Submitted to the Annals of Applied Statistics

A DYNAMIC SCREENING ALGORITHM FOR HIERARCHICAL BINARY MARKETING DATA

BY YIMEI FAN[1], YUAN LIAO[2,*], ILYA O. RYZHOV[3,†] AND KUNPENG ZHANG[3,‡]

[1]*Mathematics, University of Maryland,*

[2]*Economics, Rutgers University,* [*]*yuan.liao@rutgers.edu*

[3]*Robert H. Smith School of Business, University of Maryland,* [†]*iryzhov@umd.edu;* [‡]*kpzhang@umd.edu*

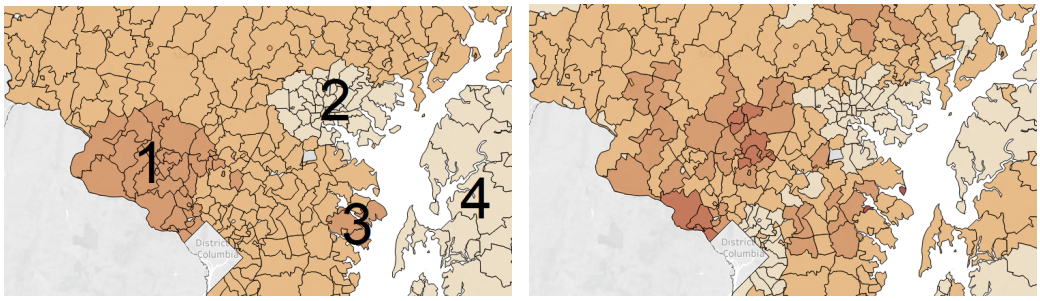
In many applications of business and marketing analytics, predictive models are fit using hierarchically structured data: common characteristics of products, customers, or webpages are represented as categorical variables, and each category can be split up into multiple subcategories at a lower level of the hierarchy. The model may thus contain hundreds of thousands of binary variables, necessitating the use of variable selection to screen out large numbers of irrelevant or insignificant features. We propose a new dynamic screening method, based on the distance correlation criterion, designed for hierarchical binary data. Our method can screen out large parts of the hierarchy at the higher levels, avoiding the need to explore many lower-level features and greatly reducing the computational cost of screening. The practical potential of the method is demonstrated in a case application on user-brand interaction data from Facebook.

1. Introduction. We consider a class of problems in business and marketing analytics in which large-scale statistical predictive models are fit using hierarchically structured data. These data consist of categorical features modeled using large numbers of binary (dummy) variables; many of these categories, however, are subcategories of features at higher levels in the hierarchy, and can themselves be subdivided further at lower levels. Hierarchical aggregation represents common characteristics of large numbers of features, and is widely applicable in revenue management, marketing and other business applications. Consider the following examples:

1. *Customer relationship management*. Negative comments by users on social media platforms can be predicted based on those users' past observed interactions with other brands, topics, and groups of topics on the platform. A firm can then act preventively, e.g., by displaying offers to certain users, in order to reduce the risk of negative word of mouth and maintain a healthy brand image.
2. *Demand modeling*. A retailer sells a wide variety of products. When modeling customer demand as a function of the price, the retailer may also include dummy variables that classify products by department (e.g., tools, electronics, clothes), then describe different categories of products within a given department (e.g., hammers, saws, drills), and finally add features at the individual product level.
3. *Non-profit fundraising*. A non-profit organization is sending out written appeals during a quarterly fundraiser. The non-profit may model donor location at the state level, as well as the level of three- and five-digit zip codes (the latter being used as a stand-in for donor income when detailed demographic information is unavailable).

The size of the feature space in these examples grows dramatically as more levels are added to the hierarchy. In a practical application, we may have tens or hundreds of thousands

Keywords and phrases: sure screening, hierarchical data, distance correlation.



(a) 3-digit zip codes. (b) 5-digit zip codes.

FIG 1. *Example showing household income by 3-digit and 5-digit zip code.*

of binary variables representing hundreds or thousands of categories. At the same time, most (but not all) of the features at the disaggregate levels may have no effect on the dependent variable of interest; moreover, the presence of these features adds noise that confounds our ability to make accurate statistical predictions (Fan, Han and Liu, 2014). Ideally, the hierarchical structure itself can help to resolve this problem, by specifying exactly how much detail is needed to make accurate predictions for different segments. For example, if we are modeling the demand for tools, it may be sufficient to include a single variable for saws, but necessary to distinguish between several individual brands of hammers.

To illustrate, Figure 1 shows a map of the Baltimore-Washington area, segmented into 5-digit zip codes, which can be aggregated by their first three digits. Figure 1(a) shows average household income by 3-digit block, while Figure 1(b) colors each 5-digit zip code separately. Block 1 in Figure 1(a) has higher average income than the surrounding area, blocks 2 and 4 have lower income, and block 3 appears to have high income due to a single very wealthy 5-digit zip code. We could model all four blocks at the 3-digit level, with additional variables as needed at the 5-digit level to capture the main sources of variability. Outside these blocks, there are wide areas with similar incomes that could be modeled at the 3-digit or even state level. Although some useful zip codes may be lost in aggregation, we have still identified several important regions and modeled them in varying degrees of detail. This is a natural and practical way to approach a large problem with a multi-layered feature space. For example, in B2B pricing, managers overseeing the sales of 50,000 distinct products would seek to identify certain product families where differentiation is critically important, as opposed to others that can safely be represented in aggregate.

The difficulty is that, in large problems, we do not know which segments should be aggregated, and how much aggregation can be used. In such situations, statistical model selection (also known as variable selection) becomes an extremely useful practical tool for reliably recovering a sparse set of significant features, while removing large numbers of insignificant features.[1] Model selection is also very useful for practical computation: when both the sample size n and the size p of the feature space are large, traditional estimation procedures may run into severe computational difficulties (Kleiner et al., 2014). Reducing the feature space mitigates this difficulty and improves predictive power.

The main contribution of this paper is a new model selection algorithm that takes a hierarchical data structure and extracts from it a subset of features that is also hierarchically ordered. The method explores the hierarchy from higher (more aggregate) to lower (more

[1] In this paper, words such as "significant," "important," "relevant" etc. implicitly refer to predictive power. We do not consider causal relationships in this paper, and assume throughout that our goal is to identify features that improve prediction, while removing those that do not.

disaggregate) levels. Informally, we estimate the relevance of a candidate feature based on the data. If the estimated relevance is sufficiently high, the feature is accepted and its children (subcategories) become candidates; however, if the relevance is too low, the feature is screened out together with all of its descendants. This approach leads to very substantial computational savings on large problems, as most of the disaggregate features are screened out at higher levels without ever being directly examined.

To obtain these gains, we first assume that the relevance of a feature can be captured by a measure of its marginal dependence on the response variable. This assumption is the foundation of an entire stream of methodological literature in statistics, known as *sure independence screening* or SIS (Fan and Lv, 2008). We leverage this literature in our work, using the distance correlation (DC) criterion of Székely, Rizzo and Bakirov (2007) as our measure of dependence. The statistical literature has shown that this criterion is valid for a very general class of models, so we do not need to impose any particular functional form on the relationship between the predictors and response; we can handle any statistical model where the data are binary. In the process, however, we prove that DC is equivalent to classical Pearson correlation in the binary setting, which allows the criterion to be computed more efficiently and provides a conceptual bridge between these two notions of correlation.[2]

Second, in order for us to exploit hierarchical data structures, the hierarchy has to be informative to begin with. We allow the set of relevant features to deviate from this structure to some degree, similar to how block 3 in Figure 1(a) only appears to be relevant because of one 5-digit zip code. Such an "indirectly relevant" feature would also be discovered by our method; however, once an aggregate feature appears to be irrelevant, we will not proceed further down the hierarchy. We prove that, if the marginal DCs of the features obey the hierarchy in this way, our procedure will recover all of the relevant features under a standard set of assumptions from the statistical literature. It follows that, by automatically screening out all descendants of an irrelevant feature, we provably reduce the number of false positives.

The dynamic DC-based algorithm (DDC) shows significant advantages in numerical experiments.[3] We study both simulated and real data, the latter consisting of historical interactions with various topics by users on Facebook, which we use to predict whether users will write negative comments for a particular brand. Although $n > p$ in this dataset, analysis is non-trivial since $n \sim O\left(10^5\right)$ and $p \sim O\left(10^4\right)$. We find that DDC scales much better to this large dataset, in terms of both predictive power and computational efficiency, than do various statistical benchmark methods. These results also constitute empirical evidence that hierarchical structure can be very effective in practical applications, as DDC significantly outperforms those benchmarks that do not consider this structure at all.

2. Literature Review. We place our work in the context of the vast literature on variable selection. Most of these references pertain to statistical and machine learning methodology; however, variable selection is increasingly used in business analytics and operations research applications (Rudin et al., 2012; Bertsimas et al., 2016; Ryzhov, Han and Bradić, 2016; Li, Netessine and Koulayev, 2018), where predictive models work together with optimization or other decision-theoretic tools. We do not directly study optimization in this paper, but our work could complement, e.g., Xue, Wang and Ettl (2016) or Qu et al. (2019), where statistical models are embedded inside optimization problems.

Within the statistical literature, our paper is closest to the work on sure independence screening (SIS), a methodology first proposed by Fan and Lv (2008) for linear regression

[2]If the data are not binary, our overall approach is still potentially applicable. We will simply have to calculate DC according to its original definition.

[3]In compliance with the journal's data disclosure policy, we have made our data and code available for replication at this website: https://github.com/ddcfs2019/DDC

problems and subsequently extended to GLMs (Fan and Song, 2010), nonparametric models (Fan, Feng and Song, 2011), survival models (Zhao and Li, 2012) and other settings (Zhu et al., 2011). Unlike regularization-based approaches such as Lasso (Tibshirani, 1996), SIS treats the problem of model selection separately from estimation. As is typical in the model selection literature, one first assumes that $p \gg n$, but that the response variable is only influenced by a small subset $\mathcal{A}$ of the features. The fundamental assumption of SIS is that this influence is reflected in the marginal dependence of the response on individual features, which can be measured using, e.g., Pearson correlation; thus, one screens out a feature if its estimated marginal dependence is sufficiently small. It is possible to consider more complex forms of joint dependence (Fan, Samworth and Wu, 2009; Barut, Fan and Verhasselt, 2016), but this increases computational cost, so the SIS literature predominantly focuses on marginal dependence.

Different statistical settings necessitate the use of different measures of dependence in order to prove the validity of the screening procedure, i.e., that the procedure recovers $\mathcal{A}$ with high probability. For example, Fan and Lv (2008) proves this using Pearson correlation, but is restricted to linear regression (OLS) models. In GLMs, Fan and Song (2010) proposed solving a marginal maximum likelihood problem for every feature (the streamwise selection method of Zhou et al., 2006 uses a similar idea). Overall, many criteria can be found in the literature (for example, Li et al., 2012 proposed to use Kendall rank correlation), but they tend to be valid for some models and not others. An important exception is the so-called *distance correlation* (DC) criterion of Székely, Rizzo and Bakirov (2007), which provides a very general measure of dependence (Székely and Rizzo, 2009; Székely and Rizzo, 2012), and can be estimated in a purely data-driven way without having to assume any specific relationship between the features and response (Huo and Székely, 2016). It was proved by Li, Zhong and Zhu (2012) that, when DC is used for screening, SIS is valid across a very general class of statistical models. Because we do not wish to restrict ourselves to any specific class a priori, we also adopt DC as the screening criterion for our procedure; in the process, we discover that it admits substantial computational simplifications in our motivating setting of binary data.

The SIS literature has not considered hierarchical data structures. The work by Hao and Zhang (2014) is perhaps the closest to our paper in that regard: it assumes a linear regression model with interaction terms whose values are products of pairs of "base" features, and one performs model selection, using screening or some other approach (Hao and Zhang, 2016), in two stages, so that interactions can only be selected if one or both of the base components are. This work cannot be directly applied to our setting, as we do not use linear regression and the hierarchy in our problem may be multi-layered. Similar data structures can also be handled using group Lasso methods (Yuan and Lin, 2006), which jointly perform selection and estimation using regularized optimization. Specifically, Zhao, Rocha and Yu (2009), Bach et al. (2012), and Yan and Bien (2017) all consider group structures that could potentially be applied to hierarchical data, while Kim and Xing (2010, 2012) explicitly study tree structures. However, all group Lasso methods, in order to handle our setting, would need to enumerate and include a separate penalty term for every subtree in the hierarchy, which would not scale well to multi-layered business data in which both n and p could be large.

Another approach, based on hypothesis testing, was proposed by Yekutieli (2008). In the language of our paper, the decision to screen out a feature can be made based on a hypothesis test (Fan and Fan, 2008), with the null hypothesis being that the feature is uncorrelated with the response. As in our work, one then refrains from testing any descendants of a feature that has already been screened out. Using the methodology of Benjamini-Hochberg false discovery rate (FDR) control (Ferreira and Zwinderman, 2006), one can guarantee that the proportion of false positives among the selected features is kept below some desired threshold.

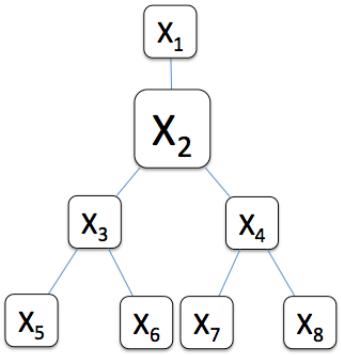


FIG 2. *Illustration of a hierarchical data structure.*

This approach has much of the same computational appeal as our own, and we extensively compare against it in numerical experiments. However, we find that our proposed approach performs more robustly and scales better to large problems; perhaps it is worth noting that FDR control requires weak dependence between hypotheses (which need not be the case in our setting) in order to derive guarantees (Liu and Shao, 2014; Fan and Han, 2017).

3. Data and Model. Let there be n observations $(\mathbf{x}_1, y_1), ...(\mathbf{x}_n, y_n)$ that are independent and identically distributed. We let $\mathbf{X} = (X_1, ..., X_p)$ denote a generic feature vector, with p being the number of features, while Y is used to denote a generic response. We assume that Y and each component of $\mathbf{X}$ are binary-valued (zero/one). Let $F(y \,|\, \mathbf{X}) = \mathbb{P}(Y = y \,|\, \mathbf{X})$ be the conditional probability of observing the response $y \in \{0, 1\}$ given $\mathbf{X}$. Without specifying any particular regression model (thus, $\mathbb{E}(Y)$ does not have to be linear in $\mathbf{X}$), we define the sets of "relevant" and "irrelevant" features as

$$\mathcal{A} = \{j \leq p : F(Y|\mathbf{X}) \text{ functionally depends on } X_j \text{ for some } Y.\}$$

$$\mathcal{A}^c = \{j \leq p : F(Y|\mathbf{X}) \text{ is functionally independent of } X_j \text{ for any } Y.\}$$

The goal is to identify $\mathcal{A}$ while removing as much of $\mathcal{A}^c$ as possible.

We now impose a hierarchical structure on the features. For $j = 1, ..., p$, we use $\mathcal{P}(j)$ to denote its "parent," which is understood as a set containing a single index. For features that belong to the top layer of the hierarchy, we may have $\mathcal{P}(j) = \emptyset$ as a special case. We further define $\mathcal{C}(j)$ to be the index set of all the "children" of the jth feature (i.e., $k \in \mathcal{C}(j)$ if and only if $\mathcal{P}(k) = j$), and $\mathcal{D}(j)$ to be the index set of all the descendants of the jth feature. Thus, $\mathcal{C}(j) \subseteq \mathcal{D}(j)$. For instance, in the example shown in Figure 2, we have $\mathcal{P}(2) = \{1\}$, $\mathcal{C}(2) = \{3, 4\}$ and $\mathcal{D}(2) = \{3, 4, 5, 6, 7, 8\}$.

The hierarchical structure affects the composition of $\mathcal{A}$ through what we call the *extinction property*. Informally, this property says that, if X_j and Y are "weakly" correlated, in a sense that will be precisely defined later, then X_k and Y are also "weakly" correlated for all $k \in \mathcal{D}(j)$. A very strong version of such a property, which can be understood using the notation introduced thus far, is provided as an example.

EXAMPLE (strong extinction property). If $j \in \mathcal{A}^c$, then $k \in \mathcal{A}^c$ for all $k \in \mathcal{D}(j)$.

In words, all descendants of irrelevant features are also irrelevant, so that the set $\mathcal{A}$ is also hierarchically ordered. The weaker extinction property, which we will introduce later and

use in our analysis, relaxes this requirement, but has a similar form: if a feature appears to be only weakly relevant (but no longer has to be a member of $\mathcal{A}^c$), the same is true for its descendants.

Such assumptions have an intuitive appeal in many areas of application. For instance, consider a large online retailer using data to quantify and predict the demand for large numbers of products. The response Y represents whether the customer buys the product ($Y = 1$) or not ($Y = 0$), with $F(1 \mid \mathbf{X})$ being the probability of a sale (a stand-in for demand) given a large number of binary product attributes in $\mathbf{X}$. Thus, one of the features in the top layer of the hierarchy may be "electronics," and the children of this feature may be, respectively, "phones," "cameras," "tablets" and "TVs." Different features may have different numbers of children; for example, if "tools" is another feature in the top layer of the hierarchy, its children will be completely different from those of the "electronics" feature.

The features that are children of "cameras" may be "SLR" and "digital," with further categorization by size one level down. The features that are children of "tablets" may include various operating systems. The children of "TVs" may be different sizes, which can be further broken down by brand. The extinction property implies that, for instance, if a certain size of TV is only weakly correlated with the purchase probability, individual brands of TVs of that same size should not be strongly correlated. However, the weaker form of the property will allow size to be completely irrelevant while some individual brands are (weakly) relevant.

4. Methodology. We now describe our new dynamic screening algorithm for identifying features in $\mathcal{A}$. First, Section 4.1 reviews the DC criterion used by our procedure and proves its equivalence to Pearson correlation for binary data. By using DC as the foundation for our procedure, we do not need to parametrize $F(Y \mid \mathbf{X})$, and thus the proposed method is model-free. Section 4.2 formally states the dynamic algorithm, while Section 4.3 provides a descriptive example illustrating how the procedure exploits the hierarchical structure.

4.1. *Distance Correlation.* We begin by describing the distance correlation (Székely, Rizzo and Bakirov, 2007), which we adopt as the criterion for the relevance of a feature. Let X and Y be scalar random variables with respective characteristic functions $\phi_X(t)$ and $\phi_Y(t)$, and let $\phi_{X,Y}(s,t)$ be their joint characteristic function. The distance covariance between X and Y is given by

$$\operatorname{dcov}(X,Y) = \left(\int \left| \phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t) \right|^2 \left(\pi^2 s^2 t^2\right)^{-1} ds\, dt \right)^{\frac{1}{2}}. \tag{1}$$

The distance correlation is defined as

$$\operatorname{dcorr}(X,Y) = \frac{\operatorname{dcov}(X,Y)}{\sqrt{\operatorname{dcov}(X,X)\operatorname{dcov}(Y,Y)}},$$

and is always non-negative. Li, Zhong and Zhu (2012) showed that this criterion is a very general and powerful measure of dependence: when using distance correlation to evaluate the relevance of features, screening procedures asymptotically recover $\mathcal{A}$ across a very general class of statistical models. Thus, by adopting this criterion for our study, we do not need to assume any particular form for $F(Y \mid \mathbf{X})$. Other choices of criteria would not allow such flexibility; to give one example, the criterion developed in Fan and Song (2010) for screening in GLMs can only be computed with full knowledge of the likelihood function.

Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. samples from the joint distribution of (X,Y). Székely, Rizzo and Bakirov (2007) proposed, and proved the consistency of, the estimator

$$\widehat{\operatorname{dcov}}(X,Y) = \left(\widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3 \right)^{\frac{1}{2}}, \tag{2}$$

$$\widehat{\operatorname{dcorr}}(X,Y) = \frac{\widehat{\operatorname{dcov}}(X,Y)}{\sqrt{\widehat{\operatorname{dcov}}(X,X)\,\widehat{\operatorname{dcov}}(Y,Y)}}, \tag{3}$$

where

$$\widehat{S}_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|\cdot|Y_i - Y_j|$$

$$\widehat{S}_2 = \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|\right)\cdot\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i - Y_j|\right)$$

$$\widehat{S}_3 = \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|X_i - X_l|\cdot|Y_j - Y_l|\,.$$

This estimator is purely data-driven and can be computed without any knowledge of F.

In the special case where both X and Y are binary, we find that (1) is equivalent to the absolute value of their Pearson correlation. Perhaps more surprisingly, (3) is almost surely equivalent to the absolute value of the *sample* Pearson correlation

$$\hat{r} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_x s_y}, \tag{4}$$

where $\bar{X}$ and s_x denote the sample mean and standard deviation of X. This result is stated below; the proof can be found in the Appendix.

PROPOSITION 4.1. *Suppose X, Y take values in $\{0,1\}$, with i.i.d. samples $\{X_i, Y_i\}_{i=1}^n$. Then, the following statements hold:*

(i) $\operatorname{dcov}(X,Y) = 2\,|\operatorname{cov}(X,Y)|$, $\quad \operatorname{dcorr}(X,Y) = |\operatorname{corr}(X,Y)|$;
(ii) $\widehat{\operatorname{dcov}}(X,Y) = \frac{2(n-1)}{n}\,|\widehat{\operatorname{cov}}(X,Y)|$, $\quad \widehat{\operatorname{dcorr}}(X,Y) = |\widehat{\operatorname{corr}}(X,Y)|$,

where $\widehat{\operatorname{cov}}$ and $\widehat{\operatorname{corr}}$ respectively denote the usual sample covariance and correlation.

Proposition 4.1 greatly simplifies the computation of DC, as (4) can be calculated more efficiently than (2)-(3). More importantly, it justifies the use of Pearson correlation in our model-free setting. In previous work on sure independence screening, Pearson correlation was only used in conjunction with linear regression (Fan and Lv, 2008), and other criteria were used to prove the validity of screening in other classes of models. On the other hand, DC was shown to be valid under very general assumptions on the model, which are easily satisfied when the data are binary. Thus, Proposition 4.1 can be viewed as a proof that Pearson correlation is valid in the setting under consideration.

Having defined DC, we can give a precise statement of the weaker version of Example 3. The threshold b_n defined in Assumption 4.2 represents "weak" correlation, and is chosen to converge to zero at a suitable rate as $n \to \infty$. A more detailed justification is deferred until Section 5, but we note here that our theoretical analysis relies on the weak extinction property rather than the strong one mentioned in Example 3.

ASSUMPTION 4.2 (weak extinction property). Define $b_n = \sqrt{\frac{2\log(p\vee n)}{n}}$. If $\operatorname{dcorr}(X_j, Y) = O(b_n)$ for some j, then $\operatorname{dcorr}(X_k, Y) = O(b_n)$ for $k \in \mathcal{D}_j$.

4.2. *Dynamic Distance Correlation (DDC) Algorithm.* We first give an overview of the proposed algorithm before stating it formally. The jth feature is assumed to be relevant if $\operatorname{dcorr}(X_j, Y) \geq K_n$, where K_n is a threshold to be determined. The procedure first considers features at the top level of the hierarchy and screens them based on the empirical DC, so that $\widehat{\operatorname{dcorr}}(X_j, Y) < K_n$ will cause the feature to be screened out. The key to the procedure is that, once j is screened out, we do not examine any feature in $\mathcal{D}(j)$. Conversely, if $\widehat{\operatorname{dcorr}}(X_j, Y) \geq K_n$, we select the feature (i.e., report it as being relevant), whereupon all of its children features $k \in \mathcal{C}(j)$ become "candidates" whose empirical DC is to be evaluated. The algorithm stops once there are no candidates with empirical DC above K_n. This has the effect of substantially saving computational resources when the size of $\mathcal{A}$ is small relative to p.

The precise definition of the cutoff K_n is deferred to Section 5. As will be discussed there, in order for the procedure to be valid, K_n should be slightly larger than the maximum estimation error $\max_{j \leq p} \left| \widehat{\operatorname{dcorr}}(X_j, Y) - \operatorname{dcorr}(X_j, Y) \right|$ of the distance correlations.

We now give a formal statement. Let $\mathcal{S}_\ell$ denote the index set of selected features by stage ℓ of the algorithm, and let $\mathcal{M}_\ell$ denote the index set of the current candidates at stage ℓ. These will be updated dynamically by the procedure.

Step 1 (initialization) Set $\ell = 0$, $\mathcal{S}_0 = \emptyset$, and let $\mathcal{M}_0$ be the indices of the features at the top layer only (that is, all features i satisfying $\mathcal{P}(i) = \emptyset$).

Step 2 (screening) For each $j \in \mathcal{M}_\ell$, compute $\widehat{\operatorname{dcorr}}(X_j, Y)$ and set $\mathcal{M}_\ell = \mathcal{M}_\ell \backslash \{j\}$ if $\widehat{\operatorname{dcorr}}(X_j, Y) < K_n$.

Step 3 (termination) If $\mathcal{M}_\ell = \emptyset$, return $\widehat{\mathcal{A}} = \mathcal{S}_\ell$ and stop. Otherwise, continue.

Step 4 (selection) Find

$$j_\ell = \arg\max_{j \in \mathcal{M}_\ell} \widehat{\operatorname{dcorr}}(X_j, Y), \tag{5}$$

and update

$$\mathcal{S}_{\ell+1} = \mathcal{S}_\ell \cup \{j_\ell\},$$
$$\mathcal{M}_{\ell+1} = (\mathcal{M}_\ell \backslash \{j_\ell\}) \cup \mathcal{C}(j_\ell),$$

where $\mathcal{C}(j_\ell)$ is the set of children of j_ℓ as defined in Section 3.

Step 5 (iteration) Increment ℓ by 1 and return to Step 2.

In the algorithm, $\mathcal{M}_\ell$ is the candidate set containing features to be considered in this step of iterations. Step 2 screens out all candidates whose empirical DC is insufficiently strong to claim relevance; if no candidates remain, step 3 terminates. Otherwise, step 4 adds the "most relevant" of the remaining features to the selection set. This feature, labeled as j_ℓ in (5), is no longer a candidate, but all of its children (if there are any) now become candidates. Equivalently, since relevance is determined based on the marginal DC, step 4 could add *all* of the features in $\mathcal{M}_\ell$ to the selection set; the difference between this approach and the given formulation may be viewed analogously to the difference between breadth-first and depth-first search.

The procedure returns the selection set $\widehat{\mathcal{A}}$, which is different from the *screening set*

$$\widehat{\mathcal{H}} = \{j \in \{1, 2, 3..., p\} : \widehat{\operatorname{dcorr}}(X_j, Y) \geq K_n\}, \tag{6}$$

which includes all features whose empirical DC is above the threshold. It is clear that $\widehat{\mathcal{A}} \subseteq \widehat{\mathcal{H}}$. In the finite-sample setting, there may be j and $k \in \mathcal{D}(j)$ such that $\widehat{\operatorname{dcorr}}(X_j, Y) < K_n$, but $\widehat{\operatorname{dcorr}}(X_k, Y) \geq K_n$. Such a k would be an element of $\widehat{\mathcal{H}}$ but not $\widehat{\mathcal{A}}$. This is a fundamental

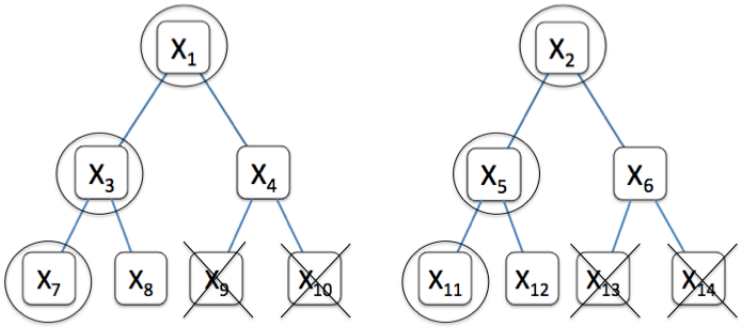


FIG 3. *Illustration of the DDC algorithm. Due to the extinction property, features* 9, 10, 13 *and* 14 *are screened out without being examined directly.*

difference between our dynamic approach and the classical SIS technique of Fan and Lv (2008). SIS arranges the empirical correlations in descending order and simply screens out a certain proportion of features ranked at the bottom. This approach requires us to estimate the marginal correlation for every feature, which may be expensive when p is large; on the other hand, our proposed algorithm automatically rules out all descendants of any feature that has been screened out in step 2. Thus, if the problem is sufficiently sparse, we will avoid having to compute empirical DCs for a substantial proportion of the feature space; consequently, $\widehat{\mathcal{A}}$ will contain fewer false positives than $\widehat{\mathcal{H}}$, which will be formally established in Section 5.

REMARK. Our work is motivated by applications in which the data are binary. Potentially, however, the above-described dynamic approach may be useful for other discrete and continuous features where an analog of the extinction property is assumed to hold. In such cases, other nonparametric measures of relevance may be useful, such as the marginal mean regression function $\mathbb{E}(Y \mid X_j)$ or the Kendall τ based robust correlation (Li et al., 2012).

4.3. *Descriptive Example.* To illustrate our algorithm, we briefly discuss a descriptive example on a hierarchy with three levels shown in Figure 3. As there are two features in the top layer, we initialize $\mathcal{M}_0 = \{1, 2\}$ and $\mathcal{S}_0 = \emptyset$.

Iteration 1: steps 2-5. We first evaluate the empirical DC for features 1 and 2. Suppose that $\widehat{\operatorname{dcorr}}(X_1, Y) > \widehat{\operatorname{dcorr}}(X_2, Y) > K_n$. Then, both features remain in the candidate set during step 2, and step 3 does not terminate. Step 4 sets $j_0 = 1$ since feature 1 has the largest DC among the candidates. We move feature 1 to the selection set, and add the elements of $\mathcal{C}(1) = \{3, 4\}$ to the candidate set, leading to

$$\mathcal{S}_1 = \{1\}, \qquad \mathcal{M}_1 = \{2, 3, 4\}.$$

Iteration 2: steps 2-5. Suppose $\widehat{\operatorname{dcorr}}(X_3, Y) > \widehat{\operatorname{dcorr}}(X_2, Y) > K_n$, but $\widehat{\operatorname{dcorr}}(X_4, Y) < K_n$. Then, step 2 screens out feature 4, whence $\mathcal{M}_1 = \{2, 3\}$, but step 3 does not terminate. Step 4 sets $j_1 = 3$, whence feature 3 is moved to the selection set and the elements of $\mathcal{C}(3) = \{7, 8\}$ become candidates, leading to the update

$$\mathcal{S}_2 = \{1, 3\}, \qquad \mathcal{M}_2 = \{2, 7, 8\}.$$

Iteration 3: steps 2-5. Suppose $\widehat{\operatorname{dcorr}}(X_2, Y) > \widehat{\operatorname{dcorr}}(X_7, Y) > K_n$ but $\widehat{\operatorname{dcorr}}(X_8, Y) < K_n$. Then, step 2 screens out feature 8, whence $\mathcal{M}_2 = \{2, 7\}$. Step 3 does not terminate, step

4 sets $j_2 = 2$, whence feature 2 is selected and the new candidates $\mathcal{C}(2) = \{5, 6\}$ are added. The resulting update is

$$\mathcal{S}_3 = \{1, 2, 3\}, \qquad \mathcal{M}_3 = \{5, 6, 7\}.$$

Iteration 4: steps 2-5. Suppose $\widehat{\text{dcorr}}(X_5, Y) > \widehat{\text{dcorr}}(X_7, Y) > K_n$ but $\widehat{\text{dcorr}}(X_6, Y) < K_n$. At the end of this iteration, we will have

$$\mathcal{S}_4 = \{1, 2, 3, 5\}, \qquad \mathcal{M}_4 = \{7, 11, 12\}.$$

Iteration 5: steps 2-5. Suppose $\widehat{\text{dcorr}}(X_{11}, Y) > \widehat{\text{dcorr}}(X_7, Y) > K_n$ but $\widehat{\text{dcorr}}(X_{12}, Y) < K_n$. At the end of this iteration, we will have

$$\mathcal{S}_5 = \{1, 2, 3, 5, 11\}, \qquad \mathcal{M}_5 = \{7\}.$$

Note that the candidate set shrinks in this iteration since $j_5 = 11$ and $\mathcal{C}(11) = \emptyset$.

Iteration 6: steps 2-5. Since $\widehat{\text{dcorr}}(X_7, Y) > K_n$, feature 7 is selected. As $\mathcal{C}(7) = \emptyset$, we obtain

$$\mathcal{S}_6 = \{1, 2, 3, 5, 7, 11\}, \qquad \mathcal{M}_5 = \emptyset.$$

Iteration 7: steps 2-3. Since the candidate set $\mathcal{M}_5$ is empty, step 3 terminates.

Observe that the procedure never calculates the DCs for features 9, 10, 13, and 14, since their parent features were screened out in earlier iterations. This leads to increased computational savings when the hierarchy has many layers.

5. Theoretical Analysis. We begin by choosing the threshold

$$K_n = \frac{a_0}{\min_{j \le p} \widehat{\text{dcov}}(X_j, X_j)^{1/2} \widehat{\text{dcov}}(Y, Y)^{1/2}} b_n, \tag{7}$$

where b_n is as in Assumption 4.2 and $a_0 > 3.5$ is a constant (we can take, for instance, $a_0 = 3.51$). We will first motivate this definition and explain why it is needed for theoretical analysis, then discuss practical concerns.

In the analysis of threshold-based screening methods (i.e., the entire literature on sure independence screening), the main technical issue is to choose K_n in such a way that, with high probability,

$$\max_{j \le p} \left| \widehat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| < K_n.$$

In order to do this, it is necessary to control for the so-called "uniform deviation"

$$S_n = \max_{j \le p} \left| \widehat{\text{dcov}}(X_j, Y) - \text{dcov}(X_j, Y) \right|,$$

which represents the noise in the problem. Accordingly, we select the threshold to dominate the noise, that is, we choose $b_n \to 0$ so that $P(S_n < a_0 b_n) \to 1$ for some positive constant a_0. It is well-known from the theory of moderate deviations for self-normalized sums (Peña, Lai and Shao, 2008 and Belloni et al., 2012) that the choice of b_n stated in Assumption 4.2 can uniformly control the self-normalized noise. As for the constant, we show in the Appendix that $a_0 > 3.5$ suffices to control for the distance correlations. We then standardize $a_0 b_n$ using the estimated distance covariances, leading to our proposed threshold value in (7).

From a purely practical point of view, (7) requires us to compute $\min_j \widehat{\text{dcov}}(X_j, X_j)$, which may be demanding for large p, and conflicts with our desire to avoid having to examine every feature. Several options are available. First, a practitioner may replace $\min_j \widehat{\text{dcov}}(X_j, X_j)$ with an empirical estimate, e.g., by sampling a small subset of $\{1, ..., p\}$

and calculating the smallest $\widehat{\mathrm{dcov}}(X_j, X_j)$ within this subset. A second alternative is to replace $\min_j \widehat{\mathrm{dcov}}(X_j, X_j)$ in (7) with some function of p and n that declines more slowly than the empirical DCs (i.e., an asymptotic upper bound on $\min_j \widehat{\mathrm{dcov}}(X_j, X_j)$). Finally, in the absence of any such information, a practitioner can simply treat K_n as a tunable parameter; we do exactly this in our case study (Section 6.3) and obtain good performance.

We now proceed to the theoretical analysis. The proofs of all results in this section are given in the Appendix; here, we state the necessary definitions and assumptions. The first assumption simply ensures that we are in the high-dimensional setting, as is standard in the model selection literature.

ASSUMPTION 5.1. The data $\{Y_i, X_{i1}, ..., X_{ip}\}_{i=1}^n$ are independent and identically distributed. As $n \to \infty$, the number of features, p, either stays constant or grows with n, satisfying $\log p = o\left(n^{\frac{1}{4}}\right)$.

The second assumption ensures that we are able to separate the signal in the data from the noise. For this, it is necessary for relevant features to be sufficiently strongly correlated with the response; at the same time, the variances of the features are allowed to slowly decay to zero as $n \to \infty$. In other words, it is possible for individual features to be observed rarely, as long as the sample size is large.

ASSUMPTION 5.2. The following statements hold:

(i) $\mathrm{var}(Y) \gg b_n$ and $\min_{j \le p} \mathrm{var}(X_j) \gg b_n$.
(ii) $\min_{j \in \mathcal{A}} \mathrm{dcov}(X_j, Y) \cdot \min_j \mathrm{var}(X_j)^{1/2} \gg b_n$.

Under the above conditions, we show that the empirical distance correlation converges in probability to its population counterpart uniformly in $j = 1, ..., p$. These conditions also imply $\max_{j \in \mathcal{A}^c} \widehat{\mathrm{dcorr}}(X_j, Y) = O_P(K_n)$, and that $\min_{j \in \mathcal{A}} \widehat{\mathrm{dcorr}}(X_j, Y)$ is bounded away from K_n with probability approaching 1.

Recalling that $\widehat{\mathcal{A}}$ denotes the final selection returned by the proposed algorithm, we can now state the main feature selection guarantee. Define

$$\mathcal{B} = \{j \in \mathcal{A}^c : \mathcal{D}(j) \cap \mathcal{A} \neq \emptyset\}.$$

In words, $\mathcal{B}$ collects those features j which do not have any relevance to Y on their own (that is, $j \in \mathcal{A}^c$), but which have descendants $k \in \mathcal{D}(j)$ that are relevant to Y. We can refer to such $j \in \mathcal{B}$ by the name "indirectly relevant features." Our main result shows that $\widehat{\mathcal{A}}$ captures both directly and indirectly relevant features.

THEOREM 5.3. *Under Assumptions 4.2, 5.1, and 5.2,*

$$\mathbb{P}\left((\mathcal{A} \cup \mathcal{B}) \subseteq \widehat{\mathcal{A}}\right) \to 1.$$

Additionally, the selection set $\widehat{\mathcal{A}}$ will also be structured hierarchically, that is, $j \in \widehat{\mathcal{A}}$ implies that $i \in \widehat{\mathcal{A}}$ for all i satisfying $j \in \mathcal{D}(i)$, and the DDC procedure will calculate $O\left(\sum_{j \in \widehat{\mathcal{A}}} |\mathcal{C}(j)|\right)$ empirical distance correlations before terminating.

Thus, the set $\widehat{\mathcal{A}}$ has the "sure screening property" typically studied in the statistical literature, namely that $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ with probability approaching 1. In addition, however, we also select all indirectly relevant features that have one or more relevant descendants, since we

also have $\mathcal{B} \subseteq \widehat{\mathcal{A}}$ with probability approaching 1. Intuitively, if $j \in \mathcal{B}$ and $k \in \mathcal{D}(j) \cap \mathcal{A}$ is its relevant descendant, Assumption 5.2 ensures that k is sufficiently strongly correlated with the response, while Assumption 4.2 also ensures sufficient correlation for j.

At the same time, our proposed algorithm does not conduct an exhaustive search: as soon as we find a feature j that is weakly correlated with the response, we do not continue to explore its descendants. This greatly reduces the computational complexity of the procedure, and yields interpretable results since the output $\widehat{\mathcal{A}}$ is always a tree.

Finally, we can note that, if we are willing to make additional assumptions on the correlation strength of irrelevant features, the consistency result can be strengthened. Essentially, the additional assumption ensures that $\mathcal{B}$ is empty and provides a clear separation between the correlation strengths of relevant vs. irrelevant features, enabling us to select the relevant features with no false discoveries.

COROLLARY 5.4. *Suppose that we are in the setting of Theorem 5.3, and impose the additional assumption that*

$$\max_{j \in \mathcal{A}^c} \operatorname{dcov}(X_j, Y) = \delta \sqrt{\frac{\log(p \vee n)}{n}} \tag{8}$$

for some fixed $\delta > 0$. Then, $\mathbb{P}\left(\mathcal{A} = \widehat{\mathcal{A}}\right) \to 1$.

We can further characterize the advantages of DDC relative to a more traditional sure screening approach that returns the screening set $\widehat{\mathcal{H}}$ defined in (6) with no explicit consideration of the hierarchical structure. Compared to such an approach, DDC can be guaranteed to select fewer false positives under *any* finite sample size, and achieves a smaller false discovery rate in the asymptotic regime of Theorem 5.3, as formalized in the following proposition.

PROPOSITION 5.5. *The following statements are true:*

(i) *For any finite sample size n, $\left|\widehat{\mathcal{A}} \cap \mathcal{A}^c\right| \leq \left|\widehat{\mathcal{H}} \cap \mathcal{A}^c\right|$. Specifically, the following type of false positive is excluded by DDC: if we let*

$$\widehat{\mathcal{G}} = \left\{ j \notin \mathcal{A} : \widehat{\operatorname{dcorr}}(X_j, Y) \geq K_n,\ \exists i \leq p : \widehat{\operatorname{dcorr}}(X_i, Y) < K_n,\ j \in \mathcal{D}(i) \right\},$$

then $\widehat{\mathcal{G}} \subseteq \widehat{\mathcal{H}}$, but $\widehat{\mathcal{G}} \cap \widehat{\mathcal{A}} = \emptyset$.

(ii) *Suppose that we are in the setting of Theorem 5.3. Then, $\frac{|\widehat{\mathcal{A}} \setminus \mathcal{A}|}{|\widehat{\mathcal{A}}|} \leq \frac{|\widehat{\mathcal{H}} \setminus \mathcal{A}|}{|\widehat{\mathcal{H}}|}$.*

(iii) *Suppose that we are in the setting of Corollary 5.4. Then, $\frac{|\widehat{\mathcal{A}} \setminus \mathcal{A}|}{|\widehat{\mathcal{A}}|} = 0$.*

The set $\widehat{\mathcal{G}}$ in Proposition 5.5(i) contains a type of false positive that is quite likely to arise in a finite-sample setting. Since most of the features are concentrated in the bottom layer of the hierarchy (see Figure 3), this is also where we would expect to see the most false positives due to spurious correlation. However, DDC can only accept such a false positive if *all* of its irrelevant ancestor features are false positives, a less likely event. Proposition 5.5(ii) then shows that, asymptotically, the false positive *rate* is reduced for DDC as compared to non-hierarchical screening.

6. Numerical Studies. We assessed the performance of the DDC algorithm as compared to several benchmark methods, which are described in Section 6.1 together with various details related to implementation. Experiments were conducted on both simulated (Section 6.2) and real (Section 6.3) data. In compliance with the journal's data disclosure policy, both types of data and code are available for replication.[4]

Experiments were conducted in Python; thus, computation times are reported for Python code and statistical packages. We used sparse matrix representations with the real data, but found that this did not improve performance on the synthetic data.

6.1. *Experimental Setup.* We implemented the DDC algorithm (treating the threshold K_n as a tunable parameter) together with four benchmarks. The first three are general model selection methods, while the fourth handles hierarchical data structures similarly to DDC.

Lasso (Tibshirani, 1996; Van de Geer, 2008) assumes a particular regression model (we chose logistic regression because the response is binary) and optimizes a penalized likelihood function that encourages eliminating features by setting their regression coefficients equal to zero. Lasso has one tunable parameter that governs the tradeoff between model accuracy and model sparsity.

Streamwise regression or SR (Zhou et al., 2006) performs a univariate (marginal) regression for each individual feature, analogously to the screening approach of Fan and Song (2010). Screening is performed dynamically using two parameters, an initial "budget" W_0 and an "investment" α_Δ that is added to (subtracted from) the budget every time a feature is rejected (accepted); thus we are more likely to accept a feature if the budget is greater. As with Lasso, we used logistic regression as the model inside SR.

Sure independence screening (SIS) uses the same screening criterion (DC/Pearson correlation) that we use in DDC, but calculates this criterion for every feature, without considering the hierarchy, and simply selects a proportion d of features with the highest estimated correlations. Although Fan and Lv (2008) gives several suggestions for how to choose d, we simply treated it as a tunable parameter.

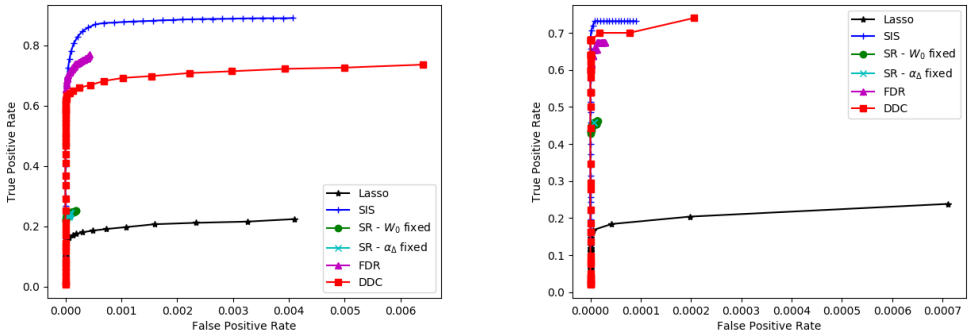
Hierarchical false discovery rate control or FDR (Yekutieli, 2008) explicitly considers the hierarchy by screening features in the same order as DDC. For each feature j, FDR tests the null hypothesis that X_j is uncorrelated with Y using the usual t-statistic; if the null hypothesis is not rejected, then FDR never examines the descendants of j. The procedure has one tunable parameter which essentially governs the threshold for rejecting the hypothesis.

The Lasso method, logistic regression (used inside SR), and k-fold cross-validation (used in Section 6.3) were implemented using the well-known, off-the-shelf Python package `scikit-learn` (`https://scikit-learn/stable`). All experiments were conducted on a machine with 32 GB of memory and a 4GHz Intel Core i7 processor. The specific performance metrics used for each type of experiment will be discussed in Sections 6.2-6.3.

6.2. *Simulated Data.* We generated multiple hierarchical binary data structures satisfying the extinction property. In order to demonstrate how well DDC and benchmark methods scale to larger problems, we considered problems of two sizes: in the first, the hierarchy has five levels and $p \approx 5,500$, and in the second, the hierarchy has six levels with $p \approx 170,000$ features. The sample sizes are $n = 100$ and $n = 1,000$ respectively for the two problem types; note that $p \gg n$ in both cases. The reported results are averaged over 500 randomly generated datasets in the first example, and 50 datasets in the second.

In both cases, the following procedure was applied to generate hierarchical data. The top level of the hierarchy consists of five features, all of which are relevant (correlated with the

[4] `https://github.com/ddcfs2019/DDC`



(a) Small data. (b) Large data.

FIG 4. *Average TPR/FPR curves.*

response). For every feature in level $i = 1, 2, ..., L$, where L is the number of layers in the hierarchy, we generated 2^{i-1} children, resulting in exponential growth of the feature space. For any relevant feature $i \in \mathcal{A}$, its first child is always hard-coded as being relevant, while its other children are irrelevant (members of $\mathcal{A}^c$). Thus, $|\mathcal{A}| = 5L$.

For relevant features $i \in \mathcal{A}$, correlation was ensured in the following manner. First, a quantity κ_i was generated as follows: if feature i belongs to the top layer of the hierarchy, we let κ_i be uniform on the interval $[-0.25, 0.25]$; otherwise, κ_i is uniform on the interval $\left[-|\kappa_{\mathcal{P}(i)}|, |\kappa_{\mathcal{P}(i)}|\right]$. In this way, the correlation is decreasing as we move toward the disaggregate levels of the hierarchy (as we would expect to see in an application). Then, κ_i was used to set the distribution

$$P\left(X_i = 1 \,|\, Y = 1, X_{\mathcal{P}(i)} = 1\right) = \frac{\kappa_i + \frac{1}{2}}{P\left(Y = 1\right)},$$

$$P\left(X_i = 1 \,|\, Y = 0, X_{\mathcal{P}(i)} = 1\right) = \frac{\frac{1}{2}P\left(Y = 0\right) - \kappa_i}{P\left(Y = 0\right)}.$$

To simulate X_i for $i \in \mathcal{A}$, we first sample Y from a Bernoulli distribution with success probability 0.5. Then, if $X_{\mathcal{P}(i)} = 1$, we generate the value of X_i from the above conditional distribution. If $X_{\mathcal{P}(i)} = 0$, we set $X_i = 0$ as is commonly the case in practical applications with hierarchical data (see Section 6.3 for one such application). For $i \notin \mathcal{A}$, we simply generate X_i from an independent Bernoulli distribution with success probability 0.3.

All of the methods listed in Section 6.1 were evaluated using three criteria: a) the true positive rate (TPR), or the proportion of relevant features being selected among all features in $\mathcal{A}$; b) the false positive rate (FPR), or the proportion of irrelevant features being selected among all features in $\mathcal{A}^c$; and c) computation time. In general, a better model will have higher TPR and lower FPR. Computation time is also important, because of the exponential growth of the number of candidate features.

Because all methods are tunable, we ran each method across a range of parameter values and obtained TPR/FPR values (averaged over the randomly generated datasets) for each parameter setting. We then created TPR/FPR curves from these results, shown in Figure 4. We note that SR is the only method that has two tunable parameters, but either parameter essentially changes the scaling of the other. To illustrate, we include two curves for SR in each plot, each of which varies one parameter with the other fixed, and observe that both curves are virtually identical. We also include average computation times for a single execution of each method (for SR we average across both curves) in Table 1.

Method	Small data	Large data
DDC	0.0114	0.2650
Lasso	0.0182	16.4719
SR	4.4838	202.3595
SIS	0.2448	27.5916
FDR	0.0261	0.5744

TABLE 1

Average computation times (in seconds) of all methods on simulated data.

Comparing Figures 4(a)-4(b) is instructive for understanding how different methods scale. In the smaller problem, DDC is outperformed by both SIS and FDR; however, when we move to the larger problem, the advantage of FDR completely disappears, and the performance gap between SIS and DDC shrinks considerably while DDC runs over 100 times faster than SIS. Moreover, in the larger problem, DDC is now able to achieve the highest TPR of any method, though it is at the expense of a larger FPR than SIS. From Table 1, we also see that DDC is the fastest method in both problem types and scales much better than Lasso, SR, and SIS. The FDR method is consistently the second-fastest, but experiences performance degradation on the larger problem. Finally, we note that SR and Lasso do not perform well on either problem. This may be due to the fact that both methods make the additional assumption of a logistic regression model, which may not be well-suited for these particular datasets. Although it may be possible to obtain better performance with another model, one could also argue that the need to make such an assumption is a weakness of these methods, because DDC, SIS and FDR are all able to run without assuming any specific statistical model.

Overall, we conclude that, given its computational cost, DDC is highly competitive with the benchmark methods on high-dimensional problems in which the data are structured hierarchically. Furthermore, the results suggest that DDC scales better to larger problems.

6.3. *Application: Predicting Negative Comments on Social Media.* We applied our method to the problem of predicting the incidence of negative comments about a particular brand, made by Facebook users, based on their previous interactions with other brands and topics. We used Facebook's graph API (`https://developers.facebook.com/docs/graph-api/`) to collect data on user-brand interactions.[5] Specifically, we considered user comments recorded from 4 years before to 3 months after January 1, 2014, for $31,078$ Facebook pages classified into eight categories defined by the Facebook system (brands, celebrities, communities, media etc.). These categories can be divided into 83 subcategories (for example, "communities" can be decomposed into "lifestyle," "hobbies," "fun," "sport-interest" and others), which can be further divided into 51 additional fine-grained categories (for example, "fashion" can be decomposed into "clothing," "accessories," and "jewelry"). All three types of categories are pre-defined by Facebook. These three levels form our hierarchy, with an additional (fourth) bottom layer containing one feature for each individual page; in this way, we can link negative word-of-mouth to user interest in various broad topics, as well as to user engagement with certain specific influential pages.

Next, we consider each comment in the dataset and determine its sentiment (tone) using an ensemble learning algorithm by Zhang, Bhattacharyya and Ram (2016), which has previously been validated with state-of-the-art performance on human-labeled data. The ensemble combines and understands texts from different angles, such as unstructured data, bag-of-words-based linguistic rules, and contexts; the output is a 3-class classification (positive, negative,

[5]Currently, Facebook requires firms to go through a review process in order to access much of the API's functionality. However, at the time when we downloaded the data, the entire API was publicly accessible. The dataset that we have disclosed for replication has been processed as described in this section, with all identifiable information removed.

or neutral). Based on this classification, we randomly choose a single focal brand (Walmart) and gather all users who have made negative comments on this brand during the first three months *after* January 1, 2014, thus forming a negative user set of $95,594$ users. We then repeat this process and select the same number of users who made non-negative comments about the focal brand. After combining these user sets, we have $n = 191,188$ users who have interacted with a total of $18,736$ pages.

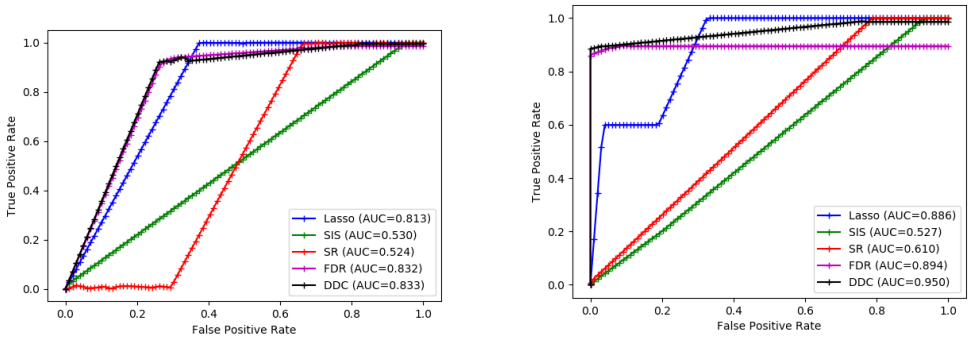
Each user is now represented by a single data point, where the binary response indicates whether they have made a negative comment specifically about the focal brand within the first three months of 2014, whereas the binary features indicate whether they have interacted with certain pages or categories of pages *before* 2014. Our goal is to use the observed past interactions, across all pages, to predict the incidence of future negative comments for the focal brand only. The binary predictors record the incidence of any type of commenting interaction (positive, negative, neutral). Thus, our dataset has a sample size of $n = 191,188$ with $p = 18,878$ features. To demonstrate how well our method scales with problem size, we also constructed a similar, smaller dataset with $n = 38,238$ users and, correspondingly, $p = 9,510$ features; note that, in this application, p increases with n since more users will interact with more brands.

Both datasets are very noisy, with many features appearing infrequently and a low proportion of data points with a response of 1. All of these factors make prediction quite challenging. In such problems, model selection has great practical value: even though $n > p$, the high level of noise creates the risk of spurious correlation, noise accumulation, and other known practical issues (Fan, Han and Liu, 2014), which can be mitigated by using a sparse model. As we demonstrate below, very sparse models can achieve very strong out-of-sample predictive power in this setting. Model sparsity also improves interpretability, since managers now only have to consider a small set of key features; DDC is particularly useful in this regard since it distinguishes between pages that can safely be represented at the aggregate level (e.g., by topic), as opposed to brands that require individual attention.

Furthermore, by reducing the number of features, model selection also reduces the computational complexity of estimating a regression model on the data. A screening approach is especially helpful in this setting, since we work with the marginal DC of each feature rather than the entire design matrix. Computational cost is important because, in practice, we may wish to repeat this analysis for many different brands (i.e., different response variables), which requires both rerunning and retuning the model. Moreover, the specific instance we consider here uses only a small portion of the data that are available to Facebook, and more computationally efficient methods would enable us to increase the problem size correspondingly.

Since the true sparse feature set $\mathcal{A}$ is unknown in this problem, we evaluate DDC and other methods according to their predictive power. We first conduct a screening step using the method of choice (DDC, Lasso, SR, SIS, or FDR). We then fit a new logistic regression model to the selection set returned by that method. This estimation step is required for DDC, SR, SIS and FDR as none of them performs estimation directly. As for Lasso, although it does perform estimation together with screening, this estimation is known to be biased, and Belloni and Chernozhukov (2013) has shown that the bias can be reduced by fitting a new model to the selected features. Thus, we perform this separate estimation step on the output of every method, so that prediction is performed using the same model class in all cases, and the only difference is in the feature set provided to the model.

As before, we run each method across a range of parameter settings. For each setting, we also consider a range of threshold values for the post-selection logistic regression model: that is, we predict $Y = 1$ if the estimated probability of this event is above the threshold. We then evaluate each combination of threshold and parameter setting using 10-fold cross-validation,



(a) Small data. (b) Large data.

FIG 5. *ROC curves for real datasets.*

and create ROC curves (Smithson and Merkle, 2013) from the results. Figure 5 shows these curves and reports the AUC, or area under the curve, for each method; this metric, which always takes values between 0.5 and 1, is widely used in practice when the data and response are binary and the proportion of positive responses is low. Table 2 reports computation times for both selection and estimation.

Again, one should compare Figures 5(a)-5(b) in order to understand how well the methods scale. On the small dataset, DDC, FDR and Lasso all achieve virtually the same performance. When we move to the large dataset, Lasso and FDR continue to perform very similarly, but DDC has now pulled ahead. DDC is consistently the fastest method: on the large dataset, it runs 7.5 times faster than FDR (combining both steps), 7.8 times faster than Lasso, over 1100 times faster than SIS, and over 2700 times faster than SR. The top three methods (DDC, Lasso and FDR) are all able to produce very sparse models, accepting under 1% of the available features. SIS can also be made to achieve this level of sparsity, since its tunable parameter directly controls the proportion of features to accept, but the accepted features appear to include many false positives (as was suggested in Proposition 5.5) since the performance of SIS on the large dataset is poor. Finally, the solutions produced by SR are less sparse, but still accept fewer than 3% of features.

Based on these results, we conclude that DDC offers significant practical potential in applications where the data follow a hierarchical structure, and both n and p are sufficiently large to merit the use of model selection to reduce the feature space, improve estimation speed, and increase predictive power. We note that the benefits of DDC are greater, relative to the benchmark methods, when the dataset is larger.

7. Conclusion. We have developed a new algorithm for model selection and screening in problems where the data are binary and structured hierarchically, which occur in many

Method	Small data		Large data	
	Selection	Estimation	Selection	Estimation
DDC	0.1332	0.0322	0.7215	0.1812
Lasso	0.1542	0.3422	0.8812	6.2348
SR	997.7151	0.3480	1340.3724	1177.1811
SIS	2.7791	80.6958	26.6560	983.6413
FDR	1.2697	0.0790	6.2700	0.6468

TABLE 2

Average computation times (in seconds) of all methods on real data.

business and marketing applications. An attractive feature of our approach is that it explores the hierarchy from top to bottom and screens features in a dynamic manner; as a result, lower-level features may not need to be examined at all if they have already been screened out at higher levels, and the computational cost is substantially reduced. The practical potential of the approach was demonstrated on both simulated and real data.

We note that our computational study considered two different types of settings. Our simulated data belong to the high-dimensional setting where $p \gg n$. However, we also give a case application in which $p < n$, but both n and p are fairly large. We emphasize that, even though this setting is not "high-dimensional" as that term is usually understood in the theoretical literature, nonetheless it is a setting where screening offers great practical value: first, it reduces the computational cost of estimating a predictive model, which can be prohibitive when both n and p are large, and second, it improves the predictive power of that model. Model selection is also very useful to managers as it leads to more interpretable results; in the context of hierarchical data, it allows decision-makers to better understand the degree of granularity needed for the aggregation structure in order to capture the statistical significance of a class of products or a customer segment. Thus, the application studied in our paper adds an important dimension to the practical study of the algorithm.

APPENDIX A: TECHNICAL PROOFS

In this section, we give the full proofs of all results that were stated in the text.

A.1. Proof of Proposition 4.1. For any two binary variables X, Y, where it is allowed that $X = Y$ as a special case, we first prove

$$\phi_{XY}(s,t) - \phi_X(s)\,\phi_Y(t) = \left(e^{is} - 1\right)\left(e^{it} - 1\right)\operatorname{cov}(X,Y).$$

For the left hand side, we have

$$\phi_{XY}(s,t) - \phi_X(s)\,\phi_Y(t) = \mathbb{E}\left(e^{isX}e^{itY}\right) - \mathbb{E}\left(e^{isX}\right)\mathbb{E}\left(e^{itY}\right) = \operatorname{cov}(e^{isX}, e^{itY}).$$

Note that $\mathbb{E}e^{isX} = e^{is}\mathbb{P}(X=1) + \mathbb{P}(X=0)$ (and similarly for Y). Then, with some algebra it can be shown that

$$\begin{aligned}
&\phi_{XY}(s,t) - \phi_X(s)\phi_Y(t) \\
&= \mathbb{E}\left[(e^{isX} - \mathbb{E}e^{isX})(e^{itY} - \mathbb{E}e^{itY})\right] \\
&= (e^{is}-1)\mathbb{P}(X=0)(e^{it}-1)\mathbb{P}(Y=0)\mathbb{P}(X=1,Y=1) \\
&\quad -(e^{is}-1)\mathbb{P}(X=0)(e^{it}-1)\mathbb{P}(Y=1)\mathbb{P}(X=1,Y=0) \\
&\quad -(e^{is}-1)\mathbb{P}(X=1)(e^{it}-1)\mathbb{P}(Y=0)\mathbb{P}(X=0,Y=1) \\
&\quad +(e^{is}-1)\mathbb{P}(X=1)(e^{it}-1)\mathbb{P}(Y=1)\mathbb{P}(X=0,Y=0) \\
&= (e^{is}-1)(e^{it}-1)\mathbb{P}(X=0)\mathbb{P}(Y=0)\mathbb{P}(X=1,Y=1) \\
&\quad -(e^{is}-1)(e^{it}-1)\mathbb{P}(X=0)\mathbb{P}(Y=1)\mathbb{P}(X=1,Y=0) \\
&\quad -(e^{is}-1)(e^{it}-1)\mathbb{P}(X=1)\mathbb{P}(Y=0)\mathbb{P}(X=0,Y=1) \\
&\quad +(e^{is}-1)(e^{it}-1)\mathbb{P}(X=1)\mathbb{P}(Y=1)\mathbb{P}(X=0,Y=0).
\end{aligned}$$

The first and third terms after the last equality above can be combined and simplified as

$$B = (e^{is}-1)(e^{it}-1)\mathbb{P}(Y=0)(\mathbb{P}(X=1,Y=1) - \mathbb{P}(X=1)\mathbb{P}(Y=1))$$

The second and fourth terms can likewise be simplified as

$$C = (e^{is}-1)(e^{it}-1)\mathbb{P}(Y=1)(\mathbb{P}(X=1,Y=1) - \mathbb{P}(X=1)\mathbb{P}(Y=1)).$$

Combining these together yields

$$B + C$$

$$= (e^{is}-1)(e^{it}-1)(\mathbb{P}(X=1,Y=1)-\mathbb{P}(X=1)\mathbb{P}(Y=1))(\mathbb{P}(Y=0)+\mathbb{P}(Y=1))$$
$$= (e^{is}-1)(e^{it}-1)(\mathbb{P}(X=1,Y=1)-\mathbb{P}(X=1)\mathbb{P}(Y=1))$$
$$= (e^{is}-1)(e^{it}-1)\mathrm{cov}(X,Y).$$

Recalling the definition of $\mathrm{dcov}(X,Y)$, we write

$$\mathrm{dcov}^2(X,Y) = \int_{\mathbb{R}^2} \|\phi_{XY}(s,t) - \phi_X(s)\phi_Y(t)\|^2 w(s,t)ds\,dt,$$

where $w(s,t) = \left(\pi^2 s^2 t^2\right)^{-1}$. We simplify this as

$$\begin{aligned}\mathrm{dcov}^2(X,Y) &= \int_{\mathbb{R}^2} (e^{is}-1)(e^{-is}-1)(e^{it}-1)(e^{-it}-1)\mathrm{cov}^2(X,Y)w(s,t)ds\,dt \\ &= A\cdot \mathrm{cov}^2(X,Y),\end{aligned}$$

where

$$\begin{aligned}A &= \int_{\mathbb{R}^2} \|(e^{is}-1)(e^{it}-1)\|^2 w(s,t)ds\,dt \\ &= \int_{\mathbb{R}^2} (2-2\cos s)\,(2-2\cos t)\,w(s,t)ds\,dt \\ &= 4. \qquad (9)\end{aligned}$$

Thus,

$$\mathrm{dcov}(X,Y) = 2|\mathrm{cov}(X,Y)|, \qquad \mathrm{dcov}(X,X) = 2\mathrm{cov}(X,X) = 2\mathrm{var}(X),$$

whence

$$\mathrm{dcorr}(X,Y) = \frac{2|\mathrm{cov}(X,Y)|}{2\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = |\,\mathrm{corr}(X,Y)|,$$

which completes the proof of statement (i) in Proposition 4.1.

We now prove statement (ii). First, we state a technical result proved in Székely, Rizzo and Bakirov (2007) that will be useful later.

LEMMA A.1. *The estimator* $\widehat{\mathrm{dcov}}\,(X,Y)$ *satisfies*

$$\widehat{dcov}^{\,2}(X,Y) = \int_{\mathbb{R}^2} \|f^n_{X,Y}(s,t) - f^n_X(s)f^n_Y(t)\|^2 w(t,s)ds\,dt,$$

where

$$f^n_{X,Y}(s,t) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle s,x_k\rangle + i\langle t,y_k\rangle\}$$

is the empirical characteristic function of the sample $(x_1,y_1),\ldots(x_n,y_n)$*, and*

$$f^n_X(s) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle s,x_k\rangle\}, \qquad f^n_Y(t) = \frac{1}{n}\sum_{k=1}^{n} \exp\{i\langle t,y_k\rangle\}.$$

Next, we prove the following technical lemma, which simplifies the computation for binary data.

LEMMA A.2. *Let $\bar{x}$ and $\bar{y}$ denote the sample averages of the binary vectors $(x_1, ..., x_n)$ and $(y_1, ..., y_n)$. The empirical characteristic function satisfies*

$$f^n_{X,Y}(s,t) - f^n_X(s) f^n_Y(t) = \frac{1}{n}\left(\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}\right)(e^{is}-1)(e^{it}-1).$$

Proof: We rewrite $f^n_{X,Y}(s,t)$, $f^n_X(s)$, and $f^n_Y(t)$ specifically for the binary case. In the following, let $\#(E)$ be the number of data points (x_k, y_k) in the sample that satisfy a condition E. For example, $\#(x_k = 1)$ is the number of such data points satisfying $x_k = 1$.

We write

$$\begin{aligned} f^n_{X,Y}(s,t) &= \frac{1}{n}\sum_{k=1}^{n} \exp(isx_k + ity_k) \\ &= \frac{1}{n}[e^{i(s+t)}\#(x_k=1, y_k=1) + e^{is}\#(x_k=1, y_k=0) \\ &\quad + e^{it}\#(x_k=0, y_k=1) + \#(x_k=0, y_k=0)] \\ &= \frac{1}{n}\left[(e^{i(s+t)} - e^{is} - e^{it})\#(x_k=1, y_k=1) + e^{is}\#(x_k=1) + e^{it}\#(y_k=1) + \#(x_k=0, y_k=0)\right]. \end{aligned}$$

The last line is obtained by adding and subtracting $e^{is}\#(x_k = 1, y_k = 1)$ and $e^{it}\#(x_k = 1, y_k = 1)$. In addition,

$$\begin{aligned} f^n_X(s) &= \frac{1}{n}\sum_{k=1}^{n} e^{isx_k} = \frac{1}{n}(e^{is}\#(x_k=1) + \#(x_k=0)) \\ &= \frac{1}{n}[(e^{is}-1)\#(x_k=1) + n] = 1 + \bar{x}(e^{is}-1), \end{aligned}$$

where the second line can be obtained by adding and subtracting $\#(x_k = 1)$. Similarly, we have $f^n_Y(t) = 1 + \bar{y}(e^{it} - 1)$. Then,

$$\begin{aligned} f^n_X(s) f^n_Y(t) &= (1 + \bar{x}(e^{is}-1))(1 + \bar{y}(e^{it}-1)) \\ &= 1 + \bar{x}(e^{is}-1) + \bar{y}(e^{it}-1) + \bar{x}\bar{y}(e^{is}-1)(e^{it}-1). \end{aligned}$$

Consequently,

$$\begin{aligned} & f^n_{X,Y}(s,t) - f^n_X(s) f^n_Y(t) \\ &= \frac{1}{n}\left(\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}\right)(e^{is}-1)(e^{it}-1) \\ &\quad + \bar{x} + \bar{y} - 1 + \frac{1}{n}(\#(x_k=0, y_k=0) - \#(x_k=1, y_k=1)) \\ &= \frac{1}{n}\left(\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}\right)(e^{is}-1)(e^{it}-1) + \frac{\#(x_k=1) + \#(x_k=0)}{n} - 1 \\ &= \frac{1}{n}\left(\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}\right)(e^{is}-1)(e^{it}-1), \end{aligned}$$

which completes the proof. □

Combining Lemmas A.1 and A.2, we have

$$\widehat{\mathrm{dcov}}^2(X,Y) = \frac{1}{n^2}\left(\sum_{k=1}^{n} x_k y_k - n\bar{x}\bar{y}\right)^2 \cdot A,$$

where A is as in (9). The desired result follows.

A.2. Proof of Theorem 5.3. Using results from moderate deviation theory for self-normalized sums, we first prove an intermediate result bounding the distance between the estimated and population DC. To begin, we define

$$b_n = \sqrt{\frac{2\log(p \vee n)}{n}}$$

and

$$\widehat{\mathrm{var}}(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$\widehat{\mathrm{var}}(X_j) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2,$$

$$\widetilde{\mathrm{var}}(X_j) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \mathbb{E}X_{ij})^2.$$

Note that this definition of $\widehat{\mathrm{var}}$ is slightly different from the usual sample variance (which we used in Proposition 4.1), in that we divide the sum by n instead of $n-1$. Since our analysis focuses on the asymptotic regime where $n \to \infty$, this is not a major issue. The benefit is that, under this definition, we have $\widehat{\mathrm{dcov}}(X,X) = 2\,|\widehat{\mathrm{var}}(X)|$, which simplifies some of the computations in the proofs that follow.

The following lemma establishes several technical results that are useful for the proof.

LEMMA A.3. *The following statements are true:*

(i) *Suppose that Z_{ij} is defined to be any one of the quantities in the set*

$$\{X_{ij}Y_i - \mathbb{E}X_{ij}Y_i, X_{ij} - \mathbb{E}X_{ij}, (X_{ij} - \mathbb{E}X_{ij})^2 - \mathbb{E}(X_{ij} - \mathbb{E}X_{ij})^2\}.$$

Also define $V_{nn,j}^2 = \sum_{i=1}^n Z_{ij}^2$ and $S_{nn,j} = \sum_{i=1}^n Z_{ij}$. Then, under Assumption 5.1,

$$\mathbb{P}\left(\max_{j\le p}\frac{\frac{1}{n}|S_{nn,j}|}{(\frac{1}{n}V_{nn,j}^2)^{1/2}} \le b_n\right) \to 1, \tag{10}$$

$$\mathbb{P}\left(\max_{j\le p}\frac{|\frac{1}{n}\sum_i X_{ij} - \mathbb{E}X_j|}{\widetilde{\mathrm{var}}(X_j)^{1/2}} \le b_n\right) \to 1. \tag{11}$$

(ii) *For any $d_0 > 0$,*

$$\mathbb{P}\left(\frac{|\frac{1}{n}\sum_i Y_i - \mathbb{E}Y|}{\mathrm{var}(Y)^{1/2}} \le d_0 b_n\right) \to 1, \tag{12}$$

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i Y_i - \mathbb{E}Y\right| \le 0.5 d_0 b_n\right) \to 1. \tag{13}$$

(iii) *For any $d_0 > 0$,*

$$\mathbb{P}\left(\max_{j\le p}\left|\frac{1}{n}\sum_i X_{ij}Y_i - \mathbb{E}X_{ij}Y_i\right| \le b_n(0.5 + d_0)\right) \to 1, \tag{14}$$

$$\mathbb{P}\left(\max_{j\le p}\left|\frac{1}{n}\sum_i X_{ij} - \mathbb{E}X_{ij}\right| \le b_n(0.5 + d_0)\right) \to 1. \tag{15}$$

(iv) *Under Assumption 5.2(i),*

$$\left|\frac{\widehat{\mathrm{dcov}}(Y,Y)}{\mathrm{dcov}(Y,Y)}-1\right|+\max_{j\le p}\left|\frac{\widehat{\mathrm{dcov}}(X_j,X_j)}{\mathrm{dcov}(X_j,X_j)}-1\right|+\max_{j\le p}\left|\frac{\widehat{\mathrm{var}}(X_j)}{\mathrm{var}(X_j)}-1\right|+\max_{j\le p}\left|\frac{\widetilde{\mathrm{var}}(X_j)}{\mathrm{var}(X_j)}-1\right|=o_P(1).$$

(v) *For any $d_0>0$,*

$$\mathbb{P}\left(\max_{j\le p}\left|\widehat{\mathrm{var}}(X_j)^{1/2}-\mathrm{var}(X_j)^{1/2}\right|\le(0.5+d_0)b_n\right)\to 1.$$

$$\mathbb{P}\left(\frac{\left|\widehat{\mathrm{dcov}}(X_j,X_j)^{1/2}-\mathrm{dcov}(X_j,X_j)^{1/2}\right|}{\widehat{\mathrm{dcov}}(X_j,X_j)^{1/2}}\le\frac{b_n\,(1+2d_0)\,(1+d_0)}{\widehat{\mathrm{dcov}}\,(X_j,X_j)^{1/2}},\quad j=1,...,p\right)\to 1$$

$$\mathbb{P}\left(\frac{\left|\widehat{\mathrm{dcov}}(Y,Y)^{1/2}-\mathrm{dcov}(Y,Y)^{1/2}\right|}{\widehat{\mathrm{dcov}}(Y,Y)^{1/2}}\le\frac{b_n\,(1+2d_0)\,(1+d_0)}{\widehat{\mathrm{dcov}}\,(Y,Y)^{1/2}},\quad j=1,...,p\right)\to 1.$$

Proof: (i) Observe that $\mathbb{E}Z_{ij}=0$ and the random variables Z_{ij} are independent across $i\le n$. Applying Lemma 5 of Belloni et al. (2012), there exists a sequence $l_n\to\infty$, and a constant $C>0$, such that for any $0<x<C\frac{n^{1/6}}{l_n}-1$, we have

$$\left|\frac{\max_{j\le p}\mathbb{P}(|S_{nn,j}/V_{nn,j}|>x)}{2(1-\Phi(x))}-1\right|\to 0,$$

where Φ denotes the standard normal cdf. Now, choose $x=\Phi^{-1}(1-\gamma_n/(2p))$ and define $\gamma_n=\left(\sqrt{\pi\log(p\vee n)}\right)^{-1}$. Then, $2p(1-\Phi(x))=\gamma_n$ and $\gamma_n=o(1)$ by construction, whence

$$\begin{aligned}\mathbb{P}\left(\max_{j\le p}\frac{\frac{1}{n}|S_{nn,j}|}{(\frac{1}{n}V_{nn,j}^2)^{1/2}}>\frac{x}{n^{1/2}}\right)&=\mathbb{P}\left(\max_{j\le p}\left|\frac{S_{nn,j}}{V_{nn,j}}\right|>x\right)\\&\le p\max_{j\le p}\mathbb{P}\left(\left|\frac{S_{nn,j}}{V_{nn,j}}\right|>x\right)\\&\le 2p(1-\Phi(x))(1+o(1))\\&=\gamma_n(1+o(1)).\end{aligned}$$

To complete the proof of (10), we now show that $x\le\sqrt{2\log p\vee n}$, which is equivalent to the inequality

$$\mathbb{P}\left(N(0,1)>\sqrt{2\log(p\vee n)}\right)\le\frac{\gamma_n}{2p}.$$

This is achieved by applying the Mill's ratio inequality (Ruben, 1962) as follows:

$$\mathbb{P}\left(N(0,1)>\sqrt{2\log(p\vee n)}\right)\le\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sqrt{2\log(p\vee n)}}e^{-\log(p\vee n)}\le\frac{1}{2p\sqrt{\pi}}\cdot\frac{1}{\sqrt{\log(p\vee n)}}.$$

Combining this with the assumption that $\log p=o\left(n^{\frac{1}{4}}\right)$, we take $l_n=n^{\frac{1}{24}}$ and obtain

$$x\le\sqrt{2\log p\vee n}<C\frac{n^{\frac{1}{6}}}{l_n}-1$$

for all $C>0$. Thus, the first statement (10) is proved.

The second statement (11) is a direct implication of the preceding by setting $Z_{ij} = X_{ij} - \mathbb{E}X_{ij}$. Thus, part (i) is proved.

(ii) Define $Z_i = \frac{Y_i - \mathbb{E}Y}{\sqrt{\mathrm{var}(Y)}}$. Then, we have $|\frac{1}{n}\sum_i Z_i| = O_P(n^{-1/2})$. It follows that $|\frac{1}{n}\sum_i Z_i| = o_P(b_n)$, implying (12) for any $d_0 > 0$. On the event where (12) holds, it follows that

$$\left|\frac{1}{n}\sum_i Y_i - \mathbb{E}Y\right| \le d_0 b_n \sqrt{\mathrm{var}(Y)} \le 0.5 d_0 b_n,$$

completing the proof of (13).

(iii) Let $Z_{ij} = X_{ij}Y_i - \mathbb{E}X_{ij}Y_i$. On the event

$$E_1 = \left\{ \max_{j\le p} \frac{\frac{1}{n}|S_{nn,j}|}{(\frac{1}{n}V^2_{nn,j})^{1/2}} \le b_n \right\},$$

we have

$$\max_{j\le p}\left|\frac{1}{n}\sum_i X_{ij}Y_i - \mathbb{E}\left(X_{ij}Y_i\right)\right| \le b_n \max_j \left(\frac{1}{n}\sum_{i=1}^n \left(X_{ij}Y_i - \mathbb{E}\left(X_{ij}Y_i\right)\right)^2\right)^{1/2}. \tag{16}$$

A crude but simple bound for the right-hand side of (16) is

$$\max_j \left(\frac{1}{n}\sum_{i=1}^n \left(X_{ij}Y_i - \mathbb{E}\left(X_{ij}Y_i\right)\right)^2\right)^{1/2} \le 2,$$

which implies that

$$\max_j \left|\frac{1}{n}\sum_{i=1}^n X_{ij}Y_i - \mathbb{E}\left(X_{ij}Y_i\right)\right| = o_P\left(1\right).$$

Thus, uniformly in $j \le p$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left(X_{ij}Y_i - \mathbb{E}X_{ij}Y_i\right)^2 &= \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i + \left(\mathbb{E}\left(X_{ij}Y_i\right)\right)^2 - \frac{2}{n}\sum_{i=1}^n X_{ij}Y_i\mathbb{E}\left(X_{ij}Y_i\right)\\ &\le \mathbb{E}\left(X_{ij}Y_i\right) - \left(\mathbb{E}\left(X_{ij}Y_i\right)\right)^2 + o_P(1)\\ &\le 0.25 + o_P(1)\end{aligned}$$

Consequently, for any $d_0 > 0$, (14) holds with probability approaching 1. Equation (15) follows from similar arguments, which are omitted.

(iv) Let $Z_{ij} = (X_{ij} - \mathbb{E}X_{ij})^2 - \mathbb{E}(X_{ij} - \mathbb{E}X_{ij})^2$. Applying part (i) proved above, on the event

$$E_2 = \left\{ \max_j \left| \frac{\frac{1}{n}\sum_i Z_{ij}}{\left(\frac{1}{n}\sum_i Z_{ij}^2\right)^{1/2}} \right| \le b_n \right\},$$

we calculate

$$\max_{j\le p}\left|\frac{\widetilde{\mathrm{var}}(X_j) - \mathrm{var}(X_j)}{\mathrm{var}(X_j)}\right| \le \left|\frac{\max_{j\le p}\frac{1}{n}\sum_i Z_{ij}}{\min_j \mathrm{var}(X_j)}\right| \le b_n \frac{\max_j \left(\frac{1}{n}\sum_i Z_{ij}^2\right)^{1/2}}{\min_j \mathrm{var}(X_j)} = o_P(1),$$

where the last equality follows from the boundedness of $\max_j |Z_{ij}|$ and from Assumption 5.2(i).

Note that, under the definition of $\widehat{\text{var}}$ used in this section, we have

$$\widehat{\text{var}}\,(X_j) = \frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2 = \widetilde{\text{var}}(X_j) - \left(\bar{X}_j - \mathbb{E}X_{ij}\right)^2.$$

Thus, for arbitrarily small d_0, we have

$$|\widehat{\text{var}}\,(X_j) - \text{var}\,(X_j)| \leq |\widetilde{\text{var}}\,(X_j) - \text{var}\,(X_j)| + b_n^2\,(0.5 + d_0)^2. \tag{17}$$

For the first term on the right-hand side, we apply (i) with $Z_{ij} = (X_{ij} - \mathbb{E}X_{ij})^2 - \mathbb{E}\left((X_{ij} - \mathbb{E}X_{ij})^2\right)$, and obtain, with probability approaching 1, the inequality,

$$|\widetilde{\text{var}}\,(X_j) - \text{var}\,(X_j)| \leq b_n\left(\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2\right)^{\frac{1}{2}}.$$

We now bound $\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2$. By the same argument as in (i), noting that $X_{ij}^2 = X_{ij}$ and therefore

$$\text{var}\,(X_j)\,(1 - 2\mathbb{E}X_{ij})^2 = \text{var}\left((X_j - \mathbb{E}X_{ij})^2\right)$$

for binary features, we obtain

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} Z_{ij}^2 &= O_P\,(b_n) + \mathbb{E}\left(Z_{ij}^2\right)\\ &= O_P\,(b_n) + \text{var}\,(X_j)\,(1 - 2\mathbb{E}X_{ij})^2\\ &\leq O_P\,(b_n) + \text{var}\,(X_{ij})\\ &< (1 + \varepsilon_n)^2\,\text{var}\,(X_j),\end{aligned}$$

where the last inequality holds, uniformly in $j \leq p$ and with probability approaching 1, for some $\varepsilon_n \to 0$, due to the fact that $b_n \ll \min_j \text{var}\,(X_j)$. Consequently, returning to (17), there exists some $\varepsilon_n' \to 0$ satisfying

$$\begin{aligned}|\widehat{\text{var}}\,(X_j) - \text{var}\,(X_j)| &\leq b_n\,(1 + \varepsilon_n)\sqrt{\text{var}\,(X_j)} + b_n^2\,(0.5 + d_0)^2\\ &\leq b_n\left(1 + \varepsilon_n'\right)\sqrt{\text{var}\,(X_j)}.\end{aligned}$$

This also implies

$$\max_{j\leq p}\left|\frac{\widehat{\text{var}}\,(X_j) - \text{var}\,(X_j)}{\text{var}\,(X_j)}\right| \leq \frac{b_n\,(1 + \varepsilon_n')}{\sqrt{\text{var}\,(X_j)}} = o_P\,(1),$$

and

$$\max_{j\leq p}\sum_{i=1}^{n}\left(X_{ij} - \bar{X}_j\right)^2 \leq \max_{j\leq p}\text{var}\,(X_j) + o_P\,(1) = O_P\,(1).$$

The same arguments also yield $\left|\frac{\widehat{\text{var}}(Y)}{\text{var}(Y)} - 1\right| = o_P\,(1)$, whence

$$\left|\frac{\widehat{\text{dcov}}\,(Y,Y)}{\text{dcov}\,(Y,Y)} - 1\right| = \left|\frac{\widehat{\text{var}}\,(Y)}{\text{var}\,(Y)} - 1\right| = o_P\,(1),$$

as well as

$$\max_{j\leq p}\left|\frac{\widehat{\operatorname{dcov}}\left(X_j,X_j\right)-\operatorname{dcov}\left(X_j,X_j\right)}{\operatorname{dcov}\left(X_j,X_j\right)}\right|=\max_{j\leq p}\left|\frac{\widehat{\operatorname{var}}\left(X_j\right)}{\operatorname{var}\left(X_j\right)}-1\right|=o_P\left(1\right).$$

(v) First, we note that the result of part (iv) implies

$$\frac{\operatorname{var}\left(X_j\right)^{1/2}}{\widehat{\operatorname{var}}\left(X_j\right)^{1/2}+\operatorname{var}\left(X_j\right)^{1/2}}\leq\frac{1}{2}\left(1+o_P\left(1\right)\right). \tag{18}$$

We then observe that

$$\left|\widehat{\operatorname{var}}\left(X_j\right)^{1/2}-\operatorname{var}\left(X_j\right)^{1/2}\right|\leq\frac{\left|\widehat{\operatorname{var}}\left(X_j\right)-\operatorname{var}\left(X_j\right)\right|}{\widehat{\operatorname{var}}\left(X_j\right)^{1/2}+\operatorname{var}\left(X_j\right)^{1/2}}. \tag{19}$$

Applying (17) to the right-hand side of (19), we obtain

$$\begin{aligned}\left|\widehat{\operatorname{var}}\left(X_j\right)^{1/2}-\operatorname{var}\left(X_j\right)^{1/2}\right|&\leq\frac{b_n\left(1+\varepsilon_n'\right)\operatorname{var}\left(X_j\right)^{1/2}}{\widehat{\operatorname{var}}\left(X_j\right)^{1/2}+\operatorname{var}\left(X_j\right)^{1/2}}\\&\leq\frac{1}{2}b_n\left(1+\varepsilon_n'\right)\left(1+\varepsilon_n''\right),\end{aligned}$$

where the last inequality follows from (18). Now note that, as defined in this section, $\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}=\sqrt{2}\widehat{\operatorname{var}}\left(X_j\right)^{1/2}$, which leads to

$$\begin{aligned}\frac{\left|\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}-\operatorname{dcov}\left(X_j,X_j\right)^{1/2}\right|}{\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}}&\leq\sqrt{2}\frac{\left|\widehat{\operatorname{var}}\left(X_j\right)^{1/2}-\operatorname{var}\left(X_j\right)^{1/2}\right|}{\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}}\\&\leq\frac{\sqrt{2}}{2}\frac{b_n\left(1+\varepsilon_n'\right)\left(1+\varepsilon_n''\right)}{\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}}\\&\leq\frac{b_n\left(1+2d_0\right)\left(1+d_0\right)}{\widehat{\operatorname{dcov}}\left(X_j,X_j\right)^{1/2}},\end{aligned}$$

where the last line follows because, for fixed $d_0>0$, we have $\frac{1}{\sqrt{2}}\left(1+\varepsilon_n'\right)\left(1+\varepsilon_n''\right)\leq\left(1+2d_0\right)\left(1+d_0\right)$ for large enough n. The same argument also implies

$$\frac{\left|\widehat{\operatorname{dcov}}\left(Y,Y\right)^{1/2}-\operatorname{dcov}\left(Y,Y\right)^{1/2}\right|}{\widehat{\operatorname{dcov}}\left(Y,Y\right)^{1/2}}\leq\frac{b_n\left(1+2d_0\right)\left(1+d_0\right)}{\widehat{\operatorname{dcov}}\left(Y,Y\right)^{1/2}},$$

as required.

We are now able to consider the distance between the estimated and population DC. The following theorem presents a bound on this distance that holds w.p. 1 asymptotically.

THEOREM A.4. *For any $d_0>0$, under Assumptions 5.1 and 5.2, we have*

$$\mathbb{P}\left(\max_{j\leq p}\left|\widehat{\operatorname{dcov}}(X_j,Y)-\operatorname{dcov}(X_j,Y)\right|>b_n(2+d_0)\right)\to 0$$

as $n\to\infty$.

Proof: For any $c_0 > 0$, define three events

$$E_1 = \left\{ \max_j |\mathbb{E}X_j Y - \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i| \leq b_n(0.5 + c_0) \right\},$$
$$E_2 = \left\{ \max_j |\mathbb{E}X_j - \bar{X}_j| \leq b_n(0.5 + c_0) \right\},$$
$$E_3 = \left\{ |\mathbb{E}Y - \bar{Y}| \leq b_n\left(0.5 + c_0\right) \right\}.$$

By Lemma A.3, all three events jointly hold with probability approaching 1.

On events E_2 and E_3, we have

$$\left|\mathbb{E}X_j \mathbb{E}Y - \bar{X}_j\bar{Y}\right| \leq \left|\mathbb{E}X_j(\mathbb{E}Y - \bar{Y})\right| + \left|\mathbb{E}X_j - \bar{X}\right| \cdot \bar{Y} \leq b_n(0.5 + 1.5c_0)$$

uniformly in $j \leq p$. On the event $E_1 \cap E_2 \cap E_3$ (that is, when all three events simultaneously hold), it follows from the triangle inequality that

$$\begin{aligned}
|\operatorname{dcov}(X_j, Y) - \widehat{\operatorname{dcov}}(X_j, Y)| &\leq 2\left|\operatorname{cov}\left(X_j, Y\right) - \widehat{\operatorname{cov}}\left(X_j, Y\right)\right| \\
&\leq 2\left|\mathbb{E}\left(X_j Y\right) - \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i\right| + 2\left|\mathbb{E}X_j \cdot \mathbb{E}Y - \bar{X}_j\bar{Y}\right| \\
&\leq 2b_n\left(0.5 + c_0\right) + 2b_n\left(0.5 + 1.5c_0\right) \\
&\leq b_n\left(2 + 3.5c_0\right).
\end{aligned}$$

uniformly in $j \leq p$. Thus, the event

$$E = \left\{ \max_j \left|\operatorname{dcov}(X_j, Y) - \widehat{\operatorname{dcov}}(X_j, Y)\right| < b_n(2 + 3.5c_0) \right\}$$

is implied by $E_1 \cap E_2 \cap E_3$, and thus holds with probability approaching one. The result holds since c_0 can be arbitrarily small, so we can take $c_0 = \frac{d_0}{3.5}$. □

Now, consider the threshold

$$K_n := \frac{b_n(3.5 + d_0)}{\min_j \widehat{\operatorname{dcov}}\left(X_j, X_j\right)^{1/2} \widehat{\operatorname{dcov}}\left(Y, Y\right)^{1/2}}.$$

The following theorem is the analog of Theorem A.4 for distance correlation (rather than covariance).

THEOREM A.5. *Under Assumptions 5.1 and 5.2, we have*

$$\mathbb{P}\left(\max_j \left|\widehat{\operatorname{dcorr}}(X_j, Y) - \operatorname{dcorr}(X_j, Y)\right| < K_n \right) \to 1,$$

where $\mathbb{P}$ represents the probability measure induced by the distribution of K_n. Additionally,

$$\mathbb{P}\left(\min_{j \in \mathcal{A}} \operatorname{dcorr}(X_j, Y) \geq 3K_n \right) \to 1$$

as $n \to \infty$.

Proof: We calculate

$$\left|\widehat{\operatorname{dcorr}}(X_j, Y) - \operatorname{dcorr}(X_j, Y)\right|$$

$$
\begin{aligned}
&= \left| \frac{\widehat{\mathrm{dcov}}(X_j, Y)}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} - \frac{\mathrm{dcov}(X_j, Y)}{\mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2}} \right| \\
&= \left| \frac{\widehat{\mathrm{dcov}}(X_j, Y)\,\mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2}}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}\,\mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2}} \right. \\
&\qquad \left. - \frac{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}\mathrm{dcov}(X_j, Y)}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}\,\mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2}} \right|. \qquad (20)
\end{aligned}
$$

Factoring out the common denominator in (20), the numerator is bounded by

$$
\begin{aligned}
&\widehat{\mathrm{dcov}}(X_j, Y)\,\mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2} - \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}\mathrm{dcov}(X_j, Y) \\
&\le \left|\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)\right| \mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2} \\
&\quad + \mathrm{dcov}(X_j, Y)\left|\mathrm{dcov}(X_j, X_j)^{1/2} - \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\right| \mathrm{dcov}(Y, Y)^{1/2} \\
&\quad + \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\mathrm{dcov}(X_j, Y)\left|\widehat{\mathrm{dcov}}(Y, Y)^{1/2} - \mathrm{dcov}(Y, Y)^{1/2}\right| \qquad (21) \\
&\le \left|\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)\right| \mathrm{dcov}(X_j, X_j)^{1/2}\mathrm{dcov}(Y, Y)^{1/2} \\
&\quad + (\mathrm{dcov}(X_j, X_j))^{1/2}(\mathrm{dcov}(Y, Y))^{1/2}\left|\mathrm{dcov}(X_j, X_j)^{1/2} - \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\right| \mathrm{dcov}(Y, Y)^{1/2} \\
&\quad + \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}(\mathrm{dcov}(X_j, X_j))^{1/2}(\mathrm{dcov}(Y, Y))^{1/2}\left|\widehat{\mathrm{dcov}}(Y, Y)^{1/2} - \mathrm{dcov}(Y, Y)^{1/2}\right|, \qquad (22)
\end{aligned}
$$

where (22) follows from (21) because $\mathrm{dcov}(X_j, Y) \le \sqrt{\mathrm{dcov}(X_j, X_j)} \cdot \sqrt{\mathrm{dcov}(Y, Y)}$. Hence,

$$
\left|\widehat{\mathrm{dcorr}}(X_j, Y) - \mathrm{dcorr}(X_j, Y)\right| \le A_1 + A_2 + A_3,
$$

where, for any $d_0 > 0$, we have

$$
A_1 = \frac{\left|\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)\right|}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \le \frac{b_n(2 + d_0)}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}}, \qquad (23)
$$

$$
\begin{aligned}
A_2 &= \frac{\left|\mathrm{dcov}(X_j, X_j)^{1/2} - \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\right| \mathrm{dcov}(Y, Y)^{1/2}}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \\
&\le \frac{\left|\mathrm{dcov}(X_j, X_j)^{1/2} - \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}\right|}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}}(1 + d_0) \qquad (24) \\
&\le \frac{b_n\,(1 + 2d_0)\,(1 + d_0)}{\widehat{\mathrm{dcov}}\,(X_j, X_j)^{1/2}}\,(1 + d_0) \qquad (25)
\end{aligned}
$$

$$
A_3 = \frac{\left|\widehat{\mathrm{dcov}}(Y, Y)^{1/2} - \mathrm{dcov}(Y, Y)^{1/2}\right|}{\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \le \frac{b_n\,(1 + 2d_0)\,(1 + d_0)}{\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \qquad (26)
$$

The inequality in (23) follows from Theorem A.4. Inequality (24) follows from part (iv) of Lemma A.3, whereas (25)-(26) follow from part (v) of Lemma A.3. Now, with probability

approaching 1, we have

$$\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \leq \mathrm{dcov}(X_j, X_j)^{1/2} + o_P(1) \leq \sqrt{2}\mathrm{var}(X_j)^{1/2} + o_P(1) \leq \frac{\sqrt{2}}{2} + d_0.$$

Similarly, $\widehat{\mathrm{dcov}}(Y, Y)^{1/2} \leq \frac{\sqrt{2}}{2} + d_0$. Recalling that $n^{-1/2} = o(b_n)$, we conclude that, for any $c_0 > 0$, we can take d_0 sufficiently small so that

$$A_2 \leq \frac{(1+d_0)^2 (1+2d_0) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}} \leq \frac{(1+d_0)^2 (1+2d_0) \left(\frac{\sqrt{2}}{2} + d_0\right) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \leq \frac{\left(\frac{\sqrt{2}}{2} + c_0\right) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}}$$

and

$$A_3 \leq \frac{(1+2d_0)(1+d_0) b_n}{\widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \leq \frac{(1+2d_0)(1+d_0) \left(\frac{\sqrt{2}}{2} + d_0\right) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \leq \frac{\left(\frac{\sqrt{2}}{2} + c_0\right) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}}.$$

Hence, we can write $c_0' = 3c_0$ and obtain

$$\begin{aligned}
\left|\widehat{\mathrm{dcorr}}(X_j, Y) - \mathrm{dcorr}(X_j, Y)\right| &\leq \frac{\left(2 + \sqrt{2} + 3c_0\right) b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \\
&\leq \frac{(3.5 + c_0') b_n}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \\
&\leq \frac{(3.5 + c_0') b_n}{\min_j \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \\
&= K_n
\end{aligned}$$

uniformly in $j \leq p$.

In addition, applying Lemma A.3 again, we obtain

$$\begin{aligned}
K_n &\leq \frac{b_n(3.5 + c_0')}{\min_j \widehat{\mathrm{dcov}}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \\
&\leq \frac{b_n(3.5 + c_0')}{\min_j \mathrm{dcov}(X_j, X_j)^{1/2} \widehat{\mathrm{dcov}}(Y, Y)^{1/2}} \max_j \left(\frac{\mathrm{dcov}(X_j, X_j)^{1/2}}{\widehat{\mathrm{dcov}}(X_j, X_j)^{1/2}} \right) \\
&\leq \frac{2b_n(3.5 + c_0')}{\min_j \mathrm{dcov}(X_j, X_j)^{1/2} \mathrm{dcov}(Y, Y)^{1/2}},
\end{aligned}$$

with probability approaching 1. On the other hand, there exists some constant $C > 0$ satisfying $\frac{1}{\sqrt{2}} \max_j \mathrm{dcov}\left(X_j, X_j\right)^{1/2} < \frac{1}{C}$. By Assumption 5.2, we also have

$$\min_{j \in \mathcal{A}} \mathrm{dcov}\left(X_j, Y\right) \cdot \min_{j \leq p} \mathrm{var}\left(X_j\right)^{1/2} \geq \frac{6}{C} \left(3.5 + c_0'\right) b_n,$$

which leads to

$$\begin{aligned}
\min_{j \in \mathcal{A}} \mathrm{dcorr}(X_j, Y) &\geq \frac{\min_{j \in \mathcal{A}} \mathrm{dcov}(X_j, Y)}{\mathrm{dcov}(Y, Y)^{1/2} \max_j \mathrm{dcov}(X_j, X_j)^{1/2}} \\
&\geq \frac{\frac{6}{C} \left(3.5 + c_0'\right) b_n}{\min_j \mathrm{dcov}\left(X_j, X_j\right)^{1/2} \mathrm{dcov}\left(Y, Y\right)^{1/2}} \frac{\sqrt{2}}{\max_j \mathrm{dcov}\left(X_j, X_j\right)^{1/2}} \\
&\geq \frac{6 \left(3.5 + c_0'\right) b_n}{\min_j \mathrm{dcov}\left(X_j, X_j\right)^{1/2} \mathrm{dcov}\left(Y, Y\right)^{1/2}} \\
&\geq 3K_n,
\end{aligned}$$

completing the proof. □

Proof of Theorem 5.3: We can now complete the proof of the main theorem. The hierarchical structure of $\widehat{\mathcal{A}}$ follows directly from the design of the algorithm. With regard to the main result, we separately prove $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ and $\mathcal{B} \subseteq \widehat{\mathcal{A}}$.

Note that, by Assumption 5.2, $\min_{j\in\mathcal{A}} \operatorname{dcorr}(X_j, Y) > 2K_n$. Hence, by Theorem A.5, with probability approaching 1, we have

$$\min_{j\in\mathcal{A}} \widehat{\operatorname{dcorr}}(X_j, Y) \geq \min_{j\in\mathcal{A}} \operatorname{dcorr}(X_j, Y) - K_n > K_n.$$

Consequently, Theorem A.5 implies

$$\mathbb{P}\left(\min_{j\in\mathcal{A}} \widehat{\operatorname{dcorr}}(X_j, Y) > K_n\right) \to 1. \tag{27}$$

Now, for any $j \in \mathcal{A}$, we aim to show $j \in \widehat{\mathcal{A}}$. If not, then there are two possibilities: either $\widehat{\operatorname{dcorr}}(X_j, Y) \leq K_n$, or there exists i such that $j \in \mathcal{D}(i)$ and $\widehat{\operatorname{dcorr}}(X_i, Y) \leq K_n$. The first case is ruled out due to (29). In the second case, we have, with probability approaching one,

$$\operatorname{dcorr}(X_i, Y) \leq \widehat{\operatorname{dcorr}}(X_i, Y) + K_n \leq 2K_n \leq Cb_n \tag{28}$$

for some constant $C > 0$. By the weak extinction property (Assumption 4.2), $\operatorname{dcorr}(X_j, Y) = O(b_n)$ since $j \in \mathcal{D}(i)$. However, by Assumption 5.2, we have $\min_{j\in\mathcal{A}} \operatorname{dcorr}(X_j, Y) \gg b_n$, whence $j \in \mathcal{A}^c$ and a contradiction is obtained. Thus, $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ asymptotically w.p. 1.

We now show $\mathcal{B} \subseteq \widehat{\mathcal{A}}$. For any $j \in \mathcal{B}$, we know that $j \notin \mathcal{A}$ but there is a descendant feature k of j such that $k \in \mathcal{A}$. Suppose that $j \notin \widehat{\mathcal{A}}$. Then, there are two possibilities: either $\widehat{\operatorname{dcorr}}(X_j, Y) \leq K_n$, or there exists i such that $j \in \mathcal{D}(i)$ and $\widehat{\operatorname{dcorr}}(X_i, Y) \leq K_n$; in the latter case, k is also a descendant of i. Thus, in either case, k is a descendant of some feature (either i or j) whose estimated distance correlation with Y is below K_n. By (28), the true distance correlation between this feature (either i or j) is bounded above by Cb_n. Then, by the weak extinction property, $\operatorname{dcorr}(X_k, Y) = O(b_n)$. However, Assumption 5.2 together with $k \in \mathcal{A}$ implies that $\operatorname{dcorr}(X_k, Y) \gg b_n$, and a contradiction is obtained.

For the final statement in Theorem 5.3, observe that, for any $j \in \widehat{\mathcal{A}}$, $\mathcal{C}(j)$ features will be added to the candidate set, and therefore $|\mathcal{C}(j)|$ calculations of empirical DC will be made in the next iteration. For the initial candidate set, the number of variables at the top level of the hierarchy is finite. Therefore, the total number of calculations of empirical DC is $O(\sum_{j\in\widehat{\mathcal{A}}} |\mathcal{C}(j)|)$.

Proof of Corollary 5.4: Combining the assumption of the corollary with Theorem A.5, we obtain

$$\mathbb{P}\left(\max_{j\notin\mathcal{A}} \widehat{\operatorname{dcorr}}(X_j, Y) \leq K_n\right) \to 1. \tag{29}$$

It can then be shown that $j \notin \mathcal{A}$ implies $j \notin \widehat{\mathcal{A}}$ using arguments similar to the previous proof.

Proof of Proposition 5.5: We first consider statement (i). First, from the definition of (6) it follows that $\widehat{\mathcal{A}} \subseteq \widehat{\mathcal{H}}$, leading to the inequality $\left|\widehat{\mathcal{A}} \cap \mathcal{A}^c\right| \leq \left|\widehat{\mathcal{H}} \cap \mathcal{A}^c\right|$.

By the definition of the DDC algorithm, for any $j \in \widehat{\mathcal{A}}$ we have $\widehat{\operatorname{dcorr}}(X_j, Y) \geq K_n$ as well as $\widehat{\operatorname{dcorr}}(X_i, Y) \geq K_n$ for any i such that $j \in \mathcal{D}(i)$. Thus, it follows that $j \neq \widehat{\mathcal{G}}$ and thus $\widehat{\mathcal{A}} \cap \widehat{\mathcal{G}} = \emptyset$. On the other hand, any $j \in \widehat{\mathcal{G}}$ satisfies $\widehat{\operatorname{dcorr}}(X_j, Y) \geq K_n$ and thus is an element of $\widehat{\mathcal{H}}$ by (6). Thus, $\widehat{\mathcal{G}} \subseteq \widehat{\mathcal{H}}$.

To show statement (ii), first recall that $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ in the asymptotic regime of Theorem 5.3, whence $\mathcal{A} \subseteq \widehat{\mathcal{H}}$ since $\widehat{\mathcal{A}} \subseteq \widehat{\mathcal{H}}$. Therefore,

$$\left|\widehat{\mathcal{H}} \cap \mathcal{A}^c\right| - \left|\widehat{\mathcal{A}} \cap \mathcal{A}^c\right| = \left|\widehat{\mathcal{H}}\right| - \left|\widehat{\mathcal{A}}\right|.$$

It follows that

$$\frac{\left|\widehat{\mathcal{A}} \setminus \mathcal{A}\right|}{\left|\widehat{\mathcal{A}}\right|} \leq \frac{\left|\widehat{\mathcal{H}} \setminus \mathcal{A}\right|}{\left|\widehat{\mathcal{H}}\right|} \tag{30}$$

because the right-hand side of (30) is obtained by adding the same constant to both the numerator and denominator of the left-hand side. Finally, statement (iii) follows directly from Corollary 5.4.

REFERENCES

BACH, F., JENATTON, R., MAIRAL, J. and OBOZINSKI, G. (2012). Structured sparsity through convex optimization. *Statistical Science* **27** 450–468.

BARUT, E., FAN, J. and VERHASSELT, A. (2016). Conditional sure independence screening. *Journal of the American Statistical Association* **111** 1266–1277.

BELLONI, A. and CHERNOZHUKOV, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.

BELLONI, A., CHEN, D., CHERNOZHUKOV, V. and HANSEN, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* **80** 2369–2429.

BERTSIMAS, D., O'HAIR, A., RELYEA, S. and SILBERHOLZ, J. (2016). An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer. *Management Science* **62** 1511–1531.

FAN, J. and FAN, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics* **36** 2605–2637.

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106** 544–557.

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *National Science Review* **1** 293–314.

FAN, J. and HAN, X. (2017). Estimation of the false discovery proportion with unknown dependence. *Journal of the Royal Statistical Society* **B79** 1143–1164.

FAN, J. and LV, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society* **B70** 849–911.

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research* **10** 2013–2038.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics* **38** 3567–3604.

FERREIRA, J. A. and ZWINDERMAN, A. H. (2006). On the Benjamini–Hochberg method. *The Annals of Statistics* **34** 1827–1849.

HAO, N. and ZHANG, H. H. (2014). Interaction Screening for Ultrahigh-Dimensional Data. *Journal of the American Statistical Association* **109** 1285–1301.

HAO, N. and ZHANG, H. H. (2016). A note on high dimensional linear regression with interactions. *The American Statistician* **71** 291–297.

HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics* **58** 435–447.

KIM, S. and XING, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the 27th International Conference on International Conference on Machine Learning* 543–550.

KIM, S. and XING, E. P. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eQTL mapping. *The Annals of Applied Statistics* **6** 1095–1117.

KLEINER, A., TALWALKAR, A., SARKAR, P. and JORDAN, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society* **B76** 795–816.

LI, J., NETESSINE, S. and KOULAYEV, S. (2018). Price to Compete... With Many: How to Identify Price Competition in High Dimensional Space. *Management Science* **64** 3971–4470.

LI, R., ZHONG, W. and ZHU, L. (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association* **107** 1129–1139.

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40** 1846–1877.

LIU, W. and SHAO, Q. M. (2014). Phase transition and regularized bootstrap in large-scale t-tests with false discovery rate control. *The Annals of Statistics* **42** 2003–2025.

PEÑA, V. H., LAI, T. L. and SHAO, Q.-M. (2008). *Self-normalized processes: Limit theory and Statistical Applications*. Springer Science & Business Media.

QU, H., RYZHOV, I. O., FU, M. C., BERGERSON, E., KURKA, M. and KOPACEK, L. (2019). Learning demand curves in B2B pricing: a new framework and case study. *Production and Operations Management (to appear)*.

RUBEN, H. (1962). A New Asymptotic Expansion For The Normal Probability Integral And Mill's Ratio. *Journal of the Royal Statistical Society* **B24** 177–179.

RUDIN, C., WALTZ, D., ANDERSON, R. N., BOULANGER, A., SALLEB-AOUISSI, A., CHOW, M., DUTTA, H., GROSS, P. N., HUANG, B., IEROME, S., ISAAC, D. F., KRESSNER, A., PASSONNEAU, R. J., RADEVA, A. and WU, L. (2012). Machine learning for the New York City power grid. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34** 328–345.

RYZHOV, I. O., HAN, B. and BRADIĆ, J. (2016). Cultivating disaster donors using data analytics. *Management Science* **62** 849–866.

SMITHSON, M. and MERKLE, E. C. (2013). *Generalized linear models for categorical and continuous limited dependent variables*. CRC Press.

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and Testing Dependence by Correlation of Distances. *The Annals of Statistics* **35** 2769–2794.

SZÉKELY, G. J. and RIZZO, M. L. (2009). Brownian Distance Covariance. *The Annals of Applied Statistics* **3** 1233–1303.

SZÉKELY, G. J. and RIZZO, M. L. (2012). On the uniqueness of distance covariance. *Statistics & Probability Letters* **82** 2278–2282.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B58** 267–288.

VAN DE GEER, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics* **36** 614–645.

XUE, Z., WANG, Z. and ETTL, M. (2016). Pricing personalized bundles: A new approach and an empirical study. *Manufacturing & Service Operations Management* **18** 51–68.

YAN, X. and BIEN, J. (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science* **32** 531–560.

YEKUTIELI, D. (2008). Hierarchical false discovery rate-controlling methodology. *Journal of the American Statistical Association* **103** 309–316.

YUAN, M. and LIN, Y. (2006). Model Selection and Estimation in Regression with Group Variables. *Journal of the Royal Statistical Society* **B68** 49–67.

ZHANG, K., BHATTACHARYYA, S. and RAM, S. (2016). Large-Scale Network Analysis for Online Social Brand Advertising. *MIS Quarterly* **40** 849–868.

ZHAO, S. D. and LI, Y. (2012). Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* **105** 397–411.

ZHAO, P., ROCHA, G. and YU, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics* **37** 3468–3497.

ZHOU, J., FOSTER, D. P., STINE, R. A., UNGAR, L. H. and GUYON, I. (2006). Streamwise Feature Selection. *Journal of Machine Learning Research* **7** 1861–1885.

ZHU, L. P., LI, L., LI, R. and ZHU, L. X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106** 1464–1475.