A dynamic screening algorithm
for hierarchical binary marketing data

Yimei Fan*  Yuan Liao†  Ilya O. Ryzhov‡

May 2, 2018

Abstract

In many applications of business and marketing analytics, predictive models are fit using hierarchically structured data: common characteristics of products and customers are represented as categorical variables, and each category can be split up into multiple subcategories at a lower level of the hierarchy. Hundreds of thousands of binary variables may be required to model the hierarchy, 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 involving a large volume of B2B transaction data.

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

*Applied Mathematics, Statistics & Scientific Computation, University of Maryland
†Economics, Rutgers University
‡Robert H. Smith School of Business, University of Maryland (email: iryzhov@rhsmith.umd.edu)
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 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.

2. **Market segmentation.** In the previous application, the dummy variables may represent attributes of the customer rather than the product. For instance, in business-to-business pricing, the seller may classify client firms based on their geographic location (which may be described at the regional, country, or city levels) or by attributes of their economic sector.

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 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 the model’s ability to make accurate predictions (Fan et al., 2014). In such situations, statistical model selection (also known as variable selection; see Fan & Lv, 2010) becomes an extremely useful practical tool for reliably recovering a sparse set of significant features, while removing large numbers of insignificant features. This improves predictive power, but also helps managers to know the degree of aggregation sufficient for making accurate
predictions. Thus, in our first motivating example, it may be sufficient to include a single variable for saws, but necessary to distinguish between several individual brands of hammers; in the second example, we would like the flexibility to define market segments as broadly or narrowly as may be required for prediction.¹

Model selection is also extremely useful for practical computation. While the theoretical literature often focuses on problems where the size $p$ of the feature space is large relative to the sample size $n$, there are also numerous practical applications where $n > p$, but both $n$ and $p$ are large (in the tens or hundreds of thousands), which may cause severe computational difficulties for traditional estimation procedures (Kleiner et al., 2014). Reducing the feature space mitigates this difficulty and leads to more easily interpretable forecasts.

In this paper, we propose a new model selection algorithm, which exploits the hierarchical structure of the data in our motivating applications to improve computational efficiency and avoid exploring the entire feature space. Our method is based on a unique property of the hierarchy: each binary feature has a single “parent” feature (e.g., “tools” is the parent of “hammers”), and any feature can be a member of the set $\mathcal{A}$ of significant variables only if its parent is also in $\mathcal{A}$. That is, if a feature is irrelevant, then all of its descendants must be irrelevant as well. This assumption, which we call the extinction property, is suitable for our motivating applications (it does not make sense for hammers to be important if tools in general are not), and thus is assumed to hold on the data-generating process.

Because our motivating applications use binary data, we adopt the distance correlation (DC) criterion of Székely et al. (2007) to test the significance of a particular candidate feature. DC measures dependence under very general assumptions (see Li et al., 2012b for a discussion) that are easily satisfied in the binary-data setting. In the process, however, we prove that DC is equivalent to classical Pearson correlation when both the response and data are binary, which allows the criterion to be computed more efficiently and provides

¹In 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.
a conceptual bridge between these two notions of correlation. Thus, while our framework can potentially be generalized, it offers especial practical advantages in the binary setting. Furthermore, since we do not impose any particular functional form on the distribution of the response (crucially, we do not assume a linear relationship), our equivalence result establishes the validity of Pearson correlation in GLMs or other models where the data are binary.\footnote{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.}

We then incorporate this criterion into a new dynamic selection procedure that explores the hierarchy from higher (more aggregate) to lower (more disaggregate) levels. We add features to a candidate set and evaluate their marginal DC/Pearson correlation with the response. If this value is above a certain threshold, the feature is accepted and its children become candidates; if the correlation is too low, the feature is screened out together with all of its descendants. This approach differs from popular benchmark methods, such as sure independence screening (Fan & Lv, 2008) and Lasso (Tibshirani, 1996), in that it achieves the extinction property for any finite sample size: once the parent has been ruled out, we do not explore any of its descendants even if their empirical correlation is high (which is likely to happen in finite-sample settings, and makes it difficult for standard methods to achieve the extinction property). We prove that, under a standard set of assumptions used in the statistical literature, the procedure recovers the exact set of significant features with probability 1 as both $n$ and $p$ become large.

We also examine the practical performance of the dynamic DC-based algorithm (DDC) in numerical experiments on both simulated and real data. The simulated experiments find that DDC is competitive against Lasso and other benchmarks in correctly identifying significant features when the data are high-dimensional. We also consider real data from a practical demand modeling application in the context of B2B transactions, and demonstrate that predictive power is greatly improved after DDC is first used for screening. Although $n > p$ in this dataset, estimation poses significant computational challenges since $p \approx 50,000$
and \( n \approx 250,000 \). DDC also scales much better to larger datasets than do the benchmark methods, and thus offers significant practical benefits.

Thus, the paper makes the following contributions. 1) We create a new dynamic selection procedure for feature screening in hierarchically structured marketing data. The procedure respects the extinction property for any finite sample size, which cannot be guaranteed by other procedures. 2) In the particular case of binary data, we show that the DC criterion is equivalent to Pearson correlation, which justifies the use of the latter in generalized linear models (GLMs) and also leads to a significant computational speedup. 3) We prove that the set of features selected by the DDC method is asymptotically equal to \( A \). 4) We demonstrate the practical benefits of DDC, in terms of selection accuracy, predictive power, and computational efficiency, on both simulated and real data, including a case application involving a large volume of B2B transactions.

2 Literature Review

In this section, 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 et al., 2016; Li et al., 2017), where predictive models work together with optimization or other decision-theoretic tools. Although we do not directly study optimization in this paper, nonetheless our work could complement, e.g., Xue et al. (2016) or Qu et al. (2018), which study price optimization under a calibrated statistical model with a fixed set of features (i.e., a set of features that have already been selected).

Within the statistical literature, our paper is closest to the work on sure independence screening (SIS), a methodology first proposed by Fan & Lv (2008) for linear regression problems and subsequently extended to GLMs (Fan & Song, 2010), nonparametric models
(Fan et al., 2011), survival models (Zhao & Li, 2012) and model-free 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 regression features. SIS further assumes that this influence is reflected in the marginal correlation between individual features and the response, i.e., features are irrelevant if their marginal correlation with the response is below a certain threshold. In the linear regression setting, Pearson correlation is used as the screening criterion, though other criteria such as Kendall rank correlation have also been suggested (Li et al., 2012). Most commonly, only the marginal correlation for each feature is considered, though Fan et al. (2009) has studied more complex forms of joint dependence, while Barut et al. (2016) has proposed conditional criteria based on side information.

Since Pearson correlation measures the strength of a linear relationship, other criteria are typically used outside the linear regression setting, making SIS more computationally intensive. For example, in GLMs, Fan & Song (2010) proposes solving a marginal maximum likelihood problem for every feature (the streamwise selection method of Zhou et al., 2006 also uses a similar idea). However, Székely et al. (2007) developed an alternative screening criterion, called the distance correlation (DC), which can replace Pearson correlation in SIS under much more general assumptions on the model (see Székely & Rizzo, 2012 for a proof of uniqueness). In Li et al. (2012), it was shown that DC-based SIS retained asymptotic consistency. Numerous extensions of DC are available for, e.g., measuring the dependence of multivariate distributions and stochastic processes (see Székely & Rizzo, 2009, Székely & Rizzo, 2014 and the references therein), and Huo & Székely (2016) extended the applicability of DC further by developing more efficient estimation procedures. We also adopt DC as the screening criterion for our procedure; in addition to its generality, it turns out to admit substantial computational simplifications in our motivating setting of binary data.

We are not aware of any prior work in statistics that has considered hierarchical data
structures in the context of model selection. One somewhat related stream of literature is the work on group Lasso (Yuan & Lin, 2006), which uses a penalized optimization framework that combines selection and estimation (see Nardi & Rinaldo, 2008 for a theoretical treatment, and Meier et al., 2008, Roth & Fischer, 2008 for extensions to GLMs). A hierarchical Lasso method was introduced by Zhou & Zhu (2010) for group variable selection; however, the term “hierarchical” in this work refers to the structure of the Lasso penalties rather than the data. For the data structure considered in the present paper, group Lasso methods do not lead to the desired extinction property.

The recent work on interaction screening (Hao & Zhang, 2014) is perhaps the closest to our paper with regard to the data structure under consideration. This work assumes a linear regression model with “interaction” features whose values are products of pairs of “base” features. It is then assumed that an interaction can only be significant if one or both of the base components are, which bears some resemblance to the extinction property in our paper; one could then apply screening or other model selection methods (Hao & Zhang, 2016) in two stages, so that interactions are considered only after the base components have been selected. The work by Fan et al. (2015) extends this idea to classification problems where the features are generated from Gaussian mixture distributions. See also Bien et al. (2013) and She & Jiang (2016) for regularization-based approaches to this type of model selection problem (as well as Zhao & Leng, 2016 for a theoretical analysis). These methods cannot be directly applied to our setting, as we do not use linear regression and the hierarchy in our problem may be multi-layered.

3 Data and Model

Let there be $n$ observations $(x_1, y_1), \ldots, (x_n, y_n)$ that are independent and identically distributed. We let $X = (X_1, \ldots, X_p)$ 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 com-
ponent 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

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

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

We let \( \mathbf{X}_\mathcal{A} = \{X_j : j \in \mathcal{A}\} \) and \( \mathbf{X}_{\mathcal{A}^c} = \{X_j : j \in \mathcal{A}^c\} \) represent the subvectors consisting of relevant and irrelevant variables respectively. The goal is to identify \( \mathcal{A} \), and at the same time, achieve the extinction property to be defined below.

**Figure 1: Illustration of a hierarchical data structure.**

We now impose a hierarchical structure on the features. For \( j = 1, \ldots, p \), 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 \( j \)th 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 \( j \)th feature. Thus, \( \mathcal{C}(j) \subseteq \mathcal{D}(j) \). For instance, in the example shown in Figure 1, we have \( \mathcal{P}(2) = \{1\} \), \( \mathcal{C}(2) = \{3,4\} \) and \( \mathcal{D}(2) = \{3,4,5,6,7,8\} \).

Next, we define the extinction property, which is the key condition that allows us to avoid
exploring the entire feature space.

**Assumption 3.1** (extinction property). If $j \in A^c$, then $k \in A^c$ for all $k \in D(j)$.

This condition assumes that all descendants of irrelevant features are also irrelevant, and is reasonable 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 $F(1 \mid X)$ being the probability of a sale (a stand-in for demand) given a large number of binary product attributes in $X$. 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.”

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 does not significantly affect the purchase probability, individual brands of TVs of that same size should not play a role either. Note that 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.

### 4 Methodology

We now describe our new dynamic screening algorithm for identifying features in $A$. 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 $F(Y \mid X)$ 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 et al., 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

$$ dcov(X,Y) = \left( \int |\phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t)|^2 (\pi^2 s^2 t^2)^{-1} ds dt \right)^{\frac{1}{2}}. \quad (4.1) $$

The distance correlation is defined as

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

and is always positive.

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

$$ \hat{dcov}(X,Y) = \left( \hat{S}_1 + \hat{S}_2 - 2\hat{S}_3 \right)^{\frac{1}{2}}, \quad (4.2) $$

$$ \hat{dcorr}(X,Y) = \frac{\hat{dcov}(X,Y)}{\sqrt{\hat{dcov}(X,X)\cdot \hat{dcov}(Y,Y)}}, \quad (4.3) $$

where

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

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

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

In the special case where both $X$ and $Y$ are binary, we find that (4.1) is equivalent to the absolute value of their Pearson correlation. Perhaps more surprisingly, (4.3) is almost surely
equivalent to the sample Pearson correlation

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

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) \( d\text{cov} (X, Y) = 2 |\text{cov} (X, Y)|, \quad d\text{corr} (X, Y) = |\text{corr} (X, Y)|; \)

(ii) \( \hat{d}\text{cov} (X, Y) = \frac{2(n - 1)}{n} |\hat{\text{cov}} (X, Y)|, \quad \hat{d}\text{corr} (X, Y) = |\hat{\text{corr}} (X, Y)|, \)

where \( \hat{\text{cov}} \) and \( \hat{\text{corr}} \) respectively denote the usual sample covariance and correlation.

Proposition 4.1 greatly simplifies the computation of DC, as (4.4) can be calculated more efficiently than (4.2)-(4.3). More importantly, it justifies the use of Pearson correlation outside the traditional setting of linear regression. It is well-known that Pearson correlation measures the strength of a linear relationship; however, since the distribution function \( F \) may represent a GLM or other nonlinear model, it is not immediately clear that Pearson correlation is still a valid measure of dependence in our setting. On the other hand, the validity of DC holds under much more general assumptions on the model (Li et al., 2012b), 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.

### 4.2 Dynamic Distance Correlation (DDC) Algorithm

We first give an overview of the proposed algorithm before stating it formally. The \( j \)th feature is assumed to be relevant if \( d\text{corr} (X_j, Y) \geq K_n \), 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 \( \hat{d}\text{corr} (X_j, Y) < K_n \) 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 \( \hat{\text{dcorr}}(X_j, Y) \geq K_n \), 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 such as time and memory 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 consistent, \( K_n \) should be slightly larger than the maximum estimation error \( \max_{j \leq p} \left| \hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| \) of the distance correlations.

We now give a formal statement. Let \( S_\ell \) denote the index set of selected features by stage \( \ell \) of the algorithm, and let \( M_\ell \) denote the index set of the current candidates at stage \( \ell \). These will be updated dynamically by the procedure.

**Step 1** (initialization) Set \( \ell = 0 \), \( S_0 = \emptyset \), and let \( M_0 \) 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 \( j \in M_\ell \), compute \( \hat{\text{dcorr}}(X_j, Y) \) and set \( M_0 = M_0 \setminus \{ j \} \) if \( \hat{\text{dcorr}}(X_j, Y) < K_n \).

**Step 3** (termination) If \( M_\ell = \emptyset \), return \( \hat{\mathcal{A}} = S_\ell \) and stop. Otherwise, continue.

**Step 4** (selection) Find

\[
 j_\ell = \arg \max_{j \in M_\ell} \hat{\text{dcorr}}(X_j, Y),
\]

(4.5)

and update

\[
 S_{\ell+1} = S_\ell \cup \{ j_\ell \},
\]

\[
 M_{\ell+1} = (M_\ell \setminus \{ j_\ell \}) \cup \mathcal{C}(j_\ell),
\]
where $C(j_\ell)$ is the set of children of $j_\ell$ as defined in Section 3.

**Step 5** (iteration) Increment $\ell$ by 1 and return to Step 2.

In the algorithm, $M_\ell$ 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 $j_\ell$ in (4.5), 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 $M_\ell$ 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 $\hat{A}$, which is different from the *screening set*

$$\hat{B} = \{j \in \{1, 2, 3, \ldots, p\} : \hat{d}_{\text{corr}}(X_j, Y) \geq K_n\},$$

which includes all features whose empirical DC is above the threshold. It is clear that $\hat{A} \subseteq \hat{B}$. In the finite-sample setting, there may be $j$ and $k \in D(j)$ such that $\hat{d}_{\text{corr}}(X_j, Y) < K_n$, but $\hat{d}_{\text{corr}}(X_k, Y) \geq K_n$. Such a $k$ would be an element of $\hat{B}$ but not $\hat{A}$. This is a fundamental difference between our dynamic approach and the classical SIS technique of Fan & Lv (2008). SIS arranges the empirical correlations in descending order and simply screens out a certain proportion of features ranked at the bottom. Due to sampling error, this approach may violate the extinction property since some features may be screened out, but their descendants may still be selected. Furthermore, SIS requires us to estimate the marginal correlation for every feature, which may be expensive when $p$ is large. Our proposed algorithm avoids both of these issues, since screening out a feature in step 2 will automatically rule out all of its descendants. In this way, if the problem is sufficiently sparse, we will avoid having to compute empirical DCs for a substantial proportion of the feature.
Remark 4.1. 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 $\tau$ based robust correlation (Li et al., 2012a).

4.3 Descriptive Example

To illustrate our algorithm, we briefly discuss a descriptive example on a hierarchy with three levels shown in Figure 2. As there are two features in the top layer, we initialize $M_0 = \{1, 2\}$ and $S_0 = \emptyset$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Illustration of the DDC algorithm. Due to the extinction property, features 9, 10, 13 and 14 are screened out without being examined directly.}
\end{figure}

Iteration 1: steps 2-5. We first evaluate the empirical DC for features 1 and 2. Suppose that $\hat{\text{dcorr}}(X_1, Y) > \hat{\text{dcorr}}(X_2, Y) > K_n$. 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

$$S_1 = \{1\}, \quad M_1 = \{2, 3, 4\}.$$
Iteration 2: steps 2-5. Suppose $\hat{d}_{\text{corr}}(X_3, Y) > \hat{d}_{\text{corr}}(X_2, Y) > K_n$, but $\hat{d}_{\text{corr}}(X_4, Y) < K_n$. Then, step 2 screens out feature 4, whence $M_1 = \{2, 3\}$, 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 $C(3) = \{7, 8\}$ become candidates, leading to the update

$$S_2 = \{1, 3\}, \quad M_2 = \{2, 7, 8\}.$$  

Iteration 3: steps 2-5. Suppose $\hat{d}_{\text{corr}}(X_2, Y) > \hat{d}_{\text{corr}}(X_7, Y) > K_n$ but $\hat{d}_{\text{corr}}(X_8, Y) < K_n$. Then, step 2 screens out feature 8, whence $M_2 = \{2\}$. Step 3 does not terminate, step 4 sets $j_2 = 2$, whence feature 2 is selected and the new candidates $C(2) = \{5, 6\}$ are added. The resulting update is

$$S_3 = \{1, 2, 3\}, \quad M_3 = \{5, 6, 7\}.$$  

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

$$S_4 = \{1, 2, 3, 5\}, \quad M_4 = \{7, 11, 12\}.$$  

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

$$S_5 = \{1, 2, 3, 5, 11\}, \quad M_5 = \{7\}.$$  

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

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

$$S_6 = \{1, 2, 3, 5, 7, 11\}, \quad M_5 = \emptyset.$$
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. It is clear that the selected set satisfies the extinction property.

5 Theoretical Analysis

The main result of this section is given in Theorem 5.1, which shows that the set \( \hat{\mathcal{A}} \) returned by the DDC procedure is asymptotically equal to the true set \( \mathcal{A} \) of relevant features. The proof is given in the Appendix; below, we give the necessary definitions and assumptions.

Let us define the threshold

\[
K_n = \frac{a_0}{\min_{j \leq p} \hat{\text{dco}}(X_j, X_j)^{1/2}\hat{\text{dco}}(Y, Y)^{1/2}} \sqrt{\frac{2 \log(p \vee n)}{n}}.
\]

(5.1)

where \( a_0 > 2.5 \) is a constant (we can take, for instance, \( a_0 = 2.51 \)). 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, it is necessary to control for the so-called “uniform deviation”

\[
S_n = \max_{j \leq p} \left| \hat{\text{dco}}(X_j, Y) - \text{dco}(X_j, Y) \right|
\]

which represents the noise in the problem. Accordingly, we select the threshold to dominate the noise; specifically, we choose \( b_n \to 0 \) to satisfy \( \mathbb{P}(S_n < a_0 b_n) \to \infty \) for some positive constant \( a_0 \). From the moderate deviation theory for self-normalized sums (Peña et al., 2008...
and Belloni et al., 2012), it is well-known that the choice

\[ b_n = \sqrt{\frac{2 \log p \vee n}{n}} \]

can uniformly control the self-normalized noise. With regard to the constant, we show in the Appendix that \( a_0 > 2.5 \) 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 (5.1).

From a purely practical point of view, (5.1) requires us to compute \( \min_j \hat{\text{dcov}}(X_j, X_j) \), 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 \( \min_j \hat{\text{dcov}}(X_j, X_j) \) with an empirical estimate, e.g., by sampling a small subset of \( \{1, \ldots, p\} \) and calculating the smallest \( \hat{\text{dcov}}(X_j, X_j) \) within this subset. A second alternative is to replace \( \min_j \hat{\text{dcov}}(X_j, X_j) \) in (5.1) with some function of \( p \) and \( n \) that declines more slowly than the empirical DCs (i.e., an asymptotic upper bound on \( \min_j \hat{\text{dcov}}(X_j, X_j) \)). 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 6.2) and obtain good performance.

We now state several regularity conditions needed to prove the main result. The first assumption simply ensures that we are in the high-dimensional setting, as is standard in the model selection literature; the second assumption ensures that we are able to separate the signal in the data from the noise.

**Assumption 5.1.** The data \( \{Y_i, X_{i1}, \ldots, X_{ip}\}_{i=1}^n \) are independent and identically distributed. As \( n \to \infty \), the number of features, \( p \), either stays constant or grows with \( n \), satisfying \( b_n = o(1) \).

**Assumption 5.2.** The following statements hold:

(i) \( \text{var} (Y) \gg b_n \) and \( \min_{j \leq p} \text{var} (X_j) \gg b_n \).

(ii) \( \min_{j \in A} \hat{\text{dcov}}(X_j, Y) \cdot \min_j \text{var}(X_j)^{1/2} \gg b_n \).
(iii) \( \max_{j \in A^c} \text{dcorr} \left( X_j, Y \right) = 0 \).

Condition (i) in Assumption 5.2 allows the variances to slowly decay to zero as \( n \to \infty \) (in other words, individual features may be observed rarely, as long as the sample size is large). Condition (ii) requires all relevant features to be sufficiently strongly correlated with the response, while condition (iii) assumes that irrelevant features are uncorrelated with the response.

It is straightforward to relax condition (iii) to allow irrelevant features to be weakly correlated with \( Y \), as long as this behavior can be clearly separated from that of the relevant features. For instance, one could replace condition (iii) above with

\[
\max_{j \in A^c} \text{dcorr} \left( X_j, Y \right) = \delta \sqrt{\frac{\log(p \vee n)}{n}} \tag{5.2}
\]

for some fixed \( \delta > 0 \). Virtually all of the literature on sure screening requires relevant features to be more strongly correlated with \( Y \) than irrelevant features, so (5.2) must decay more quickly than \( \min_{j \in A} \text{dcorr} \left( X_j, Y \right) \); however, as long as this holds, the analysis proceeds in the same way, so we use condition (iii) in our proofs for simplicity.

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

Recalling that \( \widehat{A} \) denotes the final selection returned by the proposed algorithm, we can now state the main feature selection consistency result. We use the notation \( |A| \) to denote the cardinality of a finite set \( A \).

**Theorem 5.1.** Under Assumptions 3.1, 5.1, and 5.2,

\[
P \left( \widehat{A} = A \right) \to 1.
\]
In addition, the DDC procedure will calculate $O\left(\sum_{j \in \hat{A}} |C(j)|\right)$ empirical distance correlations before terminating.

6 Numerical Studies

We assess the performance of the DDC algorithm on both simulated (Section 6.1) and real (Section 6.2) data. All experiments were conducted in the R environment (thus, computation times are reported for R code and statistical packages) using sparse matrix representations where possible.

6.1 Simulated Data

We generated multiple hierarchical binary data structures satisfying the extinction property. Two examples are presented; 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 $n = 1000$ respectively for the two examples. 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 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 $\kappa_i$ was generated as follows: if feature $i$ belongs to the top layer of the hierarchy, we let $\kappa_i$ be uniform on the interval $[-0.25, 0.25]$; otherwise, $\kappa_i$ is uniform on the interval $[-|\kappa_{\mathcal{P}(i)}|, |\kappa_{\mathcal{P}(i)}|]$. In this way, the correlation is decreasing as we move toward the disaggre-
gate levels of the hierarchy (as we would expect to see in an application). Then, $\kappa_i$ was used to set the distribution

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

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

To simulate $X_i$ for $i \in \mathcal{A}$, we first sample $Y$ from a Bernoulli distribution with success probability 0.5. Then, if $X_{P(i)} = 1$, we generate the value of $X_i$ from the above conditional distribution. If $X_{P(i)} = 0$, we set $X_i = 0$ as is commonly the case in practical applications with hierarchical data (see Section 6.2 for one such application). For $i \notin \mathcal{A}$, we simply generate $X_i$ from an independent Bernoulli distribution with success probability 0.3.

We implemented the DDC algorithm together with three benchmarks: Lasso (Tibshirani, 1996); the streamwise regression (SR) approach of Zhou et al. (2006); and sure independence screening or SIS (Fan & Lv, 2008). Lasso is a well-known and widely-used method for reducing high-dimensional feature spaces, but tends to run slowly when both $n$ and $p$ are moderately large. The streamwise regression (SR) method performs a univariate (marginal) regression for each individual feature, analogously to the screening approach of Fan & Song (2010). We used univariate logistic regression as the criterion for SR because the response variable is binary. SIS uses the same screening criterion (Pearson correlation) that we use in DDC; however, SIS estimates this criterion for every feature, without considering the hierarchy, and simply selects $d$ features that appear to have the highest correlation. The quantity $d$ is a tunable parameter; Fan & Lv (2008) gives several suggestions for how to choose it. We experimented with all of them and found that $d = \lfloor n/\log n \rfloor$ produced the best results. For DDC, we used the threshold $K_n = (\log \log n) \sqrt{\frac{\log p}{n}}$.

All methods 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 $\mathcal{A}^c$;
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.

Table 1 presents some summary statistics across 500 simulated datasets in the first example, while Figure 3 shows the empirical distributions of TPR achieved by the four methods. DDC tends to achieve higher TPR than the other benchmarks, while the SR method achieves the lowest FPR. It should be noted, however, that FPR is generally much lower than TPR for all four methods, and the differences in FPR between them are extremely small. Fur-
Table 1: Performance of model selection methods on first simulated example (500 datasets). The SD column gives the estimated standard deviation of performance on a single dataset.

Table 2 and Figure 4 present analogous results for the second simulated example ($p \approx 170,000$). Lasso, SIS and SR perform well in terms of FPR, although this metric is small for all four methods. With regard to TPR, Lasso generally underperforms, while DDC, SIS and SR are competitive (DDC has a slight advantage overall, but its worst-case performance is slightly below that of SR and SIS). However, SR and SIS experience a very notable increase in computational cost: SIS runs 70 times more slowly than DDC, while for SR this factor is over 400.

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. The computation times suggest that DDC may scale better to problems with massive data sizes; we explore such a setting in greater depth in the following case study.

Table 2: Performance of model selection methods on second simulated example (50 datasets).
6.2 Application to B2B Transaction Data

We also implemented our method on two historical datasets provided by Vendavo, Inc., a firm specializing in business-to-business (B2B) pricing science. These data cover a large volume of B2B transactions involving numerous products. The response variable is binary, since the customer in each transaction may either accept or reject the deal. In both datasets, products are aggregated using a hierarchy with four levels; an individual product ID belongs to a ProductLevel1, ProductLevel2, and ProductLevel3, with an additional disaggregate level where one feature per product is added. The first dataset contains approximately...
Table 3: Performance of model selection methods on first pricing dataset. All numbers are averaged over 5 folds.

6,000 distinct products, \( p \approx 8000 \) features and \( n \approx 10^4 \) transactions, whereas the second dataset covers approximately 58,000 distinct products, uses \( p \approx 64,000 \) features, and records \( n \approx 2.5 \times 10^5 \) transactions. Both datasets are very noisy, with a low proportion of positive responses and many low-volume products that appear infrequently. All of these factors make prediction quite challenging.

Model selection is of great practical value in this application, as it serves three purposes. First, model selection helps to reduce the computational complexity of estimating a regression model on the data; recent statistical literature (Kleiner et al., 2014; Bradić, 2016) has observed that traditional estimation methods may work poorly in the large-sample setting with the advent of “massive” datasets in which both \( n \) and \( p \) are large. A screening approach is particularly helpful in this setting, since we work with the marginal DC of each feature rather than the entire design matrix. Second, model selection improves the interpretability of the resulting model, as managers are now able to see the exact level of detail required to capture the effect of a class of products. Third, as we demonstrate below, a sparser model will have better empirical predictive power in this setting, while standard models may still be subject to spurious correlation, noise accumulation, and other known practical issues (Fan et al., 2014).

Since the true sparse feature set \( A \) is unknown in this problem, we evaluate DDC and other methods according to their predictive power. Thus, we first conduct a screening step using the method of choice (DDC, Lasso, SR or SIS). We then run a logistic regression model on the selection set \( \hat{A} \) returned by that method; this estimation step is required for
all screening techniques as they do not directly perform estimation, and is recommended for Lasso as a way of reducing the estimation bias (Belloni & Chernozhukov, 2013). Using 5-fold cross-validation, we then calculate the AUC, or area under the ROC curve (Smithson & Merkle, 2013) for the estimated post-selection model. 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. All methods are tuned to optimize their out-of-sample predictive power; for DDC, we treat the threshold $K_n$ as a tunable parameter. We also report computation times for both selection and estimation, which is important for understanding how well the different methods scale to larger data.

Table 3 shows results for the smaller dataset (10 thousand transactions). Here, all four methods achieve similar predictive power, with DDC having a slight lead. However, DDC runs about 30% faster than SIS, about twice as fast as Lasso, and over 500 times faster than the streamwise method. All models select progressively smaller proportions of the features in each layer, in line with our expectation that more disaggregate levels contain more irrelevant features. The computational cost of estimation is generally negligible compared to that of screening, for all three methods.

Table 4 shows analogous results on the larger dataset (250 thousand transactions). DDC maintains approximately the same level of predictive power as before; however, the three benchmark methods all experience significant performance degradation due to selecting too many features. In particular, both Lasso and SIS now produce models that are essentially guessing the outcome, with no predictive power. By contrast, DDC produces the sparsest model, and screens out many more features at the more disaggregate levels. Furthermore, DDC is more scalable than the benchmarks, as it now runs over 21 times faster than SIS (combining both steps), over 22 times faster than Lasso, and over 3700 times faster than SR.

Based on these results, we conclude that DDC offers significant practical potential in applications where the data has 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
Table 4: Performance of model selection methods on second pricing dataset. All numbers are averaged over 5 folds.

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, an issue that arises in multiple 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.

References


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

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

For the left hand side, we have

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

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

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

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

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

The second and fourth terms can likewise be simplified as

\[ C = (e^{is} - 1)(e^{it} - 1)P(Y = 1)(P(X = 1, Y = 1) - P(X = 1)P(Y = 1)) \]

Combining these together yields

\[ B + C \]
\[ = (e^{is} - 1)(e^{it} - 1)(P(X = 1, Y = 1) - P(X = 1)P(Y = 1))(P(Y = 0) + P(Y = 1)) \]
\[ = (e^{is} - 1)(e^{it} - 1)(P(X = 1, Y = 1) - P(X = 1)P(Y = 1)) \]
\[ = (e^{is} - 1)(e^{it} - 1)cov(X, Y). \]

Recalling the definition of \( dcov(X, Y) \), we write

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

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

\[ dcov^2(X, Y) = \int_{\mathbb{R}^2} (e^{is} - 1)(e^{-is} - 1)(e^{it} - 1)(e^{-it} - 1)cov^2(X, Y)w(s, t) ds dt \]
\[ = A \cdot cov^2(X, Y), \]

32
where

\[
A = \int_{\mathbb{R}^2} |(e^{is} - 1)(e^{it} - 1)|^2 w(s, t) ds \, dt \\
= \int_{\mathbb{R}^2} (2 - 2 \cos s)(2 - 2 \cos t) w(s, t) ds \, dt \\
= 2. \tag{A.1}
\]

Thus,

\[
dcov(X, Y) = 2|\text{cov}(X, Y)|, \quad dcov(X, X) = 2\text{cov}(X, X) = 2\text{var}(X),
\]

whence

\[
dcorr(X, Y) = \frac{2|\text{cov}(X, Y)|}{2\sqrt{\text{var}(X)\text{var}(Y)}} = |\text{corr}(X, Y)|,
\]

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 et al. (2007) that will be useful later.

**Lemma A.1.** The estimator \( \hat{\text{dcov}}(X, Y) \) satisfies

\[
\hat{\text{dcov}}^2 (X, Y) = \int_{\mathbb{R}^2} \| f_{X,Y}^n (s, t) - f_X^n (s)f_Y^n (t) \|^2 w(t, s) ds \, dt,
\]

where

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

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

\[
f_X^n (s) = \frac{1}{n} \sum_{k=1}^n \exp \{ i \langle s, x_k \rangle \}, \quad f_Y^n (t) = \frac{1}{n} \sum_{k=1}^n \exp \{ i \langle t, y_k \rangle \}.
\]
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

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

**Proof:** We rewrite $f_{X,Y}^n(s,t)$, $f_X^n(s)$, and $f_Y^n(t)$ 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

$$f_{X,Y}^n(s,t) = \frac{1}{n} \sum_{k=1}^{n} \exp(isx_k + ity_k)$$

$$= \frac{1}{n} \left[ e^{i(s+t)}\#(x_k = 1, y_k = 1) + e^{is}\#(x_k = 1, y_k = 0) \
+ e^{it}\#(x_k = 0, y_k = 1) + \#(x_k = 0, y_k = 0) \right]$$

$$= \frac{1}{n} \left[ (e^{i(s+t)} - e^{is} - e^{it})\#(x_k = 1, y_k = 1) + e^{is}\#(x_k = 1) + e^{it}\#(y_k = 1) + \#(x_k = 0, y_k = 0) \right].$$

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,

$$f_X^n(s) = \frac{1}{n} \sum_{k=1}^{n} e^{isx_k} = \frac{1}{n} (e^{is}\#(x_k = 1) + \#(x_k = 0))$$

$$= \frac{1}{n} \left[ (e^{is} - 1)\#(x_k = 1) + n \right] = 1 + \bar{x}(e^{is} - 1),$$

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

$$f_X^n(s)f_Y^n(t) = (1 + \bar{x}(e^{is} - 1))(1 + \bar{y}(e^{it} - 1))$$

$$= 1 + \bar{x}(e^{is} - 1) + \bar{y}(e^{it} - 1) + \bar{x}\bar{y}(e^{is} - 1)(e^{it} - 1).$$
Consequently,
\[
f^n_{X,Y}(s, t) - f^n_X(s) f^n_Y(t)
\]
\[
= \frac{1}{n} \left( \sum_{k=1}^{n} x_k y_k - n \bar{x} \bar{y} \right) (e^{is} - 1)(e^{it} - 1) + \bar{x} + \bar{y} - 1 + \frac{1}{n} \left( \#(x_k = 0, y_k = 0) - \#(x_k = 1, y_k = 1) \right)
\]
\[
= \frac{1}{n} \left( \sum_{k=1}^{n} x_k y_k - n \bar{x} \bar{y} \right) (e^{is} - 1)(e^{it} - 1) + \frac{\#(x_k = 1) + \#(x_k = 0)}{n} - 1
\]
\[
= \frac{1}{n} \left( \sum_{k=1}^{n} x_k y_k - n \bar{x} \bar{y} \right) (e^{is} - 1)(e^{it} - 1),
\]
which completes the proof.

Combining Lemmas A.1 and A.2, we have
\[
\widehat{d_{cov}}^2 (X, Y) = \frac{1}{n^2} \left( \sum_{k=1}^{n} x_k y_k - n \bar{x} \bar{y} \right)^2 \cdot A,
\]
where $A$ is as in (A.1). The desired result follows.

A.2 Proof of Theorem 5.1

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
\[
\widehat{\text{var}}(Y) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2,
\]
\[
\widehat{\text{var}}(X_j) = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2,
\]
\[ \widehat{\text{var}}(X_j) = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \mathbb{E}X_{ij})^2. \]

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

**Lemma A.3.** Let

\[ Z_{ij} \in \{X_{ij} Y_i - \mathbb{E}X_{ij} Y_i, X_{ij} - \mathbb{E}X_{ij}, (X_{ij} - \mathbb{E}X_{ij})^2 - \mathbb{E}(X_{ij} - \mathbb{E}X_{ij})^2\}. \]

Define \( V_{nn,j}^2 = \sum_{i=1}^{n} Z_{ij}^2 \) and \( S_{nn,j} = \sum_{i=1}^{n} Z_{ij} \). Then, the following statements hold:

(i) As \( n \to \infty \),

\[ \mathbb{P} \left( \max_{j \leq p} \frac{1}{n} \left| \frac{S_{nn,j}}{V_{nn,j}^{1/2}} \right| \leq b_n \right) \to 1, \quad (A.2) \]

\[ \mathbb{P} \left( \max_{j \leq p} \frac{1}{n} \sum_{i} X_{ij} - \mathbb{E}X_{ij} \right) \leq b_n \right) \to 1. \quad (A.3) \]

(ii) For any \( d_0 > 0 \),

\[ \mathbb{P} \left( \frac{1}{n} \sum_{i} Y_i - \mathbb{E}Y \right) \leq d_0 b_n \right) \to 1, \quad (A.4) \]

\[ \mathbb{P} \left( \frac{1}{n} \sum_{i} Y_i - \mathbb{E}Y \right) \leq 0.5 d_0 b_n \right) \to 1. \quad (A.5) \]

(iii) For any \( d_0 > 0 \),

\[ \mathbb{P} \left( \max_{j \leq p} \frac{1}{n} \sum_{i} X_{ij} Y_i - \mathbb{E}X_{ij} Y_i \right) \leq b_n (0.5 + d_0) \right) \to 1, \quad (A.6) \]

\[ \mathbb{P} \left( \max_{j \leq p} \frac{1}{n} \sum_{i} Y_i - \mathbb{E}X_{ij} \right) \leq b_n (0.5 + d_0) \right) \to 1. \quad (A.7) \]
(iv) Under Assumption 5.2(i),
\[
\left| \frac{\widehat{\text{dcov}}(Y)}{\text{dcov}(Y)} - 1 \right| + \max_{j \leq p} \left| \frac{\widehat{\text{dcov}}(X_j)}{\text{dcov}(X_j)} - 1 \right| + \max_{j \leq p} \left| \frac{\text{var}(X_j)}{\text{var}(X_j)} - 1 \right| + \max_{j \leq p} \left| \frac{\text{var}(X_j)}{\text{var}(X_j)} - 1 \right| = o_P(1).
\]

(v) For any \( d_0 > 0 \),
\[
\mathbb{P}\left( \max_{j \leq p} \left| \frac{\text{var}(X_j)^{1/2} - \text{var}(X_j)^{1/2}}{\text{var}(X_j)^{1/2}} \right| \leq (0.5 + d_0)b_n \right) \to 1.
\]
\[
\mathbb{P}\left( \max_j \left[ \frac{\text{dcov}(X_j)^{1/2} - \text{dcov}(X_j)^{1/2}}{\text{dcov}(X_j)^{1/2}} \right] - \sqrt{2}(0.5 + d_0)b_n \right) \leq \frac{1}{\sqrt{n} - 1} \to 1.
\]
\[
\mathbb{P}\left( \left| \frac{\text{dcov}(Y)^{1/2} - \text{dcov}(Y)^{1/2}}{\text{dcov}(Y)^{1/2}} \right| \leq \frac{\sqrt{2}d_0b_n}{\text{dcov}(Y)^{1/2}} + \frac{1}{\sqrt{n} - 1} \right) \to 1.
\]

Proof: (i) Observe that \( \mathbb{E}Z_{ij} = 0 \) and the random variables \( Z_{ij} \) are independent across \( i \leq n \). 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 \( 0 < x < Cn^{1/6}/l_n - 1 \), we have
\[
\left| \max_{j \leq p} \mathbb{P}(|S_{nn,j}/V_{nn,j}| > x) \right| \to 0,
\]
where \( \Phi \) denotes the standard normal cdf. Now, choose \( x = \Phi^{-1}(1 - \gamma_n/(2p)) \) and define \( \gamma_n = (\log(\log(p \wedge n)))^{-1} \). Then, \( 2p(1 - \Phi(x)) = \gamma_n \) and \( \gamma_n = o(1) \) by construction, whence
\[
\mathbb{P}\left( \max_{j \leq p} \left( \frac{1}{n} \left| \frac{S_{nn,j}}{V_{nn,j}} \right|^{1/2} \right) > \frac{x}{n^{1/2}} \right) = \mathbb{P}\left( \max_{j \leq p} \left| \frac{S_{nn,j}}{V_{nn,j}} \right| > x \right)
\leq p \max_{j \leq p} \mathbb{P}(|S_{nn,j}/V_{nn,j}| > x)
\leq 2p(1 - \Phi(x))(1 + o(1))
= \gamma_n(1 + o(1)).
\]

To complete the proof of (A.2), it remains to show that \( x \leq \sqrt{2\log p \wedge n} \), which is equivalent to \( \mathbb{P}(N(0, 1) > \sqrt{2\log p \wedge n}) \leq \gamma_n/(2p) \). This is achieved by applying the Mill’s ratio
inequality (Ruben, 1962). From this, (A.3) is obtained by setting \( Z_{ij} = X_{ij} - \mathbb{E}X_{ij} \). Thus, part (i) is proved.

(ii) Define

\[
Z_i = \frac{Y_i - \mathbb{E}Y}{\sqrt{\text{var}(Y)}}.
\]

Then, we have \(|\frac{1}{n} \sum Z_i| = O_P(n^{-1/2})\). It follows that \(|\frac{1}{n} \sum Z_i| = o_P(b_n)\), implying (A.4) for any \( d_0 > 0 \). On the event where (A.4) holds w.p. 1, it follows that

\[
\left| \frac{1}{n} \sum_i Y_i - \mathbb{E}Y \right| \leq d_0 b_n \sqrt{\text{var}(Y)} \leq 0.5d_0 b_n,
\]

completing the proof of (A.5).

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

\[
E_1 = \left\{ \max_{j \leq p} \frac{1}{n} \left| S_{nn,j} \right| \leq \frac{1}{n} \left( \frac{1}{n} \sum_{i,j} V_{mm,j}^2 \right)^{1/2} \leq b_n \right\},
\]

we have

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

(A.8)

A crude but simple bound for the right-hand side of (A.8) is

\[
\max_j \left( \frac{1}{n} \sum_{i=1}^n (X_{ij} Y_i - \mathbb{E}(X_{ij} Y_i))^2 \right)^{1/2} \leq 2.
\]

It follows that

\[
\frac{1}{n} \sum_{i=1}^n (X_{ij} Y_i - \mathbb{E}X_{ij} Y_i)^2 = \frac{1}{n} \sum_{i=1}^n X_{ij} Y_i + (\mathbb{E}(X_{ij} Y_i))^2 - \frac{2}{n} \sum_{i=1}^n X_{ij} Y_i \mathbb{E}(X_{ij} Y_i) \leq \mathbb{E}(X_{ij} Y_i) - (\mathbb{E}(X_{ij} Y_i))^2 + o_P(1)
\]
\[ \leq 0.25 + o_P(1) \]

uniformly in \( j \leq p \). Thus, for any \( d_0 > 0 \), (A.6) holds with probability approaching 1. Equation (A.7) follows from the same argument, which is omitted here.

(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

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

we calculate

\[
\max_{j \leq p} \left| \frac{\tilde{\text{var}}(X_j) - \text{var}(X_j)}{\text{var}(X_j)} \right| \leq \frac{\max_{j \leq p} \sum_i Z_{ij}}{\min_j \text{var}(X_j)} \leq \frac{\max_j (\frac{1}{n} \sum_i Z_{ij}^2)^{1/2}}{\min_j \text{var}(X_j)} = o_P(1),
\]

where the last equality follows from the boundedness of \( \max_j |Z_{ij}| \) and from Assumption 5.2(i). This also implies that \( \max_j \tilde{\text{var}}(X_j) = O_P(1) \).

Furthermore, again applying (i) proved above, we have

\[
\left| \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 - \tilde{\text{var}}(X_j) \right| \leq (2|\bar{X}_j| + \mathbb{E}X_{ij})(\bar{X}_j - \mathbb{E}X_{ij}) = O_P(b_n).
\]

uniformly in \( j \leq p \). This also implies

\[
\max_j \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \leq \max_j \tilde{\text{var}}(X_j) + o_P(1) = O_P(1).
\]

Thus,

\[
\max_j |\tilde{\text{var}}(X_j) - \text{var}(X_j)| \leq \max_j |\tilde{\text{var}}(X_j) - \text{var}(X_j)| + \max_j \left| \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 - \tilde{\text{var}}(X_j) \right| + O_P \left( \frac{1}{n} \right) \max_j \frac{1}{n} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2
\]

39
\[ = \text{O}_P(b_n). \]

Since \( \min_j \var(X_j) \gg b_n \) by Assumption 5.2(i), we have

\[
\max_{j \leq p} \left| \frac{\tilde{\var}(X_j) - \var(X_j)}{\var(X_j)} \right| = \text{o}_P(1).
\]

The same argument also yields \( \left| \frac{\tilde{\var}(Y)}{\var(Y)} - 1 \right| = \text{o}_P(1) \). Now, we calculate

\[
\left| \frac{\tilde{\text{dcov}}(Y)}{\text{dcov}(Y)} - 1 \right| = \left| \frac{2(n-1)}{n} \frac{\tilde{\var}(Y)}{2\var(Y)} - 1 \right| \leq \left| \frac{\var(Y)}{\var(Y)} - 1 \right| (1 + n^{-1}) + \frac{1}{n} = \text{o}_P(1)
\]

and

\[
\max_{j \leq p} \left| \frac{\tilde{\text{dcov}}(X_j) - \text{dcov}(X_j)}{\text{dcov}(X_j)} \right| \leq \max_j \left| \frac{\tilde{\var}(X_j)}{\var(X_j)} - 1 \right| (1 + n^{-1}) + \frac{1}{n} = \text{o}_P(1),
\]

whence (iv) is proved.

(v) We first observe that

\[
\left| \frac{n-1}{n} \tilde{\var}(X_j) - \var(X_j) \right| = \left| (\bar{X}_j - \E X_j)(1 - \E X_j - \bar{X}_j) \right|
\leq \left| \bar{X}_j - \E X_j \right| \cdot |1 - 2 \bar{X}_j + \text{o}_P(1)|
\leq (1 + \text{o}_P(1)) \left| \bar{X}_j - \E X_j \right|
\]

uniformly in \( j \leq p \). Applying (i) proved above, we may pick \( c_0 > 0 \) satisfying \( (0.5 + c_0)(1 + c_0) < 0.5 + d_0 \), and obtain

\[
\max_j \left| \frac{\tilde{\var}(X_j) - \var(X_j)}{\var(X_j)^{1/2}} \right| \leq \max_j \frac{1}{n} \frac{\tilde{\var}(X_j)}{\var(X_j)^{1/2}} + (1 + c_0) \max_j \frac{\left| \bar{X}_j - \E X_j \right|}{\var(X_j)^{1/2}}
\leq (1 + 2c_0)b_n
\]

with probability approaching 1.
Then, applying (iv) proved above, we obtain

\[
|\hat{\text{var}}(X_j)^{1/2} - \text{var}(X_j)^{1/2}| \leq \frac{|\text{var}(X_j) - \text{var}(X_j)|}{\hat{\text{var}}(X_j)^{1/2} + \text{var}(X_j)^{1/2}} \\
\leq \frac{(1 + 2c_0)b_n\hat{\text{var}}(X_j)^{1/2}}{\hat{\text{var}}(X_j)^{1/2} + \text{var}(X_j)^{1/2}} \\
\leq \frac{(1 + 2c_0)b_n}{2} (1 + c_0) \leq (0.5 + c_0)(1 + c_0)b_n \\
\leq (0.5 + d_0)b_n.
\]

Recalling that \( \hat{\text{dcov}}(X_j, X_j)^{1/2} = \sqrt{2(\frac{n-1}{n}) \hat{\text{var}}(X_j)^{1/2}} \) by Proposition 4.1, we calculate

\[
\left| \hat{\text{dcov}}(X_j, X_j)^{1/2} - \text{dcov}(X_j, X_j)^{1/2} \right| \leq \frac{\left| (\frac{2(n-1)}{n} \hat{\text{var}}(X_j))^{1/2} - (2\text{var}(X_j))^{1/2} \right|}{\text{dcov}(X_j, X_j)^{1/2}} \\
\leq \sqrt{2} (0.5 + d_0)b_n \frac{1}{\text{dcov}(X_j, X_j)^{1/2}} + \frac{1}{\sqrt{n-1}}.
\]

The same argument also implies

\[
\left| \hat{\text{dcov}}(Y, Y)^{1/2} - \text{dcov}(X_j, X_j)^{1/2} \right| \leq \frac{\left| (\frac{2(n-1)}{n} \hat{\text{var}}(Y))^{1/2} - (2\text{var}(Y))^{1/2} \right|}{\text{dcov}(Y, Y)^{1/2}} \\
\leq \sqrt{2} d_0b_n \frac{1}{\text{dcov}(Y, Y)^{1/2}} + \frac{1}{\sqrt{n-1}},
\]

completing the proof.

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.1.** Under Assumptions 5.1 and 5.2, we have

\[
P\left( \max_j \left| \hat{\text{dcov}}(X_j, Y) - \text{dcov}(X_j, Y) \right| > b_n(2 + d_0) \right) \rightarrow 0
\]

as \( n \rightarrow \infty \).
Proof: For any $c_0 < d_0/6$, define three events

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

$$E_2 = \left\{ \max_j |\mathbb{E}X_j - \bar{X}_j| \leq b_n(0.5 + c_0) \right\},$$

$$E_3 = \{ |\mathbb{E}Y - \bar{Y}| \leq b_n 0.5c_0 \}.$$

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

On events $E_2$ and $E_3$, we have

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

uniformly in $j \leq p$. Because $X_j, Y$ are binary, we have $\max_j \frac{2}{n}|\overline{\text{cov}}(X_j, Y)| \leq \frac{2}{n}$ almost surely. 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

$$|\text{dcov}(X_j, Y) - \widehat{\text{dcov}}(X_j, Y)| \leq \frac{2}{n} |\overline{\text{cov}}(X_j, Y)| + 2|\text{cov}(X_j, Y) - \widehat{\text{cov}}(X_j, Y)| \leq \frac{2}{n} + 2b_n(0.5 + c_0) + 2b_n(0.5 + 1.5c_0) \leq b_n(2 + d_0)$$

uniformly in $j \leq p$. The last inequality in (A.10) holds for large $n$ because $1/n = o(b_n)$. Thus, the event

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

is implied by $E_1 \cap E_2 \cap E_3$, and thus holds with probability approaching one. This completes the proof. \qed
Now, consider the threshold

\[ K_n := \frac{b_n(2.5 + d_0)}{\min_j \text{dcov}(X_j)^{1/2}\text{dcov}(Y)^{1/2}}. \]

The following theorem is the analog of Theorem A.1 for distance correlation (rather than covariance).

**Theorem A.2.** Under Assumptions 5.1 and 5.2, we have

\[
\mathbb{P}\left( \max_j \left| \hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| < K_n \right) \to 1
\]

and

\[
\mathbb{P}\left( \min_{j \in A} \text{dcorr}(X_j, Y) \geq 3K_n \right) \to 1
\]

as \( n \to \infty \).

**Proof:** We calculate

\[
\left| \frac{\hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y)}{\text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}} \right| \\
= \left| \frac{\hat{\text{dcorr}}(X_j, Y)}{\text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}} \right| - \left| \frac{\text{dcorr}(X_j, Y)}{\text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}} \right| \\
= \left| \frac{\hat{\text{dcorr}}(X_j, Y) \text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}}{\text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}} \right| - \left| \frac{\text{dcorr}(X_j, Y) \text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}}{\text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}} \right|. \tag{A.10}
\]

Factoring out the common denominator in (A.10), the numerator is bounded by

\[
\left| \hat{\text{dcorr}}(X_j, Y) \text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2} - \text{dcorr}(X_j, Y) \text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2} \right| \\
\leq \left| \hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| \text{dcov}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2} \\
+ \left| \text{dcorr}(X_j, Y) \right| \text{dcov}(X_j, X_j)^{1/2} - \hat{\text{dcorr}}(X_j, X_j)^{1/2}\text{dcov}(Y, Y)^{1/2}
\]
\[+ \hat{\text{dcov}}(X_j, X_j)^{1/2} \hat{\text{dcov}}(X_j, Y) \hat{\text{dcov}}(Y, Y)^{1/2} - \text{dcov}(Y, Y)^{1/2}] (A.11)\]
\[
\leq \left| \hat{\text{dcov}}(X_j, Y) - \text{dcov}(X_j, Y) \right| \text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y, Y)^{1/2} \\
+ (\text{dcov}(X_j, X_j)^{1/2} (\text{dcov}(Y, Y)^{1/2}) \left| \text{dcov}(X_j, X_j)^{1/2} - \hat{\text{dcov}}(X_j, X_j)^{1/2} \right| \text{dcov}(Y, Y)^{1/2} \\
+ \hat{\text{dcov}}(X_j, X_j)^{1/2} (\text{dcov}(X_j, X_j)^{1/2} (\text{dcov}(Y, Y)^{1/2}) \left| \hat{\text{dcov}}(X_j, X_j)^{1/2} - \hat{\text{dcov}}(X_j, X_j)^{1/2} \right| \text{dcov}(Y, Y)^{1/2} \right| (A.12)
\]

where (A.12) follows from (A.11) because \( \text{dcov}(X_j, Y) \leq \sqrt{\text{dcov}(X_j, X_j) \cdot \text{dcov}(Y, Y)}. \)

Hence,
\[
\left| \hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| \leq A_1 + A_2 + A_3,
\]

where, for any \( c_0 > 0 \), we have
\[
A_1 = \frac{\left| \hat{\text{dcov}}(X_j, Y) - \text{dcov}(X_j, Y) \right|}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y, Y)^{1/2}} \leq \frac{b_n(2 + c_0)}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y, Y)^{1/2}}, \tag{A.13}
\]
\[
A_2 = \frac{\left| \text{dcov}(X_j, X_j)^{1/2} - \hat{\text{dcov}}(X_j, X_j)^{1/2} \right|}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y, Y)^{1/2}} \leq \frac{\left( \sqrt{2}(0.5 + c_0)b_n + \frac{1}{\sqrt{n - 1}} \right)(1 + c_0)}{\text{dcov}(X_j, X_j)^{1/2}} \tag{A.14}
\]
\[
A_3 = \frac{\left| \hat{\text{dcov}}(Y, Y)^{1/2} - \text{dcov}(Y, Y)^{1/2} \right|}{\text{dcov}(Y, Y)^{1/2}} \leq \frac{\sqrt{2}c_0b_n}{\text{dcov}(Y, Y)^{1/2}} + \frac{1}{\sqrt{n - 1}} \tag{A.15}
\]

The inequality in (A.13) follows from Theorem A.1. Inequality (A.14) follows from part (iv) of Lemma A.3, whereas (A.15)-(A.16) follow from part (v) of Lemma A.3. Now, with probability approaching 1, we have
\[
\hat{\text{dcov}}(X_j, X_j)^{1/2} \leq \text{dcov}(X_j, X_j)^{1/2} + o_P(1) \leq \sqrt{2} \text{var}(X_j)^{1/2} + o_P(1) \leq \frac{\sqrt{2}}{2} + c_0.
\]
Similarly, \( \hat{\text{dcov}}(Y,Y)^{1/2} \leq \frac{\sqrt{2}}{2} + c_0 \). Recalling that \( n^{-1/2} = o(b_n) \), we conclude that, for any \( c'_0 > 0 \), there is \( c''_0 > 0 \) to satisfy the inequalities

\[
A_2 \leq \sqrt{2(0.5 + c'_0)b_n} \leq \sqrt{2(0.5 + c'_0)b_n \left( \frac{\sqrt{2}}{2} + c'_0 \right)} \leq \frac{(0.5 + c''_0)b_n}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}
\]

and

\[
A_3 \leq \sqrt{2c'_0b_n} \leq \sqrt{2c_0b_n \left( \frac{\sqrt{2}}{2} + c_0 \right)} \leq \frac{c''_0b_n}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}.
\]

Consequently, for any \( d_0 \), we can find \( c'_0, c''_0 > 0 \) so that

\[
\left| \hat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y) \right| \leq \frac{b_n(2 + c'_0 + c''_0 + 0.5 + c''_0)}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}
\]

\[
< \frac{b_n(2.5 + d_0)}{\text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}
\]

\[
\leq \frac{b_n(2.5 + d_0)}{\text{min}_j \text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}
\]

\[
= K_n
\]

uniformly in \( j \leq p \).

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

\[
K_n \leq \frac{b_n(2.5 + d_0)}{\text{min}_j \text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}}
\]

\[
\leq \frac{b_n(2.5 + d_0)}{\text{min}_j \text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}} \max_j \left( \frac{\text{dcov}(X_j, X_j)^{1/2}}{\text{dcov}(X_j, X_j)^{1/2}} \right)
\]

\[
\leq \frac{b_n(2.5 + d_0)}{\text{min}_j \text{dcov}(X_j, X_j)^{1/2} \text{dcov}(Y,Y)^{1/2}} O_P(1),
\]

with probability approaching 1. On the other hand,

\[
\min_{j \in A} \text{dcorr}(X_j, Y) \geq \frac{\min_{j \in A} \text{dcov}(X_j, Y)}{\text{dcov}(Y,Y)^{1/2} \max_j \text{dcov}(X_j, X_j)^{1/2}} \geq \sqrt{2\min_{j \in A} \text{dcov}(X_j, Y)} \geq 3K_n
\]
as long as

$$\min_{j \in A} \text{d} \text{cov}(X_j, Y) \min_j \text{var}(X_j)^{1/2} \gg b_n,$$

completing the proof.

\textbf{Proof of Theorem 5.1:} We can now complete the proof of the main theorem. Note that, by Assumption 5.2, \(\max_{j \notin A} \text{dcorr}(X_j, Y) = 0\) and \(\min_{j \in A} \text{dcorr}(X_j, Y) > 2K_n\). Hence, by Theorem A.2, with probability approaching 1, we have

$$\min_{j \in A} \hat{\text{dcorr}}(X_j, Y) \geq \min_{j \in A} \text{dcorr}(X_j, Y) - K_n > K_n.$$

Consequently,

$$\mathbb{P}\left( \min_{j \in A} \hat{\text{dcorr}}(X_j, Y) > K_n \right) \to 1, \quad (A.17)$$

$$\mathbb{P}\left( \max_{j \notin A} \hat{\text{dcorr}}(X_j, Y) < K_n \right) \to 1. \quad (A.18)$$

For all \(j \notin \hat{A}\), there are two possibilities: either \(\hat{\text{dcorr}}(X_j, Y) \leq K_n\), or there exists \(i\) such that \(j \in D(i)\) and \(\hat{\text{dcorr}}(X_i, Y) \leq K_n\). In the first case, suppose that \(j \in A\). Then, from (A.17), we have, with probability approaching 1, \(\hat{\text{dcorr}}(X_j, Y) > K_n\) which is a contradiction; consequently, it must be the case that \(j \notin A\). In the second case, we can similarly conclude that \(i \notin A\). By the extinction property, if \(i \notin A\), then \(j \notin A\) for all \(j \in D(i)\), implying that \(j \notin A\). Thus, in both cases, we have \(j \notin A\), whence \(A \subseteq \hat{A}\) with probability approaching 1.

On the other hand, consider \(j \in \hat{A}\). If \(j \notin A\), then from (A.18), with probability approaching one, we have \(\hat{\text{dcorr}}(X_j, Y) < K_n\), which contradicts the assumption that \(j \in \hat{A}\). Therefore, \(j \in A\). Combining the results,

$$\lim_{n \to \infty} \mathbb{P}\left( \hat{A} = A \right) = 1.$$
For the second statement in Theorem 5.1, observe that, for any $j \in \hat{A}$, $\mathcal{C}(j)$ 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 $O(\sum_{j \in \hat{A}} |\mathcal{C}(j)|)$. 