an associated cost $\omega_j > 0$. In this paper, we consider a very general model where $Y \mid X_1, \ldots, X_p$ follows an arbitrary distribution, and we assume that the joint distribution of $X_1, \ldots, X_p$ is known. Let $\mathscr{H}_0$ be the set of irrelevant features, that is, $j \in \mathscr{H}_0$ if and only if X_j is independent of Y conditional on the other variables $\{X_k : k \neq j\}$ (Definition 1 in Candes et al. 2018). Given a set of selected features $\mathscr{R} \subseteq \{1, \ldots, p\}$, the false discovery proportion (FDP) is defined as $|\mathscr{R} \cap \mathscr{H}_0| / |\mathscr{R}|$, that is, it is the fraction of selected features that are unnecessarily included.

Barber and Candès (2015) proposed the knockoff filter, a feature selection procedure that provably controls the false discovery rate, defined as $E(FDP)$. For each feature, they construct a knockoff feature, that is, a carefully constructed fake copy of that feature. A feature is then only selected if it shows considerably more association with the response than its knockoff counterpart. Katsevich and Ramdas (2018) showed that one can directly upper-bound the false discovery proportion, with high probability, simultaneously for an entire path of selected models, $\mathscr{R}_1, \ldots, \mathscr{R}_p$, where $\mathscr{R}_k \subseteq \mathscr{R}_{k+1}$ for all k.

However, the false discovery proportion and the false discovery rate put all features on an equal footing, and do not consider their costs $\omega_1, \ldots, \omega_p$. To overcome this shortcoming, the weighted false discovery proportion (wFDP; Benjamini & Hochberg, 1997) is defined as wFDP $(\mathscr{R}) = C(\mathscr{R} \cap \mathscr{H}_0)/C(\mathscr{R})$, that is, the fraction of the total cost that is wasted, where $C(\mathscr{A}) = \sum_{j \in \mathscr{A}} \omega_j$ is the cost of measuring the features in $\mathscr{A}$.

The weighted false discovery proportion and weighted false discovery rate are not new (Benjamini & Hochberg, 1997; Benjamini & Heller, 2007), and the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) has been generalized to the weighted false discovery rate setting. A related criterion is the penalty-weighted false discovery rate (Ramdas et al. 2019), which can be controlled with the p-filter. However, the aforementioned procedures only provably control the corresponding criteria under restrictive dependence assumptions on the p-values (Benjamini & Yekutieli, 2001). Under arbitrary dependence, the reshaping process (Benjamini & Yekutieli, 2001; Blanchard & Roquain, 2008; Ramdas et al. 2019) needs to be applied, which can greatly reduce power. Basu et al. (2018) proposed a procedure that has asymptotic control of a related quantity, namely E[C$(\mathscr{R} \cap \mathscr{H}_0)$]/E[$C(\mathscr{R})$], in a mixture model under certain regularity conditions.

In this work, we adapt the ideas of knockoffs (Barber & Candès, 2015) and simultaneous inference (Goeman & Solari, 2011; Katsevich & Ramdas, 2018) to the setting where features have costs. The key to our method, which we call *cheap knockoffs*, is to construct multiple knockoffs for each feature, with more expensive features having more knockoffs. A feature is selected only if it beats all of its knockoff counterparts; thus, costlier features have more competition. This procedure yields a path of selected feature sets $\mathscr{R}_1, \ldots, \mathscr{R}_p$ for which wFDP $(\mathscr{R}_k)$ is bounded by a certain computable quantity with high probability, regardless of how k is chosen. Unlike existing works on the weighted false discovery rate control (Benjamini & Hochberg, 1997; Benjamini & Heller, 2007; Ramdas et al. 2019), our method provably bounds the weighted false discovery proportion under arbitrary dependence among features. Yu et al. (2022) recently proposed a predictive modelling method in high-dimensional cost-constrained linear regression problems. Different from their focus which is on good prediction performance under budget constraints, our method aims at recovering the true set of features (as defined in $\mathscr{H}_0^c$) with wFDP control.

2 | CHEAP KNOCKOFFS

2.1 | A review of model-X knockoffs and simultaneous inference

Our method is based on the model-X knockoff procedure (Candes et al. 2018) and its multiple knockoff extension (Roquero Gimenez & Zou, 2018), which provably control the false discovery rate for an arbitrary sample size n and a number of features p. For simplicity, we focus on the following linear model setting

$$\mathrm{E}[Y \mid X_1, \ldots, X_p] = \sum_{j=1}^{p} \beta_j X_j, (X_1, \ldots, X_p)^T \sim N\left(0, \Sigma\right). \tag{1}$$

We start by briefly reviewing the model-X knockoff approach in the simultaneous inference setting, applied specifically in the linear model (1). Throughout this paper, we denote $\mathbf{X} \in \mathbb{R}^{n \times p}$ as a data matrix and $\mathbf{y} \in \mathbb{R}^n$ as a response vector, where $(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip}, \mathbf{Y}_i) \in \mathbb{R}^p \times \mathbb{R}$ are independently and identically distributed as $(X_1, \ldots, X_p, Y)$ for $i = 1, \ldots, n$.

1. For each variable X_j, construct a knockoff variable $\widetilde{X}_j$ that satisfies:

    a. $\mathrm{E}\left(\widetilde{X}_j\right) = \mathrm{E}(X_j)$

    b. $\mathrm{Cov}\left(\widetilde{X}_j, \widetilde{X}_k\right) = \mathrm{Cov}(X_j, X_k)$ for all k

    c. $\mathrm{Cov}\left(\widetilde{X}_j, X_k\right) = \mathrm{Cov}(X_j, X_k) - s_j 1\left\{j = k\right\}$ for some $s_j \geq 0$.

    The knockoff variables $\widetilde{X} = \left(\widetilde{X}_1, \ldots, \widetilde{X}_p\right)$ are constructed to resemble X without any knowledge of the response Y. We denote $\widetilde{\mathbf{X}} \in \mathbb{R}^{n \times p}$ as the constructed knockoff matrix of $\mathbf{X}$ in a way that $\left(\widetilde{\mathbf{X}}_{i1}, \ldots, \widetilde{\mathbf{X}}_{ip}\right)$ is a knockoff of $(\mathbf{X}_{i1}, \ldots, \mathbf{X}_{ip})$ for $i = 1, \ldots, n$.

2. For each $j \in \{1, \ldots, p\}$, compute statistics T_j and $\widetilde{T}_j$ for the variables X_j and $\widetilde{X}_j$, respectively. For example, these could be the absolute values of the coefficients of a lasso regression (Tibshirani, 1996) on the augmented design matrix $\mathbf{Z} = \left[\mathbf{X}, \widetilde{\mathbf{X}}\right] \in \mathbb{R}^{n \times 2p}$:

$$\hat{\theta}\left(\lambda\right) = \underset{\theta \in \mathbb{R}^{2p}}{\arg\min}\left(\frac{1}{2} \parallel \mathbf{y} - \mathbf{Z}\theta \parallel_2^2 + \lambda \parallel \theta \parallel_1\right), \tag{2}$$

with $T_j = \left|\hat{\theta}\left(\lambda\right)_j\right|$ and $\widetilde{T}_j = \left|\hat{\theta}\left(\lambda\right)_{j+p}\right|$. The value of λ can be fixed in advance, or selected using cross-validation. The knockoff statistics are then defined as $W_j = T_j - \widetilde{T}_j$. Barber and Candès (2015) and Candes et al. (2018) discuss other choices of T_j's and W_j's. Intuitively, a large value of W_j indicates that X_j is a genuine signal variable, that is, the distribution of Y depends on X_j, whereas a small or negative value of W_j indicates that X_j may be irrelevant.

3. For any ordering of variables $\sigma(1), \ldots, \sigma(p)$, e.g., $|W_\sigma(1)| \geq |W_\sigma(2)| \geq \ldots \geq |W_{\sigma(p)}|$, report the sets of selected variables $\mathcal{R}_k = \{\sigma(j): \sigma(j) \leq \sigma(k), W_{\sigma(j)} > 0\}$, for $k \in \{1, \ldots, p\}$.

Katsevich and Ramdas (2018) work within the simultaneous inference framework (Goeman & Solari, 2011), in which a practitioner wishes to obtain a final set of selected variables with false discovery proportion control when choosing among $\{\mathcal{R}_k, k = 1, \ldots, p\}$. To allow for such behavior, Katsevich and Ramdas (2018) form a computable upper bound $\mathcal{U}_k$ such that $\mathrm{FDP}(\mathcal{R}_k) \leq \mathcal{U}_k$ holds simultaneously over all k with some known probability.

2.2 | Multiple knockoffs based on cost

The knockoff procedure described in the previous section constructs a single knockoff variable for each feature, and then selects features based solely on the values of $W_1, \ldots, W_p$. Barber and Candès (2015) and Candes et al. (2018) discuss the possibility of constructing K knockoffs per feature for some value $K > 1$ with the goal of achieving higher statistical power and stability. This has been pursued in Roquero Gimenez and Zou (2018) and Emery et al. (2019).

We make a simple yet crucial modification to the multiple knockoff idea, allowing different features to have different numbers of knockoffs, so that an expensive irrelevant feature will have a lower chance of entering the model than a cheap irrelevant feature. Assume that the feature costs $\omega_1, \ldots, \omega_p$ are integers with $\omega_j \geq 2$. We construct $\omega_j - 1$ knockoff variables for each original variable X_j. If X_j is irrelevant, that is, $j \in \mathcal{H}_0$, then we expect it to be selected with probability $1/\omega_j$. We also incorporate costs into the construction of the sequence of selected feature sets $\mathcal{R}_k$. The cheap knockoff procedure generalizes the multiple knockoff procedure of Roquero Gimenez and Zou (2018) to the cost-conscious setting:

1. For each variable X_j with cost ω_j, denote $\widetilde{X}_j^{(1)} = X_j$ and construct the knockoff variables $\widetilde{X}_j^{(2)}, \widetilde{X}_j^{(3)}, \ldots, \widetilde{X}_j^{(\omega_j)}$ such that(a) $\mathrm{E}\left(\widetilde{X}_j^{(\ell)}\right) = \mathrm{E}(X_j)$ for $\ell \in \{2, \ldots, \omega_j\}$.

(b) $\mathrm{Cov}\left(\widetilde{X}_j^{(\ell)}, \widetilde{X}_k^{(m)}\right) = \mathrm{Cov}(X_j, X_k) - s_j \mathbb{1}\left\{j = k\right\}\mathbb{1}\left\{\ell \neq m\right\}$ for all $\ell \in \{1, \ldots, \omega_j\}, m \in \{1, \ldots, \omega_k\}, j, k \in \{1, \ldots, p\}$, and some constant $s_j \geq 0$.

We denote $\widetilde{\mathbf{X}}_j^{(\ell)} \in \mathbb{R}^n$ as the constructed knockoff variables of $\mathbf{X}_j$, such that $\left(\widetilde{\mathbf{X}}_{ij}^{(\ell)}\right)_{j=1,\ldots,p}^{\ell=1,\ldots,\omega_j}$ satisfies the condition above for $(\mathbf{X}_{ij})_{j=1,\ldots,p}$ for $i = 1, \ldots, n$.

2. For each $j \in \{1, \ldots, p\}$, compute the statistics $T_j^{(1)}$ (corresponding to the original variable) and $T_j^{(2)}, \ldots, T_j^{(\omega_j)}$ (corresponding to the $\omega_j - 1$ knockoff variables). For example, these could be the absolute values of the coefficients of the following lasso regression:

$$\left\{\hat{\theta}_j^{(\ell)}(\lambda)\right\}_{j \le p, \ell \le \omega_j} = \underset{\theta_j^{(\ell)}: j \le p,\ \ell\ \le \omega_j}{\arg\min} \left(\frac{1}{2}\left\| \mathbf{y} - \sum_{j=1}^{p} \sum_{\ell=1}^{\omega_j} \widetilde{\mathbf{X}}_j^{(\ell)} \theta_j^{(\ell)} \right\|_2^2 + \lambda \sum_{j=1}^{p} \sum_{\ell=1}^{\omega_j} \left|\theta_j^{(\ell)}\right|\right), \tag{3}$$

with $T_j^{(\ell)} = \left|\hat{\theta}_j^{(\ell)}(\lambda)\right|$. The value of λ in (3) can be selected using cross-validation. We define

$$\kappa_j = \underset{1 \le \ell \le \omega_j}{\arg\max}\ T_j^{(\ell)}. \tag{4}$$

3. For any ordering of variables $\sigma(1), \ldots, \sigma(p)$, report the sets of selected variables $\mathscr{R}_k = \{\sigma(j): \sigma(j) \le \sigma(k), \kappa_{\sigma(j)} = 1\}$, for $k \in \{1, \ldots, p\}$.

In Step 1, various methods are available for constructing multiple knockoffs given that the distribution of X is known (see, e.g., Candes et al. 2018; Roquero Gimenez & Zou, 2018). The computation of κ_j in Step 2 involves the ω_j statistics $T_j^{(1)}, \ldots, T_j^{(\omega_j)}$; $\kappa_j = 1$ indicates that the original variable beats all of its $\omega_j - 1$ knockoff copies. We show in Appendix B1 that the probability of this occurring for an irrelevant feature is inversely proportional to the feature's cost. This is the key property used to show the simultaneous control of the weighted false discovery proportion in the next section.

In principle, any ordering of variables can be used to obtain $\mathscr{R}_k$. In simulations, we consider a specific ordering such that $\tau_{\sigma(1)} \ge \tau_{\sigma(2)} \ldots \ge \tau_{\sigma(p)}$, where $\tau_j = 2\omega_j^{-1}\left\{T_j^{(\kappa_j)} - \max_{\ell \ne \kappa_j} T_j^{(\ell)}\right\}$. One reason for this specific choice of τ_j is that when $\omega_1 = \ldots = \omega_p = 2$, the above procedure is exactly the same as the standard knockoff procedure reviewed in Section 2.1. In particular, $W_j > 0$ if and only if $\kappa_j = 1$, and $|W_j| = \tau_j$. Moreover, all else being equal, we want to make use of cheap features over expensive features. For this reason, we set τ_j to be inversely proportional to the feature cost.

2.3 | Simultaneous control of the weighted false discovery proportion

Having constructed a cost-conscious path of selected variable sets $\mathscr{R}_1, \ldots, \mathscr{R}_p$, we next provide a simultaneous high-probability bound on the weighted false discovery proportion along this path. The next theorem and the remark that follows establish that the computable quantities $\overline{\mathscr{U}}(\mathscr{R}_1, c), \ldots, \overline{\mathscr{U}}(\mathscr{R}_p, c)$, defined below in (7), simultaneously upper bound wFDP

$(\mathscr{R}_1), \ldots,$ wFDP $(\mathscr{R}_p)$ with a known probability. This means that for any choice of k, with high probability our selected feature set is not too wasteful (in terms of the fraction of cost spent on irrelevant features).

Theorem 1. *For any* $\alpha \in (0,1)$, *we have*

$$\mathbb{P}\{\mathrm{wFDP}(\mathscr{R}_k) \leq \mathscr{U}(\mathscr{R}_k, c) \text{ for all } k\} \geq 1 - \alpha, \tag{5}$$

where for any constant $c > 0$

$$\mathscr{U}(\mathscr{R}_k, c) = -\log \alpha \left[\frac{1 + c\sum_{j=1}^{k} \mathbb{1}\{j \notin \mathscr{R}_k\}}{\left(\sum_{j=1}^{k} \omega_j \mathbb{1}\{j \in \mathscr{R}_k\}\right) \vee 1}\right]\left[\max_{k \in \mathscr{H}_0} \frac{\omega_k}{\log\{\omega_k - (\omega_k - 1)\alpha^c\}}\right]. \tag{6}$$

For the standard knockoff procedure described in Section 2.1, we have $\omega_1 = \ldots = \omega_p = 2$. In that case, with $c = 1$, (6) reduces exactly to the bound from applying Theorem 2 of Katsevich and Ramdas (2018) to the Selective and Adaptive SeqStep procedure (Barber & Candès, 2015) with $p_* = \lambda = 1/2$.

As mentioned in Section 2.1, our procedure can be generalized to any known distribution of X and any unknown conditional distribution of Y given X. For example, in the binary classification data example in Section 4, we consider the statistics $\left\{T_j^{(\ell)}\right\}$ derived from ℓ_1-penalized logistic regression. Following the arguments in Candes et al. (2018), we can show that Theorem 1 also holds for this choice of $\left\{T_j^{(\ell)}\right\}$.

Remark 1. The weighted false discovery proportion upper bound $\mathscr{U}(\mathscr{R}_k, c)$ depends on the unknown set $\mathscr{H}_0$. In practice, we can use an upper bound

$$\bar{\mathscr{U}}(\mathscr{R}_k, c) = -\log \alpha \left[\frac{1 + c\sum_{j=1}^{k} \mathbb{1}\{j \notin \mathscr{R}_k\}}{\left(\sum_{j=1}^{k} \omega_j \mathbb{1}\{j \in \mathscr{R}_k\}\right) \vee 1}\right]\left[\max_{k} \frac{\omega_k}{\log\{\omega_k - (\omega_k - 1)\alpha^c\}}\right]. \tag{7}$$

Moreover, if an estimated set $\hat{\mathscr{H}}_0$ satisfying $\mathscr{H}_0 \subseteq \hat{\mathscr{H}}_0$ is available, then (6) with the maximum taken over $\hat{\mathscr{H}}_0$ gives a tighter bound in (5).

Our procedure yields a sequence of sets $\mathscr{R}_k$ of selected variables, and the bound in (5) gives a specific description of the tradeoff between capturing enough of the signal variables and incurring too much cost. The simultaneous nature of the bound means that wFDP $(\mathscr{R}_k)$ is controlled regardless of the approach used to select k: the choice of k can depend on the size of $\mathscr{R}_k$, the cost of $\mathscr{R}_k$, or in fact any function of the data.

3 | SIMULATION STUDIES

We now investigate the feature selection performance of cheap knockoffs in simulation. We set $n = 200$ and $p = 30$. Each element of the design matrix $X \in \mathbb{R}^{n \times p}$ is independent and identically distributed as $N(0,1)$. The response is generated from the linear model (1) with Gaussian errors $\varepsilon \sim N\left(0, \sigma^2\right)$ and $\sigma^2 = \left(4n\right)^{-1} \| \mathbf{X}\beta \|_2^2$. We let $\beta_1 = \ldots = \beta_{10} = 2$, and $\beta_j = 0$ for $j > 10$. We set the first half of the relevant features to be expensive and the second half to be cheap, that is, $\omega_1 = \ldots = \omega_5 = 6$, and $\omega_6 = \ldots = \omega_{10} = 2$. For the irrelevant features, that is, for any $j > 10$, we set $\mathbb{P}(\omega_j = 6) = \gamma$ and $\mathbb{P}(\omega_j = 2) = 1 - \gamma$, where $\gamma \in \{0,0.25,0.5,0.75,1\}$.

We construct multiple knockoff variables using entropy maximization (Roquero Gimenez & Zou, 2018), and we compute the statistics $T_j^{(\ell)}$ as the absolute value of the lasso coefficient estimates in (3), with the tuning parameter selected using cross-validation. In Appendix A we report the wall-clock running time of cheap knockoffs in the numerical studies (Tables A1 and A2). We find that the majority of computation is spent on generating multiple knockoffs, which is challenging when p is large and (or) the feature costs are large (after dividing by their greatest common factor). In such cases, alternative construction methods could be used. For example, Roquero Gimenez and Zou (Appendix A.1; (2018)) show that an equicorrelation construction has a closed form expression, which is particularly favorable in computation since it does not depend on the number of multiple knockoffs (and equivalently, the feature costs).

We first verify the bound in Theorem 1 and compare the performance of cheap knockoffs to Katsevich and Ramdas (2018), which ignores feature costs. In particular, by carrying out Steps $1-3$ in Section 2.1 with $\omega_1 = \ldots = \omega_p = 2$ in (7), the bound in (7) coincides with the result in Katsevich and Ramdas (2018). We denote this approach as Katsevich and Ramdas (2018). For both methods, we take $\alpha = 0.2$ in (7). In Figure 1 we report both the ratio $\bar{\mathcal{U}}(\mathscr{R}_k, 1)^{-1}\text{wFDP}(\mathscr{R}_k)$ and the actual weighted false discovery proportion wFDP $(\mathscr{R}_k)$ for each $\mathscr{R}_k$ for both methods in the settings where $\gamma = 0{,}0.5$, and 1. As seen in Figure 1, the ratio $\bar{\mathcal{U}}(\mathscr{R}_k, 1)^{-1}$ wFDP $(\mathscr{R}_k)$ for our cheap knockoff procedure is mostly below 1, indicating that the bound in Theorem 1 holds. Moreover, when γ is large, the weighted false discovery proportion for the cheap knockoff procedure is lower than Katsevich and Ramdas (2018) for most values of k. Table 1 gives the estimated probability that the bound is violated, that is, $\hat{\mathbb{P}}\left(\sup_k \bar{\mathcal{U}}_k^{-1}(\mathscr{R}_k, 1)\text{wFDP}(\mathscr{R}_k) > 1\right)$, for each method for $\gamma \in \{0,0.25,0.5,0.75,1\}$.

We see that the Katsevich and Ramdas (2018) procedure which is not cost-conscious performs worse as γ increases, that is, when irrelevant variables are more likely to be expensive. Since the method ignores cost, it may erroneously select expensive irrelevant features, leading to a poor weighted false discovery proportion.

While our proposal focuses on recovering the correct set of features with simultaneous wFDP control, we show empirically that the set of features selected by cheap knockoffs usually incurs low cost without compromising prediction accuracy. Specifically, for each set of selected variables $\mathscr{R}_1, \ldots, \mathscr{R}_p$, we compute both the root-mean-square prediction error of the least squares model fit to the variables in $\mathscr{R}_k$, and the total cost $\sum_{j \in \mathscr{R}_k} \omega_j$. We see from Figure 2 that for a given budget, the cheap knockoff procedure attains smaller prediction error than the procedure in Katsevich and Ramdas (2018), which is not cost-conscious. In particular, the cheap knockoff procedure tends to select all five of the cheap relevant features before any expensive feature is let in the model, whereas Katsevich and Ramdas (2018) does not take feature cost into consideration. For $k \geq 10, \mathscr{R}_k$ for both methods includes essentially all the relevant features, thus giving similar performance.

4 | DATA APPLICATION

To gauge the performance of cheap knockoffs in a real dataset, we consider data from the National Health and Nutrition Examination Survey (NHANES) (National Center for Health Statistics, 2018, processed in Kachuee, Goldstein, et al., 2019; Kachuee, Karkkainen et al., 2019). The dataset contains 92062 samples of survey participants. We consider 30 features, which can be broadly categorized into four types: demographics, questionnaire-based, examination-based and laboratory-based. For each feature, medical experts suggest a corresponding integer-valued cost (ranging from 2 to 9) for that feature based on "the overall financial burden, patient privacy, and patient inconvenience" (Kachuee, Karkkainen, et al., (2019)). A brief summary of the 30 features can be found in Table 2. Finally, each observation is associated with a label of pre-diabetes/diabetes (as one category) or normal. The task is to select features that are closely associated with diabetes while taking feature cost into consideration.

We consider the cheap knockoff procedure as in Section 2.2, modified so that the statistics $\left\{T_j^{(\ell)}\right\}$ computed in (3) are derived from ℓ_1-penalized logistic regression (instead of ℓ_1-penalized least squares). Following the arguments in Candes et al. (2018), we can show that Theorem 1 also holds for this choice of $\left\{T_j^{(\ell)}\right\}$.

To numerically verify Theorem 1, we would need to know the true set of relevant variables. We test the cheap knockoff procedure using partially-simulated data. To form a reasonable ground truth, we start by performing logistic regression on a random set of 72,062 samples. In total, we retain 11 variables whose p-values are smaller than 0.01/30 (by Bonferroni correction). We take these as the true set of relevant variables (see Table 4 for the list of relevant variables). We next generate responses for the remaining 20,000 samples from a logistic regression model using only these selected features. The coefficient values used correspond to those from the fitted logistic regression estimates. We then randomly divide these 20,000 samples (with simulated responses) into 50 non-overlapping sets, each containing 400 samples. On each set, we run our method to obtain a path of selected variables. Finally, we compute the estimated probability that the bound in (6) is violated, that is, $\widehat{\mathbb{P}}\left(\sup_k \overline{\mathscr{U}}_k^{-1}(\mathscr{R}_k, 1)\text{wFDP}(\mathscr{R}_k) > 1\right)$ for $\alpha \in \{0.05, 0.1, \ldots, 0.5\}$. We see from Table 3 that the estimated probability is lower than the corresponding value of α, indicating that Theorem 1 holds for our proposed cost-conscious procedure.

On each of the 50 non-overlapping data subsets, we further compute wFDP and cost for the path of selected variables $\mathscr{R}_k$ returned by cheap knockoffs and the proposal in Katsevich and Ramdas (2018), which ignores feature costs. Figure 3 reports the 20th, 50th and 80th percentiles (over the 50 non-overlapping sets) of wFDP and cost and shows that our proposal effectively attains a lower wFDP and a lower cost than the proposal in Katsevich and Ramdas (2018).

Although prediction performance of the selected model is not the main theoretical focus of our proposal, we next study the prediction performance and the total cost of the selected variables. For comparison, we consider the following methods:

1. Katsevich & Ramdas(2018): the proposal of Katsevich and Ramdas (2018) applied to the 'Selective and adaptive SeqStep' method. It is equivalent to our method if we ignore the cost information, that is, we set $\omega_1 = \omega_2 = \ldots = \omega_{30} = 2$.

2. **Logistic regression:** logistic regression applied to all 30 features. This procedure is not cost-conscious, and does not perform features selection. We use this as a benchmark for classification performance.

We run these methods on all 92062 observations. Given the large sample size, we expect the training error to be a good approximation of the generalization error. Furthermore, to highlight the effects of feature costs, we consider exaggerating the feature costs by using the squares of their actual costs. From Figure 4, we see that cheap knockoffs can achieve favorable classification performance at a low feature cost. In particular, the first two panels show that for a fixed model size, the cheap knockoff procedure tends to achieve slightly worse classification performance than the procedure of Katsevich and Ramdas (2018), which is not cost-conscious. However, our method achieves this classification performance at a lower cost. The right panel shows that for a given model cost, our method can obtain favorable classification performance compared with the proposal of Katsevich and Ramdas (2018). Moreover, our method's classification performance is close to the benchmark of logistic regression, while using a much cheaper set of features.

In Figures 5 and 6, we show the path of variables selected by cheap knockoffs and that of Katsevich and Ramdas (2018). Each point represents a variable added to a model (with the feature name in the legend). For example, we see that both methods include `Gender`, `Height`, `Weight`, and `Triglyceride` when the model size is 4. However, the cheap knockoff procedure tends to select cheaper features first, adding the expensive laboratory feature `Triglyceride` last among these four features. By comparison, the proposal of Katsevich and Ramdas (2018) does not show any preference for inexpensive features. For the model with two variables, cheap knockoffs selects `Gender` and `Height`, which has lower cost and better classification performance than the model of `Height` and `Weight` selected by Katsevich and Ramdas (2018).

In addition, in Figure 6, we present the path of variables selected by cheap knockoffs applied with squared feature costs, where squaring has been performed to exaggerate the effect of the feature costs. Comparing with Figure 5, we see that cheap knockoffs tends to select less expensive features, while still attaining comparable classification performance. In particular, when the costs are squared, cheap knockoffs no longer selects `Diastolic BP(2nd)`, `Systolic BP(4th)`, `Systolic BP(1st)`, `Diastolic`

`BP(3rd)`, `Vigorous activity`, and `Upper leg length`. Among these omitted variables, only `Upper leg length` is considered relevant by the logistic regression (see Table 4).

5 | DISCUSSION

In this paper, we proposed cheap knockoffs, a procedure for performing feature selection when features have costs. Cheap knockoffs is based on the idea of constructing multiple knockoffs for each feature. In particular, cheap knockoffs forces more expensive features to compete with more knockoffs, making it harder for expensive features to be selected. Our method yields a path of selected feature sets, and we show that the weighted false discovery proportion is simultaneously bounded with high probability along this path.

An interesting yet challenging future research direction is to develop a method based on the multiple knockoffs idea that provably controls the weighted false discovery rate. The martingale-type arguments used in the original knockoff paper rely on certain symmetries that are broken when the numbers of knockoffs constructed for different features are not all equal.

Finally, an R package named `cheapknockoff`, implementing our proposed method, is available on https://github.com/hugogogo/cheapknockoff. The simulation studies in Section 3 use the `simulator` package (Bien, 2016), and the code to reproduce the simulation results (in Section 3) and the NHANES data analysis (in Section 4) are available at https://github.com/hugogogo/reproducible/tree/master/cheapknockoff. The NHANES dataset (National Center for Health Statistics, 2018) is processed in Kachuee, Goldstein et al. (2019), Kachuee, Karkkainen et al. (2019) and available at https://github.com/mkachuee/Opportunistic.

ACKNOWLEDGEMENTS

The authors thank Will Fithian for suggesting the simultaneous inference framework. All authors were supported by NIH Grant R01GM123993. Jacob Bien was also supported by NSF CAREER Award DMS-1653017 and Daniela Witten was also partially supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and a Simons Investigator Award in Mathematical Modeling of Living Systems.

DATA AVAILABILITY STATEMENT

The NHANES dataset (National Center for Health Statistics, 2018) is processed in Kachuee, Goldstein et al. (2019); Kachuee, Karkkainen et al. (2019) and is available at https://github.com/mkachuee/Opportunistic.

APPENDIX A: RUNNING TIME COMPARISON IN NUMERICAL STUDIES

TABLE A1

Wall-clock time comparison (in seconds, averaged over 100 simulated datasets) between our proposal and Katsevich and Ramdas (2018) in generating Table 1

γ	**0**	**0.25**	**0.5**	**0.75**	**1**
Cheap knockoffs (our proposal)	2.796	2.772	2.784	2.798	2.812
Katsevich and Ramdas (2018)	0.273	0.250	0.258	0.251	0.253

TABLE A2

Wall-clock time comparison (in seconds, averaged over 50 non-overlapping data subsets) between our proposal and Katsevich and Ramdas (2018) in generating Figure 3

Cheap knockoffs (our proposal)	**7.284**
Katsevich and Ramdas (2018)	2.678

APPENDIX B: PROPERTIES OF MULTIPLE KNOCKOFFS

We study the properties of the multiple knockoffs constructed in Step 1 of Section 2.2. Define

$$\widetilde{Z} = \left(\widetilde{X}_1^{(2)}, \ldots, \widetilde{X}_1^{(\omega_1)}, \widetilde{X}_2^{(2)}, \ldots, \widetilde{X}_2^{(\omega_2)}, \ldots, \widetilde{X}_p^{(2)}, \ldots, \widetilde{X}_p^{(\omega_p)}\right)^T \in \mathbb{R}^{\sum_j (\omega_j - 1)}$$

as the random vector of all knockoff features, and

$$Z = \left(\widetilde{X}_1^{(1)}, \widetilde{X}_1^{(2)}, \ldots, \widetilde{X}_1^{(\omega_1)}, \widetilde{X}_2^{(1)}, \widetilde{X}_2^{(2)}, \ldots, \widetilde{X}_2^{(\omega_2)}, \ldots, \widetilde{X}_p^{(1)}, \widetilde{X}_p^{(2)}, \ldots, \widetilde{X}_p^{(\omega_p)}\right)^T \in \mathbb{R}^{\sum_j \omega_j}, \tag{B1}$$

where $\widetilde{X}_j^{(1)} = X_j$ is the original feature for $j = 1, \ldots, p$. For any p tuple of permutations $\varsigma = (\varsigma_1, \ldots, \varsigma_p)$ where ς_j is a permutation on the set $\{1, \ldots, \omega_j\}$, and for any vector $\upsilon = \left(\upsilon_1^{(1)}, \ldots, \upsilon_1^{(\omega_1)}, \ldots, \upsilon_p^{(1)}, \ldots, \upsilon_p^{(\omega_p)}\right) \in \mathbb{R}^{\sum_j \omega_j}$, we define

$$\upsilon_{\mathrm{swap}(\varsigma)} = \left(\upsilon_1^{(\varsigma_1(1))}, \ldots, \upsilon_1^{(\varsigma_1(\omega_1))}, \upsilon_2^{(\varsigma_2(1))}, \ldots, \upsilon_2^{(\varsigma_2(\omega_2))}, \ldots, \upsilon_p^{(\varsigma_p(1))}, \ldots, \upsilon_p^{(\varsigma_p(\omega_p))}\right)^T \in \mathbb{R}^{\sum_j \omega_j}.$$

Therefore, $Z_{\mathrm{swap}(\varsigma)}$ denotes the random vector where each ς_j permutes the ω_j knockoff features (including the original one) corresponding to X_j

We generalize the definition of multiple model-X knockoffs (Definition 3.2 in Roquero Gimenez & Zou, 2018) to our setting in which each feature can have a different number of knockoffs:

Definition 1. Consider any cost vector $\omega = (\omega_1, \ldots, \omega_p)$, where $\omega_j > 1$ are integers. The random vector $\widetilde{Z}$ is a valid ω-knockoff of $X = (X_1, \ldots, X_p)$ if

1. $Z_{\mathrm{swap}(\varsigma)}$ and Z are identically distributed for any tuple of permutations $\varsigma = (\varsigma_1, \ldots, \varsigma_p)$;
2. $\widetilde{Z}$ and Y are conditionally independent given X.

Under the assumption that X follows a multivariate Gaussian distribution, it can be verified (see, e.g., Proposition 3.4 in Roquero Gimenez & Zou, 2018) that following Step 1 in Section 2.2, the vector $\widetilde{Z}$ is a valid ω-knockoff of X. In particular, the second property is guaranteed provided that the construction of $\widetilde{Z}$ does not use Y, as in Roquero Gimenez and Zou (2018).

The next lemma states the exchangeability property of the irrelevant features and their knockoffs, that is, we can permute an irrelevant feature and its knockoffs without changing the joint distribution of Z and Y.

Lemma 1 Exchangeability of irrelevant features and their knockoffs. Consider any tuple of permutations $\varsigma' = (\varsigma_1, \ldots, \varsigma_p)$, where ς_j is the identity permutation for $j \notin \mathscr{H}_0$, and ς_j is an arbitrary permutation over the set $\{1, \ldots, \omega_j\}$ for $j \in \mathscr{H}_0$. If $\widetilde{Z}$ is a valid ω-knockoff of X, then (Z, Y) and $(Z_{\mathrm{swap}(\varsigma)}, Y)$ are identically distributed.

Proof. By the property of a valid ω-knockoff, $Z_{\mathrm{swap}\,(\varsigma)}$ and Z are identically distributed. So it is left to show that $Y \mid Z$ and $Y \mid Z_{\mathrm{swap}\,(\varsigma)}$ are identically distributed. This can be shown using the same arguments as in the proof of Lemma 1 in Candes et al. (2018).

We denote

$$T = \left(T_1^{(1)}, \ldots, T_1^{(\omega_1)}, T_2^{(1)}, \ldots, T_2^{(\omega_2)}, \ldots, T_p^{(1)}, \ldots, T_p^{(\omega_p)}\right) \in \mathbb{R}^{\sum_i \omega_j},$$

for $T_j^{(\ell)}$ defined in Step 2 of Section 2.2. Furthermore, we define component-wise order statistics on T,

$$T_{\mathrm{ordered}} = \left(T_{1,(1)}, \ldots, T_{1,(\omega_1)}, T_{2,(1)}, \ldots, T_{2,(\omega_2)}, \ldots, T_{p,(1)}, \ldots, T_{p,(\omega_p)}\right) \in \mathbb{R}^{\sum_j \omega_j}$$

such that $T_{j,(1)} \geq T_{j,(2)} \geq \ldots \geq T_{j,(\omega_j)}$ for all j.

The following lemma characterizes the multiple knockoff statistics $\{\kappa_j\}_{j=1}^p$ computed in Step 2 of Section 2.2. It essentially states that for $j \in \mathscr{H}_0$, the statistic κ_j corresponding to the irrelevant feature X_j is uniformly distributed on the set $\{1, \ldots, \omega_j\}$, and is independent of the statistics corresponding to all other features and the component-wise order statistics T_{ordered}. This property generalizes the "coin-flip" property of the standard model-X knockoff (see, e.g., Lemma 2 in Candes et al. 2018), and is the key to the proof of Theorem 1.

Lemma 2 Multiple knockoff statistics. Suppose $\widetilde{Z}$ is a valid ω-knockoff of Z. For any $j \in \mathscr{H}_0$, the statistic κ_j is uniformly distributed on the set $\{1, \ldots, \omega_j\}$, and is independent of $\{\kappa_k\}_{k \neq j}$ and the order statistics $T_{\text{ordered.}}$.

Proof. We adapt the proof idea in B.2 of Roquero Gimenez and Zou (2018). Consider any tuple of permutations $\varsigma = (\varsigma_1, \ldots, \varsigma_p)$, where ς_j is the identity permutation for $j \notin \mathscr{H}_0$, and ς_j is an arbitrary permutation over the set $\{1, \ldots, \omega_j\}$ for $j \in \mathscr{H}_0$. We first show that $(\varsigma_1(\kappa_1), \ldots, \varsigma_p(\kappa_p), T_{\text{ordered}})$ has the same distribution as $(\kappa_1, \ldots, \kappa_p, T_{\text{ordered}})$.

We denote $\varsigma^{-1} = (\varsigma_1^{-1}, \ldots, \varsigma_p^{-1})$ where ς_j^{-1} is the inverse permutation of ς_j. Recall from Step 2 of Section 2.2, combined with the definition of Z in (B1), that $T = f(Z, Y)$ for some map f, and observe that $T_{\text{swap}(\varsigma^{-1})} = f\left(Z_{\text{swap}(\varsigma^{-1})}, Y\right)$. So by Lemma 1, we have that $T_{\text{swap}(\varsigma^{-1})}$ and T are identically distributed. For any $k_j \in \{1, \ldots, \omega_j\}$ and $t_{j\ell} \in \mathbb{R}$ for $j = 1, \ldots, p$ and $\ell = 1, \ldots, \omega_j$, we have

$$
\begin{aligned}
&\mathbb{P}\left[\bigcap_{j=1}^{p}\{\kappa_j = k_j\}, \bigcap_{j=1}^{p}\bigcap_{\ell=1}^{\omega_j}\{T_{j,(\ell)} = t_{j\ell}\}\right] \\
&= \mathbb{P}\left[\bigcap_{j=1}^{p}\left\{T_j^{(k_j)} = T_{j,(1)} = t_{j1}\right\}, \bigcap_{j=1}^{p}\bigcap_{\ell=1}^{\omega_j}\{T_{j,(\ell)} = t_{j\ell}\}\right] \\
&= \mathbb{P}\left[\bigcap_{j=1}^{p}\left\{T_j^{\left(\varsigma_j^{-1}(k_j)\right)} = T_{j,(1)} = t_{j1}\right\}, \bigcap_{j=1}^{p}\bigcap_{\ell=1}^{\omega_j}\{T_{j,(\ell)} = t_{j\ell}\}\right] \\
&= \mathbb{P}\left[\bigcap_{j=1}^{p}\{\kappa_j = s_j^{-1}(k_j)\}, \bigcap_{j=1}^{p}\bigcap_{\ell=1}^{\omega_j}\{T_{j,(\ell)} = t_{j\ell}\}\right] \\
&= \mathbb{P}\left[\bigcap_{j=1}^{p}\{s_j(\kappa_j) = k_j\}, \bigcap_{j=1}^{p}\bigcap_{\ell=1}^{\omega_j}\{T_{j,(\ell)} = t_{j\ell}\}\right],
\end{aligned}
$$

where the first and the third equalities hold from the definition of κ_j's, the second equality holds because $T_{\text{swap}(\varsigma^{-1})}$ and T are identically distributed, along with the fact that $\left(T_{\text{swap}(\varsigma^{-1})}\right)_{\text{ordered}} = T_{\text{ordered}}$. Therefore, we have shown that

$$
(\varsigma_1(\kappa_1), \ldots, \varsigma_p(\kappa_p), T_{\text{ordered}}) \text{ and } (\kappa_1, \ldots, \kappa_p, T_{\text{ordered}}) \text{ are identically distributed}. \tag{B2}
$$

For any $j \in \mathscr{H}_0$, now we further assume that ς_k is an identity permutation for all $k \neq j$, and ς_j is an arbitrary permutation on the set $\{1, \ldots, \omega_j\}$. The equality in joint distributions (B2) implies that $\varsigma_j(\kappa_j)$ has the same distribution as κ_j. Since ς_j is an arbitrary permutation on the set $\{1, \ldots, \omega_j\}$, we have that κ_j is uniformly distributed on the set $\{1, \ldots, \omega_j\}$, that is,

$$
\mathbb{P}(\kappa_j = i) = \omega_j^{-1}\ \forall i \in \{1, \ldots, \omega_j\}. \tag{B3}
$$

Furthermore, for any $i_k \in \{1, \ldots, \omega_k\}$ for $k \neq j$, and $t \in \mathbb{R}^{\sum_\ell \omega_\ell}$,

$$\mathbb{P}\left[\varsigma_j(\kappa_j) = i \mid \bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right] = \frac{\mathbb{P}\left[\varsigma_j(\kappa_j) = i, \bigcap_{k \neq j} \{\varsigma_k(\kappa_k) = i_k\}, T_{\text{ordered}} = t\right]}{\mathbb{P}\left[\bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right]}$$
$$= \frac{\mathbb{P}\left[\kappa_j = i, \bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right]}{\mathbb{P}\left[\bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right]}$$
$$= \mathbb{P}\left[\kappa_j = i \mid \bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right],$$

where the first equality holds from the Bayes formula and the fact that ς_k is the identity permutation for all $k \neq j$, and the second equality holds from (B2). Therefore, for any $i_k \in \{1, \ldots, \omega_k\}$ for $k \neq j$, and $t \in \mathbb{R}^{\sum_\ell \omega_\ell}$, we have that

$$\mathbb{P}\left[\kappa_j = i \mid \bigcap_{k \neq j} \{\kappa_k = i_k\}, T_{\text{ordered}} = t\right] = \omega_j^{-1} \ \forall i \in \{1, \ldots, \omega_j\}. \tag{B4}$$

Combining (B3) and (B4), we have that κ_j is independent of $\{\kappa_k\}_{k \neq j}$ and T_{ordered}.

APPENDIX C: PROOF OF SIMULTANEOUS WFDP BOUND

Without loss of generality, we assume that the ordering in Step 3 of Section 2.2 is such that $\sigma(j) = j$ for $j \in \{1, \ldots, p\}$. Consider

$$\mathcal{V}(\mathcal{R}_k, c) = \frac{c^{-1} + \sum_j \mathbb{1}\{j \notin \mathcal{R}_k\}}{\left(\sum_j \omega_j \mathbb{1}\{j \in \mathcal{R}_k\}\right) \vee 1} = \frac{c^{-1} + \sum_{j=1}^k \mathbb{1}\{\kappa_j > 1\}}{\left(\sum_{j=1}^k \omega_j \mathbb{1}\{\kappa_j = 1\}\right) \vee 1} \tag{C1}$$

for some constant c. Recall that

$$\text{wFDP}(\mathcal{R}_k) = \frac{\sum_j \omega_j \mathbb{1}\{j \in \mathcal{H}_0 \cap \mathcal{R}_k\}}{\left(\sum_j \omega_j \mathbb{1}\{j \in \mathcal{R}_k\}\right) \vee 1} = \frac{\sum_j \omega_j \mathbb{1}\{j \in \mathcal{H}_0\} \mathbb{1}\{\kappa_j = 1\}}{\left(\sum_{j=1}^k \omega_j \mathbb{1}\{\kappa_j = 1\}\right) \vee 1}.$$

We have the following key lemma:

Lemma 3. Let $\mathcal{V}(\mathcal{R}_k, c)$ be defined as in (C1). Then for any $\alpha \in (0,1)$, there exists $x > 0$ such that

$$\mathbb{P}\left[\sup_k \frac{\text{wFDP}(\mathcal{R}_k)}{\mathcal{V}(\mathcal{R}_k, c)} \geq x\right] \leq \alpha. \tag{C2}$$

Proof of Lemma 3. For any $x > 0$, from (C1),

$$
\begin{aligned}
&\mathbb{P}\left\{\sup_k \frac{\mathrm{wFDP}(\mathcal{R}_k)}{\mathcal{V}(\mathcal{R}_k, c)} \geq x\right\} \\
&= \mathbb{P}\left\{\sup_k\left(\sum_{j=1}^{k} \omega_j \mathbb{1}\{\kappa_j = 1\}\mathbb{1}\{j \in \mathcal{H}_0\} - x\sum_{j=1}^{k} \mathbb{1}\{\kappa_j > 1\}\right) \geq c^{-1}x\right\} \\
&\leq \mathbb{P}\left\{\sup_k\left(\sum_{j=1}^{k} \omega_j \mathbb{1}\{\kappa_j = 1\}\mathbb{1}\{j \in \mathcal{H}_0\} - x\sum_{j=1}^{k} \mathbb{1}\{\kappa_j > 1\}\mathbb{1}\{j \in \mathcal{H}_0\}\right) \geq c^{-1}x\right\} \\
&= \mathbb{P}\left[\sup_k \exp\left[\theta\left\{\sum_{j=1}^{k} \omega_j\left(\mathbb{1}\{\kappa_j = 1\} - \frac{x}{\omega_j}\mathbb{1}\{\kappa_j > 1\}\right)\mathbb{1}\{j \in \mathcal{H}_0\}\right\}\right] \geq \exp\left(c^{-1}x\theta\right)\right]
\end{aligned}
$$

for any $\theta > 0$. Define

$$
Z_k = \exp\left[\theta\left\{\sum_{j=1}^{k} \omega_j\left(\mathbb{1}\{\kappa_j = 1\} - \frac{x}{\omega_j}\mathbb{1}\{\kappa_j > 1\}\right)\mathbb{1}\{j \in \mathcal{H}_0\}\right\}\right] \tag{C3}
$$

for $k \geq 1$, and $Z_0 = 1$. Next we find a value of $\theta > 0$ such that $\{Z_k\}$ is a super-martingale with respect to a certain filtration $\mathcal{F}_k$. If such a value of θ exists, then from Ville's maximal inequality for super-martingales (Ville, 1939), we have that

$$
\mathbb{P}\left[\sup_k \frac{\mathrm{wFDP}(\mathcal{R}_k)}{\mathcal{V}(\mathcal{R}_k, c)} \geq x\right] \leq \mathbb{P}\left\{\sup_k Z_k \geq \exp\left(c^{-1}\theta x\right)\right\} \leq \frac{E(Z_0)}{\exp\left(c^{-1}\theta x\right)} = \exp\left(-c^{-1}\theta x\right). \tag{C4}
$$

So it is left to show that Z_k is a super-martingale with respect to a filtration $\mathcal{F}_k$, where $\mathcal{F}_k$ is the σ-field generated from $\{\kappa_j\}_{j \leq k, j \in \mathcal{H}_0}$. First we observe that Z_k is adapted to $\mathcal{F}_k$ for all k. By definition of a super-martingale, it is left to show that

$$
\mathrm{E}\left(\frac{Z_k}{Z_{k-1}} \mid \mathcal{F}_{k-1}\right) = \mathrm{E}\left[\exp\left\{\omega_k\theta\left(\mathbb{1}\{\kappa_k = 1\} - \frac{x}{\omega_k}\mathbb{1}\{\kappa_k > 1\}\right)\mathbb{1}\{k \in \mathcal{H}_0\}\right\} \mid \mathcal{F}_{k-1}\right] \leq 1\,.
$$

First, we observe that this holds trivially for $k \notin \mathcal{H}_0$. For $k \in \mathcal{H}_0$, we have

$$
\begin{aligned}
\mathrm{E}\left(\frac{Z_k}{Z_{k-1}} \mid \mathcal{F}_{k-1}\right) &= \mathrm{E}\left[\exp\left\{\omega_k\theta\left(\mathbb{1}\{\kappa_k = 1\} - \frac{x}{\omega_k}\mathbb{1}\{\kappa_k > 1\}\right)\right\} \mid \mathcal{F}_{k-1}\right] \\
&= \mathrm{E}[\mathbb{1}\{\kappa_k = 1\}\exp(\omega_k\theta) \mid \mathcal{F}_{k-1}] + \mathrm{E}[\mathbb{1}\{\kappa_k > 1\}\exp(-\theta x) \mid \mathcal{F}_{k-1}] \\
&= \exp(\omega_k\theta)\mathbb{P}(\kappa_k = 1 \mid \mathcal{F}_{k-1}) + \exp(-\theta x)\mathbb{P}(\kappa_k > 1 \mid \mathcal{F}_{k-1}) \\
&= \frac{\exp(\omega_k\theta)}{\omega_k} + \frac{(\omega_k - 1)\exp(-\theta x)}{\omega_k},
\end{aligned}
$$

where the last equality holds from Lemma 2.

For any fixed $\alpha \in (0,1)$, take $x = \theta^{-1}\left(-c\log\alpha\right)$, which is equivalent to $\exp\left(-c^{-1}\theta x\right) = \alpha$. Then it remains to select θ such that for all $k \in \mathcal{H}_0$

Author Manuscript
Author Manuscript
Author Manuscript

$$\mathrm{E}\left(\frac{Z_k}{Z_{k-1}} \mid \mathscr{F}_{k-1}\right) = \frac{\exp(\omega_k\theta)}{\omega_k} + \frac{\omega_k - 1}{\omega_k}\exp\left(c\log\alpha\right) \le 1, \tag{C5}$$

which is satisfied for

$$\theta \le \frac{1}{\omega_k}\log\left\{\omega_k - (\omega_k - 1)\alpha^c\right\}.$$

So we take

$$\theta^* = \min_{k \in \mathscr{H}_0} \frac{1}{\omega_k}\log\left\{\omega_k - (\omega_k - 1)\alpha^c\right\}.$$

Then (C5) holds and thus from (C4), the theorem holds with

$$x = \frac{-c\log\alpha}{\theta^*} = -c\log\alpha\left[\max_{k \in \mathscr{H}_0} \frac{\omega_k}{\log\left\{\omega_k - (\omega_k - 1)\alpha^c\right\}}\right]. \tag{C6}$$

Now we have

$$\mathscr{U}(\mathscr{R}_k, c) = x\mathscr{V}(\mathscr{R}_k, c) = -\log\alpha\left[\frac{1 + \sum_{j=1}^{k} c\mathbb{1}\{\kappa_j > 1\}}{\left(\sum_{j=1}^{k} \omega_j \mathbb{1}\{\kappa_j = 1\}\right) \vee 1}\right]\left[\max_{k \in \mathscr{H}_0} \frac{\omega_k}{\log\left\{\omega_k - (\omega_k - 1)\alpha^c\right\}}\right],$$

and the results in Theorem 1 follow.

REFERENCES

Barber RF, & Candès EJ (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics, 43(5), 2055–2085.

Basu P, Cai TT, Das K, & Sun W (2018). Weighted false discovery rate control in large-scale multiple testing. Journal of the American Statistical Association, 113(523), 1172–1183. [PubMed: 31011234]

Benjamini Y, & Heller R (2007). False discovery rates for spatial signals. Journal of the American Statistical Association, 102(480), 1272–1281.

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 (Methodological), 57(1), 289–300.

Benjamini Y, & Hochberg Y (1997). Multiple hypotheses testing with weights. Scandinavian Journal of Statistics, 24(3), 407–418.

Benjamini Y, & Yekutieli D (2001). The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics, 29(4), 1165–1188.

Bien J (2016). The simulator: An engine to streamline simulations. ArXiv e-prints: 1607.00021.

Blanchard G, & Roquain E (2008). Two simple sufficient conditions for FDR control. Electronic Journal of Statistics, 2, 963–992.

Candes 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 (Statistical Methodology), 80(3), 551–577.

Emery K, Hasam S, Noble WS, & Keich U (2019). Multiple competition based FDR control. arXiv preprint arXiv:1907.01458.

Goeman JJ, & Solari A (2011). Multiple testing for exploratory research. Statistical Science, 26(4), 584–597.

Kachuee M, Goldstein O, Karkkainen K, Darabi S, & Sarrafzadeh M (2019). Opportunistic learning: Budgeted cost-sensitive learning from data streams. arXiv preprint arXiv:1901.00243.

Kachuee M, Karkkainen K, Goldstein O, Zamanzadeh D, & Sarrafzadeh M (2019). Cost-sensitive diagnosis and learning leveraging public health data. preprint. https://arxiv.org/abs/1902.07102

Katsevich E, & Ramdas A (2018). Towards "simultaneous selective inference": Post-hoc bounds on the false discovery proportion. arXiv preprint arXiv: 1803.06790

National Center for Health Statistics (2018). National health and nutrition examination survey. https://www.cdc.gov/nchs/nhanes

Ramdas A, Barber R, Wainwright M, & Jordan M (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. The Annals of Statistics, 47(5), 2790–2821.

Roquero Gimenez J, & Zou J (2018). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. ArXiv e-prints.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.

Ville J (1939). Etude critique de la notion de collectif. Bulletin of the American Mathematical Society, 45(11),824.

Yu G, Fu H, & Liu Y (2022). High-dimensional cost-constrained regression via nonconvex optimization. Technometrics, 64, 52–64. [PubMed: 36312889]

Author Manuscript

Author Manuscript

Author Manuscript

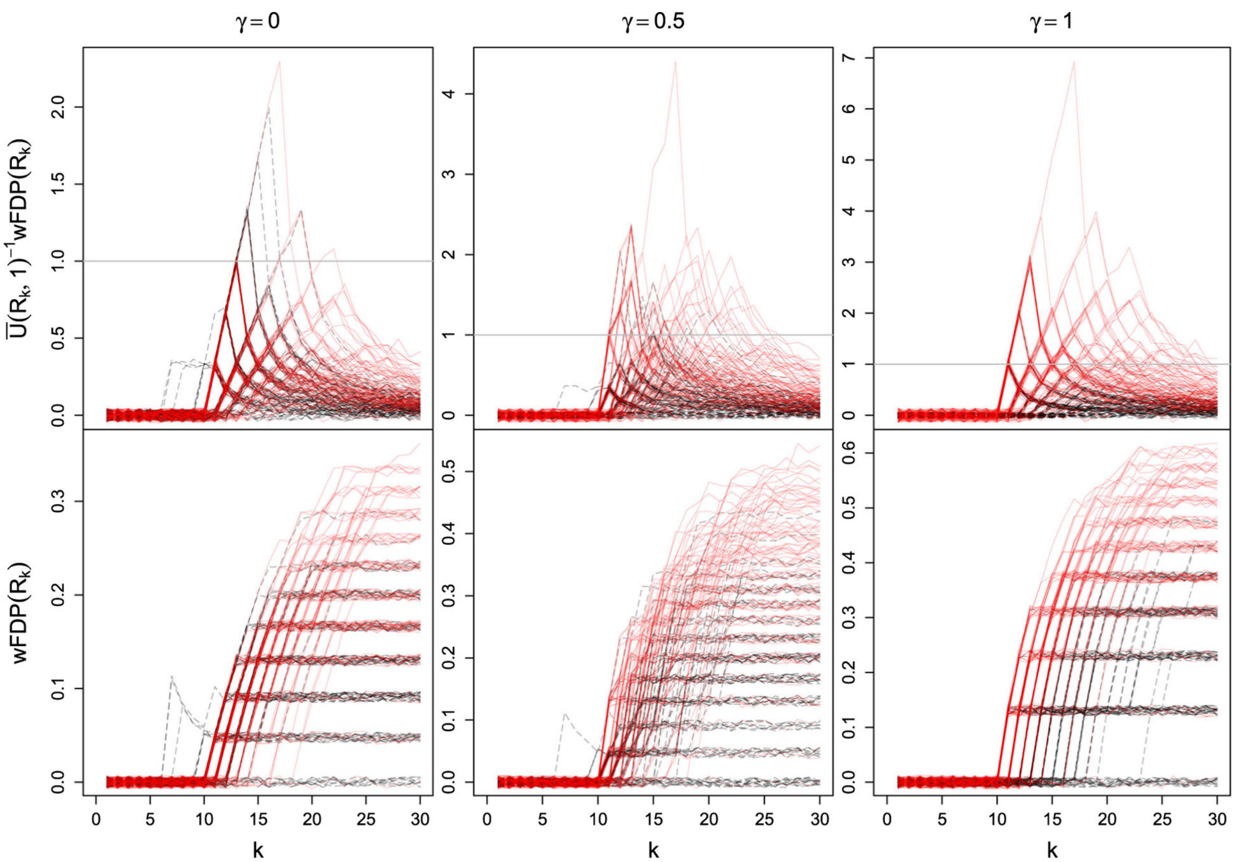

FIGURE 1.

Each line represents one of 100 simulated datasets. Jitter is applied to ease visualization. The black dashed lines represent cheap knockoffs (our proposal) which incorporates feature costs, and the red solid lines represent Katsevich and Ramdas (2018) which does not make use of feature costs. Top panel: The cheap knockoff approach controls the weighted false discovery proportion with the desired probability ($\alpha = 0.2$), whereas the Katsevich and Ramdas (2018) procedure does not. Bottom panel: The cheap knockoff approach attains a lower weighted false discovery proportion than the Katsevich and Ramdas (2018) procedure for most values of k when γ is large

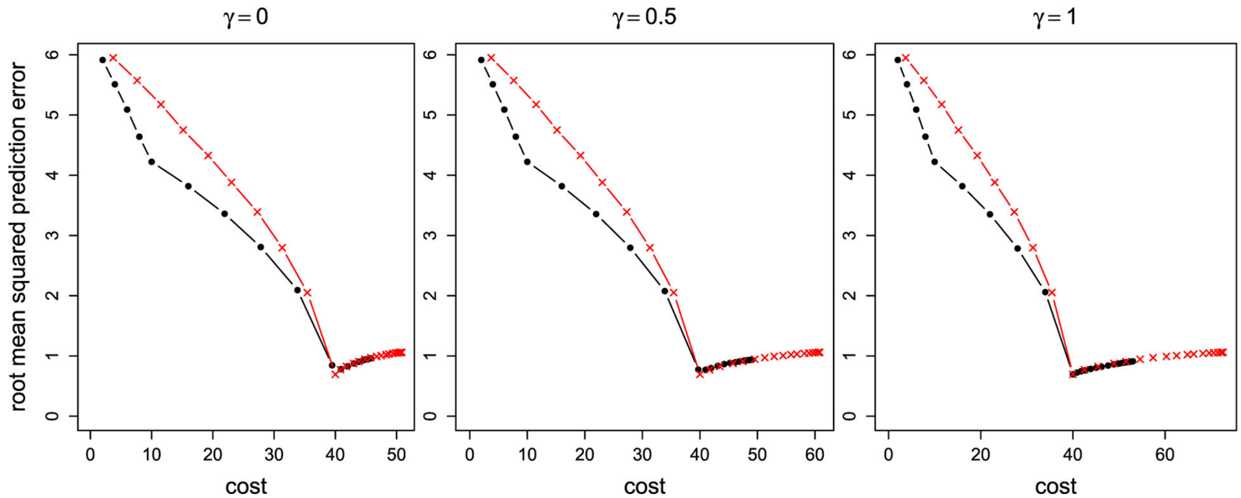

FIGURE 2.

Tradeoff between prediction accuracy and total cost (averaged over 100 simulations). The line with dots in black represents the cheap knockoff procedure, and the line with crosses in red represents Katsevich and Ramdas (2018). The cost of the model selected by our cost-conscious procedure can be much lower than that of the procedure in Katsevich and Ramdas (2018) without sacrificing predictive performance randomly divide these 20,000 samples (with simulated responses) into 50 non-overlapping sets, each containing 400 samples. On each set, we run our method to obtain a path of selected variables. Finally, we compute the estimated probability that the bound in (6) is violated, that is, $\hat{\mathbb{P}}\left(\sup_k \overline{\mathscr{U}}_k^{-1}(\mathscr{R}_k, 1)\mathrm{wFDP}(\mathscr{R}_k) > 1\right)$ for $\alpha \in \{0.05, 0.1, \ldots, 0.5\}$. We see from Table 3 that the estimated probability is lower than the corresponding value of α, indicating that Theorem 1 holds for our proposed cost-conscious procedure.

Author Manuscript
Author Manuscript
Author Manuscript

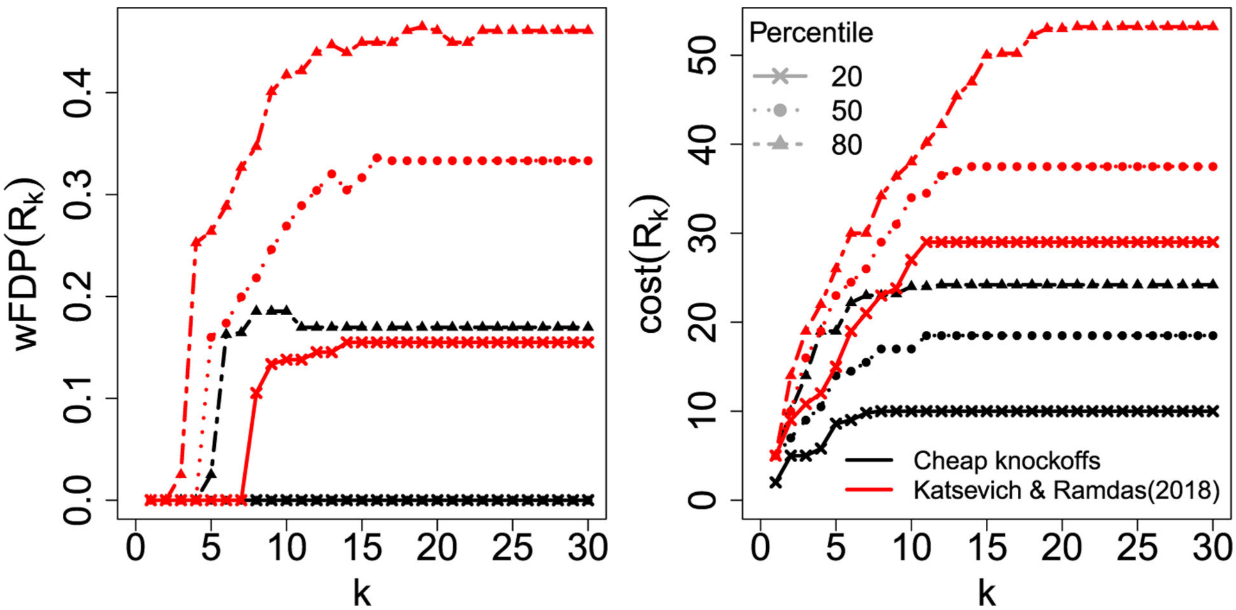

FIGURE 3.

The 20th, 50th and 80th percentiles of wFDP (left panel) and cost (right panel) over 50 non-overlapping data subsets of cheap knockoffs and the procedure in Katsevich and Ramdas (2018)

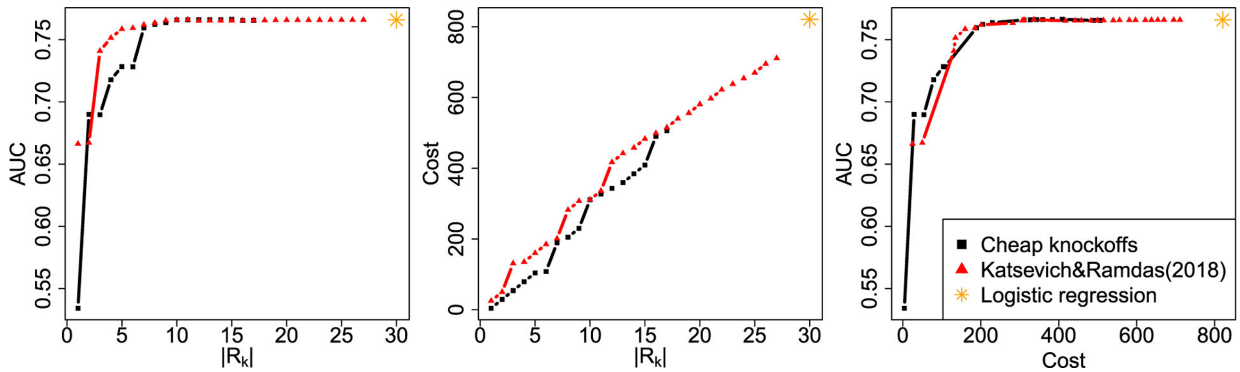

FIGURE 4.

Left: The classification performance (in terms of the area under the ROC curve) for different sizes of the selected model $\mathscr{R}_k (k = 1, \ldots, 30)$. Center: The total cost for different sizes of the selected model. Right: The classification performance versus the cost of the selected model. In all three panels of this figure, we consider the squared costs to highlight the effects of feature costs

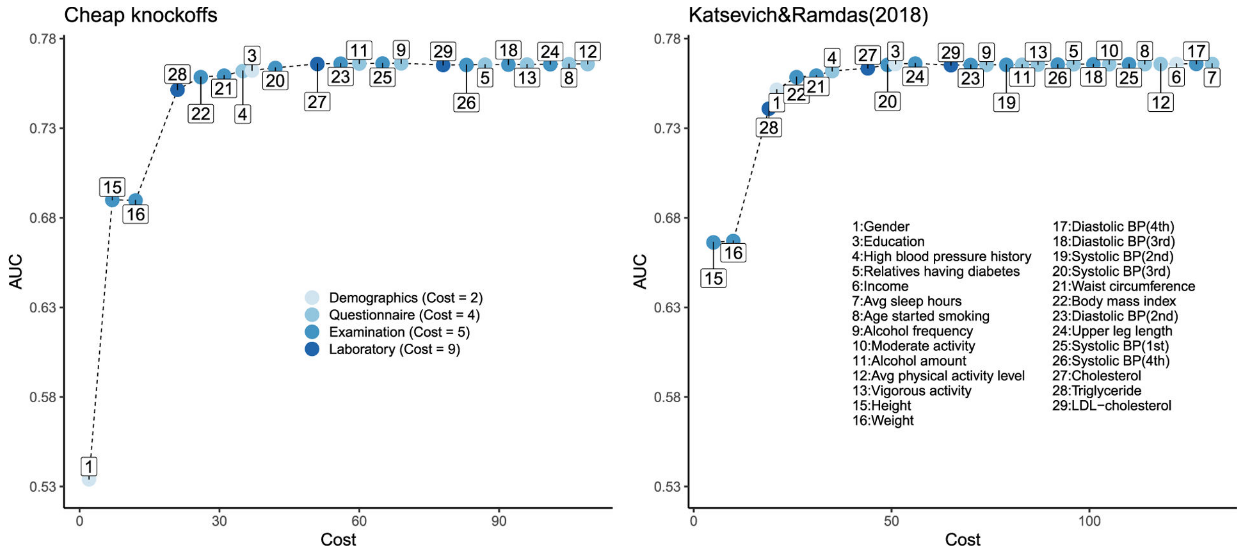

FIGURE 5.

The path of variables selected by cheap knockoffs (top) and the proposal of Katsevich and Ramdas (2018) (bottom). Each point represents a newly selected feature in the model. Variable indices are ordered from cheapest to most expensive

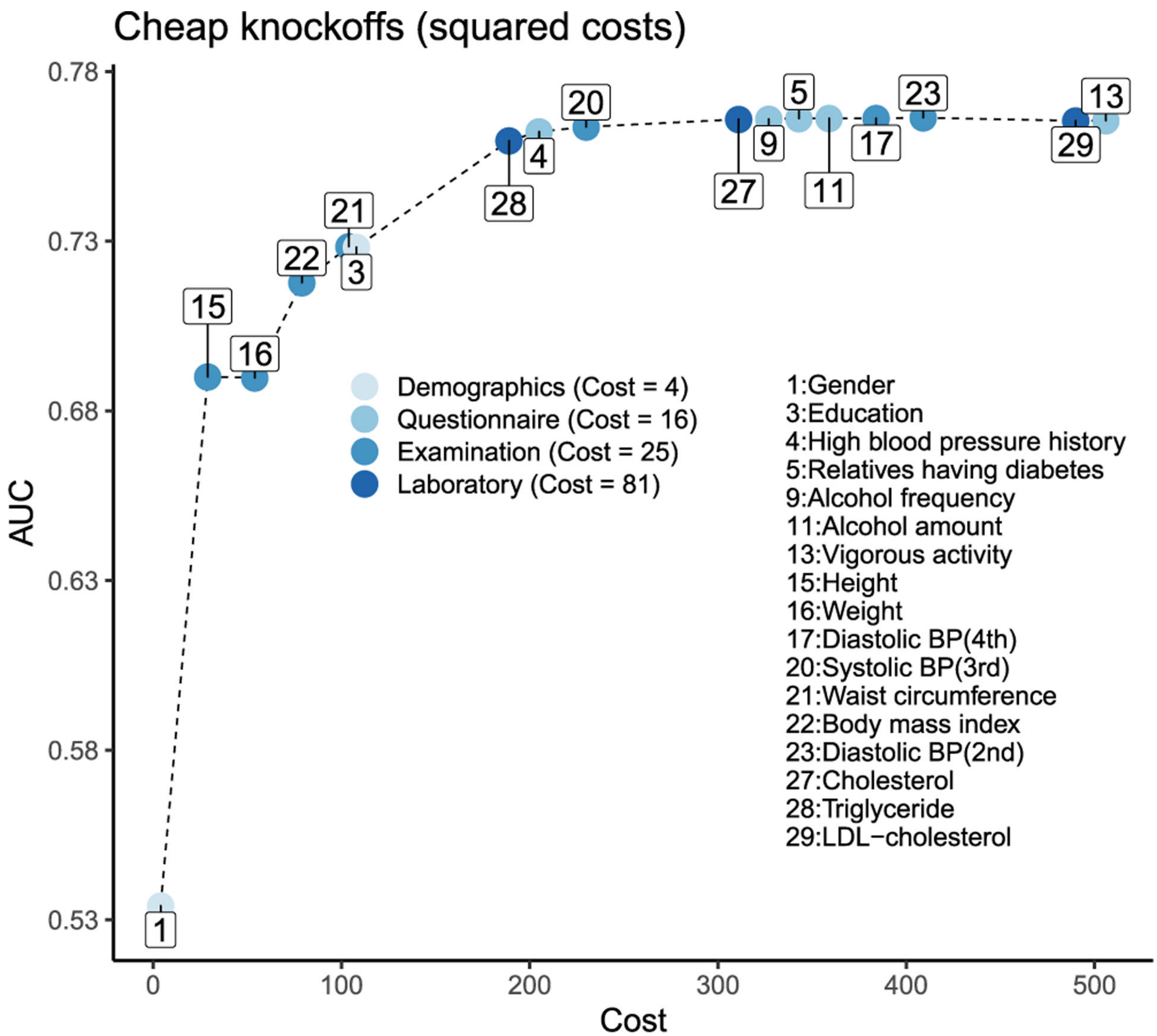

FIGURE 6.

The path of variables selected by cheap knockoffs, with squared costs. Each point represents a newly selected feature in the model. Variable indices are ordered from cheapest to most expensive

TABLE 1

Proportion of 100 simulated datasets for which $\sup_k \overline{\mathcal{U}}_k^{-1}(\mathcal{R}_k, 1)\text{wFDP}(\mathcal{R}_k) > 1$ is violated

γ	**0**	**0.25**	**0.5**	**0.75**	**1**
Cheap knockoffs (our proposal)	0.08	0.05	0.08	0.07	0.04
Katsevich and Ramdas (2018)	0.01	0.05	0.12	0.25	0.31

Note: Our proposed cost-conscious procedure successfully controls the probability below the $\alpha = 0.2$ level for all values of γ, while Katsevich and Ramdas (2018) does not control this probability when $\gamma = 0.75$ and $\gamma = 1$.

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 2

Examples of the features in the NHANES dataset

	Examples	Cost
Demographics	Age; Income; Education level	2 to 4
Questionnaire	Average sleep length (in hours)	4
Examination	Diastolic Blood pressure; Systolic Blood Pressure	5
Laboratory	Cholesterol; Triglyceride; Fibrinogen	9

Author Manuscript

Author Manuscript

Author Manuscript

TABLE 3

Proportion of 50 data subsets for which $\sup_k \overline{\mathscr{U}}_k^{-1}(\mathscr{R}_k, 1)\text{wFDP}(\mathscr{R}_k) > 1$ is violated

α	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
Cheap knockoffs	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.06

TABLE 4

NHANES dataset: Significant features in logistic regression, in the order of increasing p values (smaller than 0.01 / 30)

Name	p value
Gender	1.73×10^{-262}
Triglyceride	5.92×10^{-214}
Height	1.17×10^{-184}
Weight	1.98×10^{-102}
Waist circumference	4.09×10^{-37}
Body mass index	4.02×10^{-31}
High blood pressure history	1.51×10^{-27}
Cholesterol	4.92×10^{-24}
Education	8.16×10^{-10}
Upper leg length	3.17×10^{-5}
Systolic BP(3rd)	1.01×10^{-4}