ABSTRACT

Title of dissertation: STATISTICAL LEARNING WITH APPLICATIONS IN HIGH DIMENSIONAL DATA AND HEALTH CARE ANALYTICS

Yimei Fan, Doctor of Philosophy, 2017

Dissertation directed by: Professor Ilya Ryzhov
Department of Decision,
Operations and Information Technologies

Statistical learning has been applied in business and health care 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.

While statistical inference has been widely used for decision and policy making in health care, we particularly focused on how providers get paid for some common procedures. We explored a few rich datasets and discovered large variations among providers for how much payers/insurers have paid, aka allowed payment. Then we proposed to incorporate available providers' attributes with regression model to explain the possible reasons for those payment variations.

STATISTICAL LEARNING WITH APPLICATIONS IN HIGH DIMENSIONAL DATA AND HEALTH CARE ANALYTICS

by

Yimei Fan

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

Advisory Committee:
Professor Ilya Ryzhov, Chair/Advisor
Professor Cinzia Cirillo
Professor Guodong Gao
Professor Yuan Liao
Professor Paul Smith

Dedication

I dedicate my dissertation work to my family and many close friends. A special gratitude to my loving parents, Xiang Fan and Jingming Shi who have given me lots of support for my PhD years. My uncle Haitao, aunt Helen, cousins Kevin and Emma have given me cares and encouragement.

I would also like to dedicate my work to my colleagues, Xia Li, Ying Han, Xuan Liu, Marie Chau and Huashuai Qu who I had very positive interactions with for the past years. I will always appreciate for their academic advices and help throughout the process.

I dedicate this work and give special thanks to my best friend Yu Lu, Jie Zhang, Matthew Henricks for being my best cheerleaders and my wonderful boyfriend Micah Goldblum who has given me opinions for this dissertation and the oral exam slides.

Acknowledgments

First of all, I am grateful to my advisor, Professor Ryzhov for giving me an invaluable opportunity to work on interesting and challenging research projects for the past years. He has always been available for advice and helping me whenever I had questions. Even when he was out of town on family vacation or for conferences, he is always accessible via emails. He has constantly been encouraging me to brain storm and discover ideas for research questions and mathematical proofs. It has been a great pleasure to work with such an extraordinary individual.

I would also like to thank Professor Liao, who has been giving me a lot of help especially for the mathematical proofs and write-up of the paper we worked together. Even after he left University of Maryland, he still devoted a lot of time to revising our paper and being available on emails.

I also want to give my thank to Professor Gao, who was my research assistantship supervisor last summer for the health care project. He has been very patient and given me a lot of guidance. With his help, I was able to understand the health care system better and quicker. When I got frustrated and disappointed about the datasets we have, he kept me on the right track and helped me understand the significance of our empirical research and encouraged me not to be afraid of imperfect data. I feel honored that I could have the opportunity to work with such an incredible advise.

I would also like to acknowledge help and support from some of the staff members. Alverda MoCoy and Jessica Sadler as my program coordinators have

given me lots of help to keep my PhD and dissertation work schedules on the right track.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Lastly, thank you all.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Distance Correlation	DC
Sure Independence Screening	SIS
Dynamic DC-based algorithm	DDC
Generalized Linear Model	GLM
Forward Selection	FS
Maximal Marginal Likelihood Estimate	MMLE
Streamwise Regression	SR
Least Absolute Shrinkage and Selection Operator	LASSO
Analysis of Variance	ANOVA
Electronic Health Record	EHR
Electronic Medical Records	EMR
Center for Medicare & Medicaid Service	CMS
Coronary Artery Bypass Surgery	CABG
End-Stage Renal Disease	ESRD
Hemodialysis	HD
Self Organizing Map	SOM
Gross Domestic Product	GDP
Ordinary Least Square	OLS
Metropolitan Statistical Areas	MSA
Hirschman-Herfindahl Index	HHI
health maintenance organization	HMO
Current Procedural Terminology	CPT

Chapter 1: Introduction

1.1 Hierarchical High Dimensional Marketing Data

In business and marketing analytics, there is a class of problems 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 like a tree structure, and is widely applicable in revenue management, marketing and other business applications. Consider the following examples:

1. *Customer demand modeling.* A retailer faces the problem of pricing a wide variety of products as well as marketing campaign. Customer response varies widely depending on the attributes of a given product. When modeling customer response as a function of the price, the retailer may include dummy variables that classify products by department (e.g., tools, electronics, clothes, jewelry and accessory, etc.), then describes different categories of products within a given department (e.g., hammers, saws, drills under tools; cellphones,

TVs, audio under electronics, etc.), and finally adds features at the individual product level. This allows for considerable flexibility in modeling response curves: for example, some individual brands of hammers may not significantly impact customer response, but hammers overall may behave quite differently from other tools.

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, county, 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. Donors' willingness to participate likely varies by geographic location. Thus, the non-profit may model donor's location at the state level, as well as at the level of three- and five-digit zip codes. Model selection allows us to capture large regions that behave similarly, as well as narrow in on more specific locations (five-digit zip codes) that significantly stand out.

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 not 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.

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 $n > p$, or where both n and p are very 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 with an efficient algorithm mitigates this difficulty and leads to more easily interpretable forecasts. My work contributes to feature dimension reduction as well as reducing computation cost.

1.2 Model Selection Algorithm with Extinction Property

In this thesis, I will propose a new model selection algorithm, which exploits the hierarchical structure of the data in our motivating applications to improve computational efficiency and avoids exploring the entire feature space. The method is based on a unique property of the hierarchy: each binary feature has a single "parent" feature (e.g., "tool" is the parent of "hammers"), and any feature can be a member of the set which consists of significant variables only if its parent is also in the same set. 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. I will give a mathematical formal definition for this property in the next chapter.

Because our motivating applications use binary data, I will adopt the distance correlation (DC) criterion of Székely et al. (2007) to test the significance of a particular candidate feature. The DC criterion is valid under very general assumptions (see Li, Zhong & Zhu 2012 for a discussion) that are easily satisfied in the binary-data setting. In the process, however, I will 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 a conceptual bridge between these two notions of correlation. Thus, while the framework can potentially be generalized, it offers especially practical advantages in the binary setting.

By incorporating this criterion into a new dynamic selection procedure that explores the hierarchy from higher (more aggregated) to lower (more disaggregated) 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. Such behavior is likely to arise in the finite-sample cases, and makes it difficult for the benchmark methods to achieve the extinction property. In the next chapter I will 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 simulation 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.

In the following, in order to place my work in the context of the vast literature on variable and model selection, I will give a broad review of current methodology.

1.3 Sure Independence Screening

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, it is worth briefly pointing out that this methodology is seeing increased use in business analytics and operations research applications (Rudin et al. 2012, 2014, Bertsimas et al. 2016, Ryzhov et al. 2016, Li et al. 2017). Especially, variable selection plays an important role for high dimensional statistical modeling and learning. When the dimension of the feature space is extremely high, the remarkable results of well-known Lasso, SCAD, Dantzig selector (Tibshirani 1996, Fan & Li 2001, Candes & Tao 2007) and other popular methods will be challenged. Sure Independence Screening (SIS) (Fan & Lv 2008) offers a model free way to reduce dimension of the ultrahigh feature space (say, $\exp\{O(n^{\xi})\}$ for some $\xi > 0$) to a moderate scale that is below the sample size ($o(n)$). In the general asymptotic framework, SIS is shown to have the sure screening property. The general idea of how SIS works is based on correlation learning, which screens out the irrelevant features such that the selected model size is above a specified threshold. In the following, I will formalize the algorithm. Let $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ be the response variable, let $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T$ be an $n \times p$ matrix with i.i.d predictors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$. The

problem context is to estimate a p-vector of parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ in the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2..., \varepsilon_n)^T$ is an n-vector of i.i.d random variable. For SIS, usually we first standardize $\mathbf{X}$, and then we calculate a p-vector ω by componentwise regression, that is,

$$\omega = \mathbf{X}^T\mathbf{y}. \tag{1.2}$$

In fact, ω is basically the marginal correlation between predictors and response variable scaled by the standard deviation of the response variable. Then by sorting ω in descending order, for any given $\gamma \in (0, 1)$, we define a sub-model $\mathcal{M}_\gamma$ as follows:

$$\mathcal{M}_\gamma = \{1 \leq i \leq p : |\omega_i| \text{ is among the first } [\gamma n] \text{ largest of all components of } \omega\}, \tag{1.3}$$

where $[\gamma n]$ is the integer part of γn. SIS is a straightforward way to shrink the full model from p dimension to a sub-model $\mathcal{M}_\gamma$ with size $d = [\gamma n] < n$. Also it is noticeable that the ranking of ω is invariant under scaling, so implementing SIS is identical to selecting predictors with top $[\gamma n]$ absolute value of Pearson correlation coefficient with dependent variable. Moreover each predictor is used independently for deciding whether if it should be included in a further model. Fan & Lv (2008) has showed the sure screening property holds for SIS, that is,

$$\mathbf{P}(\mathcal{M}_* \subset \mathcal{M}_\gamma) \to 1 \text{ as } n \to \infty \tag{1.4}$$

for some given γ, where $\mathcal{M}_*$ is the true model. The theorem which gives the full asymptotic behavior regarding the accuracy of SIS as well as how the sequence $\gamma = \gamma_n$ should be chosen under four assumptions is stated and proved in Fan & Lv (2008). In applications, the choice of d can be conservative. For instance, Fan & Lv (2008) chose $d = n - 1$ or $d = n/\log(n)$ for numerical studies.

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, Peng, Zhang & Zhu 2012). Only the marginal correlation for each feature is typically considered, though Fan et al. (2009) and Barut et al. (2016) have investigated more complex forms of dependence.

Subsequently, SIS has been extended to nonparametric models (Fan et al. 2011), survival models (Zhao & Li 2012), model-free settings (Zhu et al. 2011) and GLMs (Fan & Song 2010), which I am going to describe in the following section.

Outside the linear regression setting, SIS becomes more computationally intensive; for example, in GLMs, Fan & Song (2010) propose 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 theoretical treatment). In Li, Zhong & Zhu (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). Huo & Székely (2016) extended the applicability of DC further by developing more efficient estimation procedures. I also adopt DC as the screening criterion for my procedure; in addition to its generality, it turns out to admit substantial computational simplifications in our motivating setting of binary data.

1.4 Generalized Sure Independence Screening

In SIS, the context is the linear model with independent Gaussian predictors and errors, Fan & Song (2010) has proposed a more general independent screening procedure with maximal marginal likelihood estimates (MMLE) in generalized linear models. The generalized linear model has random response variable Y from an exponential family and the density function has the canonical form

$$f_Y(y;\theta) = \exp\{y\theta - b(\theta) + c(y)\}. \tag{1.5}$$

To summarize MMLE, Fan & Song (2010) used $\mathcal{M}_* = \{1 \leq j \leq p_n : \beta_j^* \neq 0\}$ for the true sparse model and the dimension of the feature space is p_n. The size of the true model is $s_n = |\mathcal{M}_*|$ and the parameters are denoted as $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, ..., \beta_{p_n}^*)$. The MMLE $\hat{\beta}_j^M$ is the minimizer for componentwise regression, that is,

$$\hat{\boldsymbol{\beta}}_j^M = \left(\hat{\beta}_{j,0}^M, \hat{\beta}_j^M\right) = \text{argmin}_{\beta_0,\beta_j} \mathbf{P}_n l\left(\beta_0 + \beta_j x_j, Y\right) \text{ for } j = 1, ..., p_n. \tag{1.6}$$

where $l(\beta_0 + \beta_1 X, Y) = -(y\theta - b(\theta) + c(y))$ and $\mathbf{P}_n$ is the empirical measure taking average of any objective function $f(X, Y)$ over the full sample. That is, $\mathbf{P}_n f(X, Y) = \frac{1}{n}\sum_{i=1}^{n} f(x_i, y_i)$. Therefore $\hat{\boldsymbol{\beta}}_j^M$ $\mathbf{P}_n l\left(\beta_0 + \beta_j x_j, Y\right)$ is the empirical estimate of

$$\boldsymbol{\beta}_j^M = \left(\beta_{j,0}^M, \beta_j^M\right) = \text{argmin}_{\beta_0,\beta_j} \mathbf{E} l\left(\beta_0 + \beta_j x_j, Y\right) \text{ for } j = 1, ..., p_n. \tag{1.7}$$

We rank all the $\hat{\beta}_j^M$ and obtain a set of variables x_j's such that $x_j \in \hat{\mathcal{M}}_{\gamma_n}$, where

$$\hat{\mathcal{M}}_{\gamma_n} = \{x_j : |\hat{\beta}_j^M| \geq \gamma_n, 1 \leq j \leq p_n\}. \tag{1.8}$$

For logistic regression, the MMLE will be the following

$$\hat{\boldsymbol{\beta}}_j^M = \left(\hat{\beta}_{j,0}^M, \hat{\beta}_j^M\right) = \text{argmin}_{\beta_0,\beta_j} \frac{1}{n}\left[\sum_{i=1}^{n} \log\left(1 + \exp^{(\beta_0 + x_{ij}\beta_j)}\right) - y_i\left(\beta_0 + x_{ij}\beta_j\right)\right]. \tag{1.9}$$

In a similar variable screening problem for generalized linear models, Fan & Song (2010) implemented a screening procedure by sorting marginal likelihood ratios. The formulation of the procedure is as follows. Using the same notation as for MMLE, let

$$L_{j,n} = \mathbf{P}_n\left[l\left(\hat{\beta}_0^M, Y\right) - l\left(\mathbf{X}_j^T \hat{\boldsymbol{\beta}}_j^M, Y\right)\right], j = 1, ..., p_n, \tag{1.10}$$

where $\hat{\boldsymbol{\beta}}_j^M$ is defined by the MMLE procedure, $\mathbf{X}_j = (1, X_j)^T$ and $\hat{\beta}_0^M = \text{argmin}_{\beta_0} \mathbf{P}_n l(\beta_0, Y)$. By sorting all $L_{j,n}$ in descending order, we select a set of variables x_j's, such that

$x_j \in \hat{\mathcal{L}}_{\nu_n}$, where

$$\hat{\mathcal{L}}_{\nu_n} = \{x_j : L_{j,n} \geq \nu_n, 1 \leq j \leq p_n\}. \tag{1.11}$$

Compared with the MMLE screening procedure, the marginal likelihood ratio incorporates the contribution of features to the likelihood increments. Fan & Song (2010) proved the sure screening properties of both algorithms.

Another screening procedure called Streamwise Regression (SR) which is very similar to marginal likelihood ratio screening was proposed by Zhou et al. (2006). Streamwise regression is an online statistical learning procedure which doesn't assume a fixed size for features. Instead it can handle infinite feature size. Feature X_i becomes available at time t_i or at step i after feature X_{i-1} assuming $t_i > t_{i-1}$. Each feature X_i will only be selected if the corresponding t statistic relates to componentwise regression on X_i has p-value below a pre-specified threshold α_i which is updated right before X_i comes in the selecting procedure. The threshold α_i is updated in a dynamic way such that the procedure can control the False Discovery Rate (FDR). The threshold α_i corresponds to the probability of including a spurious feature at time t_i or step i and it is adjusted using the wealth, which is the current acceptable number of future false positives, denote as w_i. w_i will be increased at step i when X_i is selected into the model such that more future false positives are permitted without changing the bound for FDR. On the other hand, w_i will be decreased to save to add future features. The slightly increased threshold α_i will increase the probability of incorrect inclusion of feature (overfitting). The algorithm is described as follows:

Step 1: Initialize $W_0 = 0.5$, $\alpha_\Delta = 0.5$. Let selection set $\mathcal{S} = \emptyset$, $w_1 = W_0$ and $i = 1$.

Step 2: **While** $CPU.time.used < max.CPU.time$

$$\alpha_i = w_i/2i$$

if p value of component regression on $X_i < \alpha_i$, then

$$\mathcal{S} = \mathcal{S} \cup X_i$$

$$w_{i+1} = w_i + \alpha_\Delta - \alpha_i$$

else

$$w_{i+1} = w_i - \alpha_i$$

end if

$i = i + 1$

end while

1.5 Other Feature Selection Methods

In the classification setting, which will also apply to my real application example in Chapter 3, the simplest way to remove irrelevant features is to evaluate the relevance of each feature separately, as in the SIS method. One of the ways to

measure relevance is to use mutual information between feature X_j and Y, that is,

$$I(X,Y) = \sum_{x_j} \sum_{y} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)} \tag{1.12}$$

As an example, if the feature is binary, then the mutual information between x_j and Y is the following

$$I_j = \sum_{c} p_{jc}\pi_c \log \frac{p_{jc}}{p_j} + (1 - p_{jc})\pi_c \log \frac{1 - p_{jc}}{1 - p_j} \tag{1.13}$$

where $\pi_c = p(y = c)$, $p_{jc} = p(x_j = 1|y = c)$, and $p_j = p(x_j = 1) = \sum_c \pi_c p_{jc}$. In general, mutual information can be regarded as reduction of entropy on the class distribution once feature j is observed.

However, when there are interaction effects in the true model, screening using mutual information will fail. Hao & Zhang (2014) proposed a forward-selection-based screening algorithm with interaction effects in the ultrahigh dimension setting. This work on interaction screening is perhaps the closest to my work 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 my work. In the following I will describe the interaction screening algorithm. The

model setup is a regression model with first and second order feature terms

$$Y_i = \beta_0 + \mathbf{x}_i^T \beta^{(1)} + \mathbf{z}_i^T \beta^{(2)} + \varepsilon_i,\ 1 \leq i \leq n, \tag{1.14}$$

where Y_i is a real-valued response, $\mathbf{x}_i = (X_{i1}, X_{i2}, ..., X_{ip})$ is a p-dimensional vector and the vector $\mathbf{z}_i = (X_{i1}^2, X_{i1}X_{i2}, ..., X_{i1}X_{ip}, X_{i2}^2, X_{i2}X_{i3}, ..., X_{ip}^2)^T$. Before the interaction screening, Hao & Zhang (2014) defined the index sets of linear and order-2 terms as

$$\mathcal{P}_1 = \{1, 2, ..., p\}$$

$$\mathcal{P}_2 = \{(k, l) : l \leq k \leq l \leq p\}$$

and the true model regarding the main effects and order-2 effects is

$$\mathcal{T}_1 = \{j : \beta_j \neq 0, j \in \mathcal{P}_1\},$$

$$\mathcal{T}_2 = \{(j, k) : \beta_{jk} \neq 0, (j, k) \in \mathcal{P}_2\}.$$

The interaction screening algorithm iFORT is based on Forward Selection (FS) with standard BIC and modified BIC for high dimensional data. There are two stages. For the first stage, only main effects are selected by FS, during the second stage, the algorithm uses FS again on all the interaction with respect to all the main effects left at the first stage. The iFORT algorithm with formal notation and steps is described as

Stage 1. Define $\mathcal{C} = \mathcal{P}_1$. Implement FS on $\mathcal{C}$. Generate the solution path $\{\mathcal{S}_t^{(1)}, t =$

$1, 2, ...\}$ and incorporate BIC to get the final selected main effects $\hat{\mathcal{M}} = \{j_1, j_2, ..., j_{t_1}\}$.

Stage 2. Update $\mathcal{C} = \hat{\mathcal{M}} \cup \{(k, l) : k \in \hat{\mathcal{M}} \text{ and } l \in \hat{\mathcal{M}}\}$. Generate the solution path $\{\mathcal{S}^{(2)}_{t_1+t}, t = 1, 2, ...\}$ and incorporate BIC to get the final selected main and second order effects $\mathcal{S}^{(2)}_{t_1+t_2}$. In their work, Hao & Zhang (2014) proved the sure screening of interactions that when certain conditions are satisfied:

$$\mathbf{P}\left(\mathcal{T} \subset \mathcal{S}^{(2)}_{t_1+t_2}\right) \to 1 \text{ as } n \to \infty. \tag{1.15}$$

Hao & Zhang (2014) also proposed another FS based algorithm under marginality principle, which was called iFORM and which only has one stage compared with iFORT. This algorithm is more similar to the algorithm in my work. In detail, iFORM starts with empty candidate selection set. Whenever a main effect is selected and entered into the candidate set, the candidate set will be updated by incorporating all the possible interactions with current main effects in the candidate set. Generally, iFROM doesn't separate the process of selecting main effects from selecting interaction terms as iFORT does, but it still satisfies the sure screening property. For work similar to the interaction screening, 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). However, 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.

1.6 Lasso and Other Related Models for Feature Selection

In statistical learning, lasso (least absolute shrinkage and selection operator) is a regression model being used for feature selection which was introduced by Tibshirani (1996). Unlike ridge regression with least square loss function and l_2-norm regularization, which shrinks the estimators toward zero and includes all the variables will be included in the final model, lasso uses l_1 regularization and does both parameter shrinkage and variable selection by zeroing out some coefficients. The objective of lasso is to solve

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{ \frac{1}{n} \left\| y - \beta_0 - X\boldsymbol{\beta} \right\|_2^2 \right\} \text{ subject to } \|\boldsymbol{\beta}\|_1 \leq t, \tag{1.16}$$

equivalent to

$$\min_{\beta_0,\boldsymbol{\beta}} \left\{ \frac{1}{n} \left\| \mathbf{Y} - \beta_0 - \mathbf{X}\boldsymbol{\beta} \right\|_2^2 \right\} + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.17}$$

Furthermore, lasso has also been used for GLM, which solves

$$\min_{\beta_0,\boldsymbol{\beta}} \left(\frac{1}{n} Deviance\left(\beta_0, \boldsymbol{\beta}\right) + \lambda \sum_{j=1}^{p} |\beta_j| \right). \tag{1.18}$$

Specifically, for the logistic regression problem that we used in the application as a benchmark method, with two classes $y = \{0, 1\}$, the formula is the following:

$$\text{argmin}_{\beta_0,\boldsymbol{\beta}} \sum_{i=1}^{n} \left[\log\left(1 + \exp^{(\beta_0 + \mathbf{X}_i\boldsymbol{\beta})}\right) - y_i\left(\beta_0 + \mathbf{x}_i\boldsymbol{\beta}\right) \right] + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{1.19}$$

The main difference between lasso and ridge regression as mentioned above is the different regularization terms. Ridge only shrinks the magnitude of the coefficients and lasso imposes sparsity among the coefficients and thus, makes the fitted model more interpretable. Lasso penalizes them more uniformly. In a forecasting setting with a powerful predictor, the predictor's effectiveness is shrunk by the Ridge as compared to the Lasso. Elastic net (Zou & Hastie 2005) was introduced as a compromise between lasso and ridge and therefore the penalty is a mix of l_1 and l_2 norms. The elastic net estimator is the solution of

$$\text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2 \|\boldsymbol{\beta}\|^2 + \lambda_1 \|\boldsymbol{\beta}\|_1. \tag{1.20}$$

A hierarchical lasso method was introduced by Zhou & Zhu (2010) for group variable selection. In some cases, we need to select grouped variables or factors and the most common example is multi-factor analysis of variance (ANOVA). In other cases, like multinomial logistic regression, each feature is associated with C different weights, one per class. For the case just mentioned, if l_1 regularization of the form $\|\boldsymbol{\beta}\| = \sum_j \sum_c |\beta_{jc}|$ is used, it might end up with some elements of $\beta_{j,:}$ being zero and some not. To prevent the problem described here, elastic net could be used to encourage group selection based on its l_2 norm term. Other than that, Yuan & Lin (2006) proposed a group lasso which can also be used to handle the problem. Consider the regression problem with C groups:

$$\mathbf{Y} = \sum_{j=1}^{C} \mathbf{X}_j \boldsymbol{\beta}_j + \epsilon \tag{1.21}$$

where $\mathbf{Y}$ is an $n \times 1$ vector, $\epsilon \sim N_n(0, \sigma^2 I)$, $\mathbf{X}_j$ is a $n \times p_j$ matrix for the jth group, $\boldsymbol{\beta}_j$ is a coefficient vector of size p_j, $j = 1, ..., C$. Then given positive definite matrices $K_1, ..., K_C$ for C groups, the group lasso estimate is to minimize

$$\frac{1}{2}\|\mathbf{Y} - \sum_{j=1}^{C} \mathbf{X}_j \boldsymbol{\beta}_j\|^2 + \lambda \sum_{j=1}^{C} \|\boldsymbol{\beta}_j\|_{K_j}. \tag{1.22}$$

The K norm is defined for a vector $\eta \in R^d$, $d \geq 1$, and a symmetric $d \times d$ positive definite matrix K, by

$$\|\eta\|_K = (\eta^{'} K \eta)^{1/2}.$$

However, for the specific data in my research, I haven't been aware of any prior work in statistics that has specifically considered hierarchical data structures in the context of model selection. Group Lasso (Yuan & Lin 2006) 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). In Zhou & Zhu (2010), the term "hierarchical" refers to the structure of the lasso penalties rather than the data. For the data structure considered in my research, group lasso methods do not lead to the desired extinction property.

1.7 Statistical Learning and Data Mining in Healthcare Analytics

In recent years, the health care sector has undergone a massive shift by digitalizing massive amounts of data about individual patients. A prime driver of this

trend is the adoption of electronic health records (EHR), providing access to individuals medical and treatment history. Since EHRs are real time, patients information is available instantly and securely to authorized users. With EHR data, data mining and statistical learning techniques are used to uncover patterns and build predictive models to benefit patients, payers, providers, and policy makers.

The applications of medical and operational healthcare data are varied. One of the main ways data mining techniques help improve patient care and reduce health care costs is by detecting high-risk or high-cost patients (Bates et al. 2014). There are various ways to access these data. Similar to EHRs, electronic medical records (EMR) have been used to uncover drug-to-drug interactions outside of clinical trials (Tatonetti et al. 2012). In addition to EMRs, insurance claim datasets with more than millions of data points, consist of the billing codes, claim payments, claim codes which specify diagnoses, procedures, and drugs that physicians, pharmacies, hospitals, and other health care providers submit to payers (e.g., insurance companies, Medicare). These datasets have important significance for health care management using operation research and data mining methods. Prediction models are used to forecast medical costs using patients historical claims and diagnoses. Some other advanced data-driven techniques, like clustering, can lead to better understanding of medical practice, as well as patient segmentation according to their health risk, socioeconomic status, etc. The advantage of claim datasets over EMR is that they give a holistic view of how patients interact with the healthcare system, and they present a better overview of the interactions among payers, providers, and patients. There are other types of data which contain customers enrollment records,

with demographic as well as socioeconomic attributes, that may be used to study customers behavior in the healthcare system. For example, discrete choice model describes, explains, and predicts customers choice of health plans, and quantifies price elasticity regarding out-of-pocket premiums, deductible amounts, or copayments. Moreover, the classification model is used to predict customers switching behavior for marketing and consumer management.

In the following, I will give a brief literature review with respect to how statistical learning and data mining techniques are applied in health care.

1.7.1 Classification

Classification is one of the essential parts of supervised learning. The most commonly used classification algorithm is logistic regression which gives not only the class label as well as likelihood of each class for the prediction outcomes. In healthcare, logistic regression is applied to predict the likelihood a patient will develop a specific infection or other complication, or the likelihood a patient will be readmitted within 30 days of discharge. In fact, readmission is a serious issue for hospitals, organizations like Center for Medicare & Medicaid Service (CMS), and other payers. CMS has been addressing and standardizing the readmission measures to correctly reflect patient risk as well as population health; these readmission metrics serve as a measure for quality of care. Hospitals readmission reduction programs are using data analytic tools to find causes or risk factors for readmission and help indicate solutions for reducing readmission rate. Price et al. (2013) has

found post-coronary artery bypass surgery (CABG) readmissions may be reduced through careful postoperative surveillance for readmission risk factors (eg, abnormal serum creatinine or unplanned reoperations) and/or for frequent causes of readmission (eg, pleural effusions). From payers perspectives, classification algorithms are used for detecting churning behavior of health plan enrollees and discovering important attributes which affect churning behavior. Boonen et al. (2016) determined that switching behavior depends on health plan price, quality, and demographic and socioeconomic variables such as age, health, education, and supplementary or group insurance coverage. Young people are more sensitive to price, whereas older people are more sensitive to quality. While searching for health plan information, sensitivity toward price has a larger impact overall than service quality. In addition, switching propensity is affected by educational level. Other similar research regarding switching in the Medicaid program (Buchmueller et al. 2005) and Federal Employee Health Benefit Program (Atherly et al. 2005). Buchmueller et al. (2005) found families and individuals who make active choices upon entering the Medicaid program are at substantially lower risk of disenrollment than those who are auto-assigned. Interactions between enrollee ethnicity and provider language proficiency suggest enrollees satisfaction depends on the cultural competence of providers. Differential disenrollment by health risk status results in adverse selection for certain types of plans. Atherly et al. (2005) found individuals switch out of plans with premium increases and benefit decreases relative to other plans in the market. Switching is negatively associated with age, and individuals in preferred provider organizations are less likely to switch, but are more responsive to premium increases than those

in the managed-care sector.

1.7.2 Regression

In their 2015 CMS report, U.S. health care spending grew 5.3 percent in 2014, reaching \$3.0 trillion or $\$9,523$ per person. As a share of the nation's Gross Domestic Product (GDP), health spending accounted for 17.5 percent. It is essential to develop good prediction models and analyze the determinants for efficient health care utilization and cost. For prediction models, the classical linear regression model is not suitable to handle skewed and heavy-tailed outcomes (cost or utilization) in healthcare data. One of the widely used alternatives is to log-transform original dependent variables ($\ln(y)$). In Manning & Mullahy (2001), Their work compared several different regression models, including ordinary least square (OLS) on $\ln(y)$ and other generalized linear models. They conclude that no single model is best under all conditions (skewed data, heteroscedasticity, heavy tail, etc.). Buntin & Zaslavsky (2004) included a two-part model whose first part is to predict the probability of healthcare utilization, on top of those methods mentioned in Manning & Mullahy (2001) for Medicare cost. Their work finds all the models produced similar results. Cantoni & Ronchetti (2006) proposed a robust approach which is an extension of maximum likelihood techniques and showed the approach has less noise and excellent efficiency properties even with some deviation from underlying distribution of data.

On top of prediction models, detecting driven factors efficiently reduces health

costs and helps to understand the cost variation better. According to a report published on New Yorker written by Atul Gawande, health care cost varies across regions by three hundred percent or more. In my work, I focused on payment, particularly, allowed payment variation among physicians for some common services with private payers. CMS has a comprehensive payment model which explains almost fully the reimbursement variation in Medicare (Barnes et al. 2016) while with private payers it is different. Payment among private payers depends on the market power of provider and payer. We hypothesize the correlation between price and providers market power is positive and the correlation between price and payers market power is negative. In Welch et al. (1993), studied the variation of allowed charges in 317 Metropolitan Statistical Areas (MSA) within Medicare and found areas with high inpatient admission rates tend to have high inpatient expenditure, while areas with high inpatient expenditure have high outpatient expenditure. Barnes et al. (2016) discovered the charged amount variation among providers is present not only across different regions of United States but also among providers in the same community. Baker, Bundorf, Royalty & Levin (2014) presented the noticeable physician payment variation for five common services across and within MSA and county level and studied the relationship between variation and physician practice competition measured by Hirschman-Herfindahl Index (HHI).

1.7.3 Discrete Choice Model

Discrete choice models theoretically or empirically predict choices made by people among a finite set of options, which are used to examine and understand customers behavior in marketing and econometrics. The model statistically relates the choice made by the individual to the attributes of the individual and available alternatives. In healthcare, consumer choice of health insurance is a critical issue for providing efficient healthcare delivery. In practice, adequately rational self-selection into the/a health insurance market is important for efficiency of market competition. In academics, studying consumer choice with underlying individual characteristics, plans characteristics, risk preference, and other related factors, helps researchers to recover certain parameters that might be intrinsic interests for welfare analysis and plan design. Specifically, recent academic attention has been emphasized that price and quality are the two most important factors for health plan choices (Kolstad & Chernew 2009). Some studies have focused on how quality information (Beaulieu 2002), plan rating (Jin & Sorensen 2006), or report cards (Wedig & Tai-Seale 2002) affect consumers choice. These studies compared consumers choice before and after the availability of quality information. Others studied switching costs and price elasticity of out-of-pocket premiums related to consumers plan choice (Strombom et al. 2002). By decomposing and quantifying switching cost with respect to some time-varying effects and other financial characteristics, cost of switching is significantly large (Handel 2013). Switching cost has been studied in other industries, like electronic brokerage market, where it also plays a critical role for switching be-

havior; it varies among different brokerage firms, which implied the importance for firms regulation for retention of consumers (Chen & Hitt 2002). On the other side, there have been studies of consumers choice based on adverse selection. With the existence of adverse selection, the market is not totally efficient because of asymmetric information. A long term academic preoccupation has been to try to prove asymmetric information experimentally or empirically (Einav & Finkelstein 2011).

1.7.4 Unsupervised Learning Algorithm

Unsupervised learning techniques are becoming increasingly popular in healthcare. Particularly, clustering algorithms are used to discover similar patterns among high-utilizing patients or high-cost patients. Not only can we locate those patients if they fall into particular clusters, the characteristics of these separate clusters can also be learned. In Alsayat & El-Sayed (2016), the work proposed an efficient two stage clustering algorithm based on K-means algorithm with Self Organizing Map (SOM) which is efficient due to its unsupervised learning and topology- preserving properties. Liao et al. (2016) applied K-means and hierarchical clustering to identify cost change patterns of patients with end-stage renal disease (ESRD) who initiated hemodialysis (HD) and found the K-means clustering algorithm appeared to be the most appropriate in healthcare claims data with highly skewed cost information. In Marlin et al. (2012), it applied a probabilistic clustering model to identify temporal patterns from the physiologic time series data contained in EHRs, which is designed to mitigate the effects of temporal sparsity inherent in EHR records data.

1.8 Outline of Thesis

In Chapter 2, I will present the Dynamic Distance Correlation (DDC) algorithm for feature screening in hierarchically structured marketing data. The DDC procedure respects the extinction property for any finite sample size, which cannot be guaranteed by other procedures. I will first describe the hierarchical data and model. Then 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 GLMs and also leads to a significant computational speedup. After that, I will prove that the set of features selected by the DDC method is asymptotically equal to the true model. The proof will be provided as a separate subsection.

In Chapter 3, I will show 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. In this way, our contributions span both statistics and operations research, and are particularly applicable in business analytics and marketing.

In Chapter 4, I will present the data analysis results for physician's payment variation and propose possible insights and further work for statistical learning and analysis.

At last, conclusion will be followed after Chapter 5.

Chapter 2: Dynamic Distance Correlation Procedure

2.1 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 $\mathbf{X}$ is 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, 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.

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.

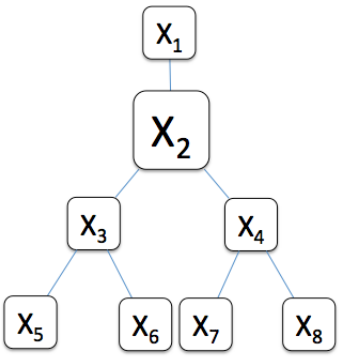


Figure 2.1: Illustration of a hierarchical data structure.

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 $\mathcal{C}(j)$ as 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 $\mathcal{D}(j)$ as 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 2.1, 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 2.1.1 (extinction property)**.** *If $j \in \mathcal{A}^c$, then $k \in \mathcal{A}^c$ for all $k \in \mathcal{D}(j)$.*

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 $F(1 \mid \mathbf{X})$ 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 "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.

2.2 Methodology

We now describe our new dynamic screening algorithm for identifying features in $\mathcal{A}$. First, Section 2.2.1 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 2.2.2 formally states the dynamic algorithm, while Section 2.2.3 provides a descriptive example illustrating how the procedure exploits the hierarchical structure.

2.2.1 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

$$\operatorname{dcov}(X,Y) = \left(\int \left| \phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t) \right|^2 \left(\pi^2 s^2 t^2\right)^{-1} ds\, dt \right)^{\frac{1}{2}}. \tag{2.1}$$

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 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 estimators

$$\widehat{\operatorname{dcov}}(X,Y) = \left(\widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3 \right)^{\frac{1}{2}}, \tag{2.2}$$

$$\widehat{\operatorname{dcorr}}(X,Y) = \frac{\widehat{\operatorname{dcov}}(X,Y)}{\sqrt{\widehat{\operatorname{dcov}}(X,X)\,\widehat{\operatorname{dcov}}(Y,Y)}}, \tag{2.3}$$

where

$$\begin{aligned}
\widehat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|\cdot|Y_i - Y_j| \\
\widehat{S}_2 &= \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|X_i - X_j|\right)\cdot\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|Y_i - Y_j|\right) \\
\widehat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}|X_i - X_l|\cdot|Y_j - Y_l|\,.
\end{aligned}$$

When both X and Y are binary, however, we find that (2.1) is equivalent to the absolute value of their Pearson correlation. Perhaps more surprisingly, (2.3) is almost surely equivalent to the *sample* Pearson correlation

$$\hat{r} = \frac{\sum_i X_i Y_i - n\bar{X}\bar{Y}}{(n-1)s_x s_y}, \tag{2.4}$$

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 2.2.1. *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) $\operatorname{dcov}(X,Y) = 2\,|\operatorname{cov}(X,Y)|, \qquad \operatorname{dcorr}(X,Y) = |\operatorname{corr}(X,Y)|;$

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

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

The practical implications of Proposition 2.2.1 are twofold. First, with binary data, this result justifies the use of Pearson correlation outside linear regression

(as the validity of DC holds under much more general assumptions). Second, the computation of DC is greatly simplified as (2.4) can be calculated more efficiently than (2.2)-(2.3).

2.2.2 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 such as time and memory when the size of $\mathcal{A}$ is small relative to p.

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.

First, define the cutoff

$$K_n = (\log \log n) \cdot \sqrt{\frac{\log(p \vee n)}{n}}, \tag{2.5}$$

where $p \vee n := \max\{p, n\}$. As will become clear in Section 2.3, the cutoff value is chosen to be slightly larger than the maximum estimation error of the distance correlations $\max_{j \leq p} \left| \widehat{\text{dcorr}}\left(X_j, Y\right) - \text{dcorr}\left(X_j, Y\right) \right|$.

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 $\widehat{\text{dcorr}}\left(X_j, Y\right)$ and set $\mathcal{M}_\ell = \mathcal{M}_\ell \backslash \{j\}$ if $\widehat{\text{dcorr}}\left(X_j, Y\right) < K_n$.

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

Step 4 (selection) Find

$$j_\ell = \arg\max_{j \in \mathcal{M}_\ell} \widehat{\text{dcorr}}\left(X_j, Y\right), \tag{2.6}$$

and update

$$\begin{aligned} \mathcal{S}_{\ell+1} &= \mathcal{S}_\ell \cup \{j_\ell\}, \\ \mathcal{M}_{\ell+1} &= \left(\mathcal{M}_\ell \backslash \{j_\ell\}\right) \cup \mathcal{C}\left(j_\ell\right), \end{aligned}$$

where $\mathcal{C}\left(j_\ell\right)$ 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 (2.6), 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*

$$\widehat{\mathcal{B}} = \{j \in \{1, 2, 3..., p\} \, : \, \widehat{\text{dcorr}}\,(X_j, Y) \geq K_n\},$$

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

Remark 2.2.1. 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 τ based robust correlation (Li, Peng, Zhang & Zhu 2012).

2.2.3 Descriptive Example

To illustrate our algorithm, we briefly discuss a descriptive example on a hierarchy with three levels shown in Figure 2.2. As there are two features in the top layer, we initialize $\mathcal{M}_0 = \{1, 2\}$ and $\mathcal{S}_0 = \emptyset$.

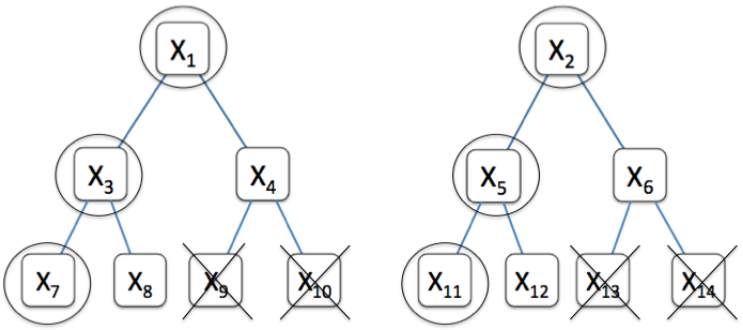


Figure 2.2: Illustration of the DDC algorithm. Due to the extinction property, features 9, 10, 13 and 14 are screened out without being examined directly.

Iteration 1: steps 2-5.

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

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

Iteration 2: steps 2-5.

Suppose $\widehat{\text{dcorr}}(X_3, Y) > \widehat{\text{dcorr}}(X_2, Y) > K_n$, but $\widehat{\text{dcorr}}(X_4, Y) < K_n$. 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 $\widehat{\text{dcorr}}(X_2, Y) > \widehat{\text{dcorr}}(X_7, Y) > K_n$, but $\widehat{\text{dcorr}}(X_8, Y) < K_n$. 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

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

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.

2.3 Theoretical Analysis

The main result of this section is given in Theorem 2.3.1, which shows that the set $\widehat{\mathcal{A}}$ 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 state several regularity conditions. The first assumption simply ensures that we are in the high-dimensional setting, as is standard in the model selection literature.

Assumption 2.3.1. *As $n \to \infty$, the number of features, p, either stays constant or grows with n, satisfying* $(\log\log n)\sqrt{\frac{\log(p\vee n)}{n}} = o(1)$.

With Assumption 2.3.1, we can control dimension of feature space p is less than $\exp(n)$. My work is mostly focused on high dimension data, however, with Assumption 2.3.1, if p is fixed and $n \to \infty$, asymptotically we are in a low dimensional setting.

Assumption 2.3.2. *The following statements hold:*

(i) $\min_{j\in\mathcal{A}} \operatorname{dcov}(X_j, Y) > 2(\log\log n)\sqrt{\frac{\log(p\vee n)}{n}}$;

(ii) $\max_{j\in\mathcal{A}^c} \operatorname{dcov}(X_j, Y) = 0$.

Condition (i) plays an important role in separating the signal from the noise, as it requires relevant features to be sufficiently strongly correlated with the response. The factor $2(\log\log n)$ is determined as a slowly increasing sequence so that we only require signals to be slightly larger than the noise level $\sqrt{\frac{\log p\vee n}{n}}$. In the literature on high-dimensional variable selection, it is straightforward to allow irrelevant features

to be "weakly associated" with the outcome, as long as this behavior can be clearly separated from that of the relevant features. In that case, we could replace condition (ii) above with a weaker condition

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

for some fixed $\delta > 0$; similarly, we could also replace $\log \log n$ in Assumptions 2.3.1-2.3.2 by an even slower-growing function of n. As these modifications are not essential, we keep the current form of the assumptions for simplicity.

Assumption 2.3.3. *There exists* $C > 0$ *such that* $\mathrm{var}\left(Y\right) \geq C$ *and* $\mathrm{var}\left(X_j\right) \geq C$ *for all* j*.*

Recall that the threshold value used to screen features is given by (2.5). Under the above regularity conditions, we show that the empirical distance correlation converges in probability to its population counterpart uniformly in $j = 1, ..., p$. These conditions also imply

$$\max_{j \in \mathcal{A}^c} \widehat{\mathrm{dcorr}}\left(X_j, Y\right) = O_{\mathbb{P}}\left(K_n\right),$$

and that $\min_{j \in \mathcal{A}} \widehat{\mathrm{dcorr}}\left(X_j, Y\right)$ 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 consistency result. We use the notation $|A|$ to denote the cardinality of a finite set A.

Theorem 2.3.1. *Under Assumptions 2.1.1, 2.3.1, 2.3.2 and 2.3.3,*

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

In addition, the DDC procedure will calculate $O\left(\sum_{j\in\widehat{\mathcal{A}}} |\mathcal{C}(j)|\right)$ *empirical distance correlations before terminating.*

It is worth pointing out that Theorem 2.3.1 is somewhat stronger than analogous results in the SIS literature, which typically guarantee (Fan & Lv 2008) that $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ w.p. 1 asymptotically. In our case, the hierarchical structure allows us to guarantee equality.

2.4 Proofs

In this section, we give the full proofs of all results that were stated in the previous section. First, I will prove Proposition 2.2.1 and then prove Theorem 2.3.1.

2.4.1 Proof of Proposition 2.2.1

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

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

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

$$\begin{aligned} & B + C \\ = \; & (e^{is} - 1)(e^{it} - 1)(\mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1)\mathbb{P}(Y = 1))(\mathbb{P}(Y = 0) + \mathbb{P}(Y = 1)) \\ = \; & (e^{is} - 1)(e^{it} - 1)(\mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1)\mathbb{P}(Y = 1)) \\ = \; & (e^{is} - 1)(e^{it} - 1)\mathrm{cov}(X, Y). \end{aligned}$$

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 $w(s, t) = (\pi^2 s^2 t^2)^{-1}$. 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

$$A = \int_{\mathbb{R}^2} \left|(e^{is} - 1)(e^{it} - 1)\right|^2 w(s, t) ds\, dt$$

$$= \int_{\mathbb{R}^2} (2 - 2\cos s)\,(2 - 2\cos t)\, w(s,t) ds\, dt$$
$$= 4. \tag{2.7}$$

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 2.2.1.

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

Lemma 2.4.1. *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

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

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

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

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

Lemma 2.4.2. *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

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

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,

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

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

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

Consequently,

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

which completes the proof. □

Combining Lemmas 2.4.1 and 2.4.2, 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 (2.7). The desired result follows.

2.4.2 Proof of Theorem 2.3.1

In the proof below, we shall apply the following Bernstein inequality for bounded data.

Lemma 2.4.3. *Let $X_1, ..., X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for all positive a,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > a\right) \leq \exp\left(-\frac{\frac{1}{2}na^2}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(X_i^2\right) + \frac{1}{3}Ma}\right)$$

Next, we prove an intermediate result bounding the distance between the estimated and population DC.

Theorem 2.4.1. *Under Assumption 2.3.1, and supposing i.i.d. samples of (X_j, Y), we have*

$$\mathbb{P}\left(\max_j |\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)| > 4\sqrt{\frac{3.3\log(p \vee n)}{n}}\right) \to 0, \quad \text{as } n \to \infty,$$

$$\mathbb{P}\left(\max_j |\widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j)| > 4\sqrt{\frac{3.3\log(p \vee n)}{n}}\right) \to 0,$$

and

$$\mathbb{P}\left(|\widehat{\mathrm{dcov}}(Y,Y) - \mathrm{dcov}(Y,Y)| > 4\sqrt{\frac{3.3\log(p\vee n)}{n}}\right) \to 0.$$

Proof: We recall from Proposition 2.2.1 that

$$\mathrm{dcov}(X_j,Y) = 2\left|\mathrm{cov}(X_j,Y)\right|, \quad \widehat{\mathrm{dcov}}(X_j,Y) = \frac{2(n-1)}{n}\left|\widehat{\mathrm{cov}}(X_j,Y)\right|.$$

Define $a_n = \sqrt{\frac{3.3\log(p\vee n)}{n}}$, and events

$$\begin{aligned} E_1 &= \left\{\max_j |\mathbb{E}X_jY - \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i| \le \frac{a_n}{2}\right\}, \\ E_2 &= \left\{\max_j |\mathbb{E}X_j - \bar{X}_j| \le a_n\right\}, \\ E_3 &= \left\{|\mathbb{E}Y - \bar{Y}| \le \sqrt{\log\log(p\vee n)/n}\right\}. \end{aligned}$$

Because X_j, Y are binary, we have $\max_j \frac{2}{n}|\widehat{\mathrm{cov}}(X_j,Y)| \le \frac{4}{n}$ almost surely. On the event $E_1 \cap E_2 \cap E_3$ (that is, when they simultaneously hold), it follows from the triangle inequality that, uniformly for $j \le p$, we have

$$\begin{aligned} &|\,\mathrm{dcov}(X_j,Y) - \widehat{\mathrm{dcov}}(X_j,Y)| \le \frac{2}{n}|\widehat{\mathrm{cov}}(X_j,Y)| + 2|\mathrm{cov}(X_j,Y) - \widehat{\mathrm{cov}}(X_j,Y)| \\ \le\ & \frac{2}{n}|\widehat{\mathrm{cov}}(X_j,Y)| + 2|\mathbb{E}X_jY - \frac{1}{n}\sum_{i=1}^n X_{ij}Y_i| + 2|\mathbb{E}X_j\mathbb{E}Y - \bar{X}_j\bar{Y}| \\ \le\ & \frac{4}{n} + a_n + 2a_n + 2\sqrt{\log\log(p\vee n)/n} \\ \le\ & 4a_n \end{aligned}$$

where the last inequality holds with all large n because $4/n + 2\sqrt{\log\log(p\vee n)/n} =$

$o(a_n)$. Thus the event

$$E = \{\max_j |\operatorname{dcov}(X_j, Y) - \widehat{\operatorname{dcov}}(X_j, Y)| < 4a_n\}$$

is implied by $E_1 \cap E_2 \cap E_3$, whence

$$\mathbb{P}(E) \geq \mathbb{P}(E_1 \cap E_2 \cap E_3) = 1 - \mathbb{P}(E_1^c \cup E_2^c \cup E_3^c) \geq 1 - \sum_{k=1}^{3} \mathbb{P}(E_k^c).$$

We now show that $\mathbb{P}(E_k^c) \to 0$ for $k = 1, 2, 3$, whence it follows that $\mathbb{P}(E) \to 1$.

To show $\mathbb{P}(E_1^c) \to 0$, define $Z_{ij} = X_{ij}Y_i - \mathbb{E}X_jY$. Then, $\mathbb{E}Z_{ij} = 0$ and the random variables Z_{ij} are independent across $i \leq n$; $|Z_{ij}| \leq 2$ since X_{ij} and Y_i are binary. Also,

$$\max_j Var(Z_{ij}) \leq \max_j \mathbb{E}(X_{ij}Y_i) \leq 1.$$

Thus, by Lemma 2.4.3,

$$\begin{aligned}
\mathbb{P}(E_1^c) &= \mathbb{P}(\max_j |\frac{1}{n}\sum_{i=1}^{n} Z_{ij}| > \frac{a_n}{2}) \\
&\leq \mathbb{P}\left(\bigcup_{j \leq p} \{|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}| > \frac{a_n}{2}\}\right) \\
&\leq p \max_j \mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} Z_{ij}| > \frac{a_n}{2}) \\
&\leq p \max_j \exp\left(-\frac{\frac{1}{2}na_n^2}{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(Z_{ij}^2\right) + \frac{2}{3}a_n}\right) \\
&\leq p \exp\left(-\frac{\frac{1}{8}na_n^2}{\max_j \frac{1}{n}\sum_{i=1}^{n} Var\left(Z_{ij}\right) + \frac{1}{3}a_n}\right) \\
&\leq (p \vee n) \exp\left(-\frac{\frac{1}{8}na_n^2}{1 + \frac{1}{3}a_n}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \exp\left(\log(p \vee n) - \frac{\frac{1}{8} n a_n^2}{1.5}\right) && (2.8)\\
&\leq \exp(-0.1 \log(p \vee n)) && (2.9)\\
&\leq n^{-0.1},
\end{aligned}
$$

where (2.8) holds true for all large n because $a_n \to 0$, and (2.9) is due to $a_n = \sqrt{\frac{3.3 \log(p \vee n)}{n}}$. Convergence to zero is obtained from the last line. It follows from the same argument that $\mathbb{P}(E_2^c) \to 0$.

In addition, for $t_n = \sqrt{\log\log(p \vee n)/n}$,

$$
\mathbb{P}(E_3^c) = \mathbb{P}(|\mathbb{E}Y - \bar{Y}| > t_n) \leq \frac{Var(Y_i)}{n t_n^2} \leq \frac{1}{n t_n^2} = \frac{1}{\log\log(p \vee n)} \to 0.
$$

Therefore, $\mathbb{P}(E) = \mathbb{P}(\max_j |\operatorname{dcov}(X_j, Y) - \widehat{\operatorname{dcov}}(X_j, Y)| < 4a_n) \to 1$. Similarly, it can be proved that $\mathbb{P}\left(\max_j |\widehat{\operatorname{dcov}}(X_j, X_j) - \operatorname{dcov}(X_j, X_j)| > 4a_n\right) \to 0$. Finally, the same argument and the proof of $\mathbb{P}(E_3^c) \to 0$ also implies

$$
\mathbb{P}\left(|\widehat{\operatorname{dcov}}(Y, Y) - \operatorname{dcov}(Y, Y)| > 4\sqrt{\frac{3.3 \log(p \vee n)}{n}}\right) \to 0,
$$

completing the proof. □

Next, we prove an analog of Theorem 2.4.1 for the distance correlation rather than the distance covariance.

Theorem 2.4.2. *Under Assumptions 2.3.1 and 2.3.3, there exists $L > 0$ satisfying*

$$
\mathbb{P}\left(\max_j \left|\widehat{\operatorname{dcorr}}(X_j, Y) - \operatorname{dcorr}(X_j, Y)\right| > L\sqrt{\frac{\log(p \vee n)}{n}}\right) \to 0, \quad \textit{as} \quad n \to \infty.
$$

Proof: We first calculate

$$\begin{aligned}
& \max_j |\widehat{\mathrm{dcorr}}(X_j, Y) - \mathrm{dcorr}(X_j, Y)| \\
= & \max_j \left| \frac{\widehat{\mathrm{dcov}}(X_j, Y)}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)}} - \frac{\mathrm{dcov}(X_j, Y)}{\sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}} \right| \\
= & \max_j \left| \frac{\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y) + \mathrm{dcov}(X_j, Y)}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)}} - \frac{\mathrm{dcov}(X_j, Y)}{\sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}} \right| \\
\leq & \max_j d_j^{\mathrm{I}} + \max_j d_j^{\mathrm{II}},
\end{aligned}$$

where

$$\begin{aligned}
d_j^{\mathrm{I}} &= \left| \frac{\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)}} \right|, \\
d_j^{\mathrm{II}} &= \left| \frac{\mathrm{dcov}(X_j, Y)}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)}} - \frac{\mathrm{dcov}(X_j, Y)}{\sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}} \right|.
\end{aligned}$$

We aim to find $b_n = o(1)$ so that

$$\mathbb{P}\left(\max_j \left| \widehat{\mathrm{dcorr}}(X_j, Y) - \mathrm{dcorr}(X_j, Y) \right| > 2b_n \right) \to 0, \quad \text{as} \quad n \to \infty.$$

Observe that

$$\mathbb{P}\left(\max_j |\widehat{\mathrm{dcorr}}(X_j, Y) - \mathrm{dcorr}(X_j, Y)| > 2b_n \right) \leq \mathbb{P}\left(\max_j d_j^{\mathrm{I}} > b_n \right) + \mathbb{P}\left(\max_j d_j^{\mathrm{II}} > b_n \right). \tag{2.10}$$

We will handle the two probabilities on the right-hand side of (2.10) separately and prove they both converge to zero when n is large enough.

To prove the first probability goes to zero, note that Assumption 2.3.1 implies $\sqrt{\log(p)/n} = o(1)$. Then, by Assumption 2.3.3 and Theorem 2.4.2, we have

$$\begin{aligned}
\min_j \widehat{\mathrm{dcov}}(X_j, X_j) &\geq \min_j \mathrm{dcov}(X_j, X_j) - \max_j |\widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j)| \\
&\geq C - o_{\mathbb{P}}(1) \\
&> C/2
\end{aligned}$$

with probability approaching one. Similarly, $\widehat{\mathrm{dcov}}(Y, Y) > C/2$ with probability approaching one.

Now define the event

$$F = \left\{ \max_j \frac{1}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)}} < \frac{2}{C} \right\}.$$

From the preceding arguments, we have $\mathbb{P}(F) \to 1$. Then,

$$\begin{aligned}
\mathbb{P}\left(\max_j d_j^{\mathrm{I}} > b_n \right) &= \mathbb{P}\left(\max_j d_j^{\mathrm{I}} > b_n, F \right) + \mathbb{P}\left(\max_j d_j^{\mathrm{I}} > b_n, F^c \right) \\
&\leq \mathbb{P}\left(\max_j |\widehat{\mathrm{dcov}}(X_j, Y) - \mathrm{dcov}(X_j, Y)| > b_n C/2 \right) \quad (2.11) \\
&+ \mathbb{P}\left(F^c \right). \quad (2.12)
\end{aligned}$$

Let b_n be such that $b_n C/2 \geq 4\sqrt{\frac{3.3\log(p \vee n)}{n}}$. We apply Theorem 2.4.1 so that the first term in (2.12) is $o(1)$. Then,

$$\lim_{n \to \infty} \mathbb{P}\left(\max_j d_j^{\mathrm{I}} > b_n \right) = 0.$$

We now analyze the quantity d_j^{II}. First, we rewrite

$$
\begin{aligned}
d_j^{\mathrm{II}} &= |\mathrm{dcov}(X_j, Y)| \cdot \left| \frac{1}{\left(\sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)} + \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} \right)} \right| \\
&\quad \cdot \Big| \mathrm{dcov}(X_j, X_j) \left(\mathrm{dcov}(Y, Y) - \widehat{\mathrm{dcov}}(Y, Y) \right) \qquad (2.13) \\
&\quad + \widehat{\mathrm{dcov}}(Y, Y) \left(\mathrm{dcov}(X_j, X_j) - \widehat{\mathrm{dcov}}(X_j, X_j) \right) \Big| \\
&\quad \cdot \frac{1}{\sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)\,\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}} \\
&\leq 2 \cdot \left| \frac{1}{\left(\min_j \sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)} + \min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} \right)} \right| \\
&\quad \cdot \Big(\max_j \widehat{\mathrm{dcov}}(Y, Y) \left| \widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j) \right| \qquad (2.14) \\
&\quad + \max_j \mathrm{dcov}(X_j, X_j) \left| \widehat{\mathrm{dcov}}(Y, Y) - \mathrm{dcov}(Y, Y) \right| \Big) \\
&\quad \cdot \frac{1}{\min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)\,\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}} \qquad (2.15)
\end{aligned}
$$

Due to Assumption 2.3.3 and the binary structure of all the variables, we have

$$
\min_j \mathrm{dcov}(X_j, X_j) \geq C, \qquad \mathrm{dcov}(Y, Y) \geq C \qquad (2.16)
$$

$$
\widehat{\mathrm{dcov}}(Y, Y) = \frac{2(n-1)}{n} \widehat{\mathrm{var}}(Y) \leq 2, \qquad \mathrm{dcov}(X_j, X_j) = 2\mathrm{var}(X_j) \leq 2. \qquad (2.17)
$$

Applying (2.16) and (2.17) to (2.15), we have

$$
\begin{aligned}
\max_j d_j^{\mathrm{II}} &\leq 4 \cdot \left| \frac{1}{\left(\min_j \sqrt{\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)} + \min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} \right)} \right| \\
&\quad \cdot \left(\max_j \left| \widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j) \right| + \left| \widehat{\mathrm{dcov}}(Y, Y) - \mathrm{dcov}(Y, Y) \right| \right)
\end{aligned}
$$

$$\cdot \frac{1}{\min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)\,\mathrm{dcov}(X_j, X_j)\,\mathrm{dcov}(Y, Y)}}.$$

Therefore,

$$\begin{aligned}
& \mathbb{P}\left(\max_j d_j^{\mathrm{II}} > b_n\right) \\
= \; & \mathbb{P}\left(\max_j d_j^{\mathrm{II}} > b_n, \min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} > \frac{C}{2}\right) \\
+ \; & \mathbb{P}\left(\max_j d_j^{\mathrm{II}} > b_n, \min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} \le \frac{C}{2}\right) \\
\le \; & \mathbb{P}\left(4\left(\max_j \left|\widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j)\right| + \left|\widehat{\mathrm{dcov}}(Y, Y) - \mathrm{dcov}(Y, Y)\right|\right) > C^3 b_n/8\right) \\
+ \; & \mathbb{P}\left(\min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} \le C/2\right).
\end{aligned}$$

Observe that $\min_j \sqrt{\widehat{\mathrm{dcov}}(X_j, X_j)\widehat{\mathrm{dcov}}(Y, Y)} > \frac{C}{2}$ with probability approaching one. Let b_n be such that $C^3 b_n \ge 136\sqrt{\frac{7\log(p\vee n)}{n}}$, then $C^3 b_n/8 - \sqrt{\frac{7\log(p\vee n)}{n}} \ge 16\sqrt{\frac{7\log(p\vee n)}{n}}$. Hence by Theorem 2.4.1,

$$\begin{aligned}
\mathbb{P}\left(\max_j d_j^{\mathrm{II}} > b_n\right) \; \le \; & \mathbb{P}\left(4\left|\widehat{\mathrm{dcov}}(Y, Y) - \mathrm{dcov}(Y, Y)\right| > C^3 b_n/8 - \sqrt{\frac{7\log(p\vee n)}{n}}\right) \\
& + \mathbb{P}\left(\max_j |\widehat{\mathrm{dcov}}(X_j, X_j) - \mathrm{dcov}(X_j, X_j)| > 4\sqrt{\frac{3.3\log(p\vee n)}{n}}\right) \\
& + o(1) \\
= \; & o(1).
\end{aligned}$$

Combining the results for d_j^{I} and d_j^{II}, we can take

$$b_n \ge \max\{136/C^3, 8/C\}\sqrt{\frac{7\log(p\vee n)}{n}}$$

. Hence there exists $L > 0$ satisfying

$$\mathbb{P}\left(\max_j |\widehat{\text{dcorr}}(X_j, Y) - \text{dcorr}(X_j, Y)| > L\sqrt{\frac{\log(p \vee n)}{n}}\right) \to 0, \quad \text{as} \quad n \to \infty,$$

which completes the proof. □

We can now complete the proof of Theorem 2.3.1. Note that, by Assumptions 2.3.2 and 2.3.3, $\max_{j \notin \mathcal{A}} \text{dcorr}(X_j, Y) = 0$ and $\min_{j \in \mathcal{A}} \text{dcorr}(X_j, Y) > 2K_n$. Hence, with probability approaching one,

$$\min_{j \in \mathcal{A}} \widehat{\text{dcorr}}(X_j, Y) \geq \min_{j \in \mathcal{A}} \text{dcorr}(X_j, Y) - L\sqrt{\frac{\log(p \vee n)}{n}} > 2K_n - L\sqrt{\frac{\log(p \vee n)}{n}} > K_n.$$

The last inequality is true since for sufficiently large n, $\log\log n\sqrt{\log(p \vee n)/n} > C_1\sqrt{\log(p \vee n)/n}$ for any $C_1 > 0$. Hence Theorem 2.4.2 implies

$$\mathbb{P}\left(\min_{j \in \mathcal{A}} \widehat{\text{dcorr}}(X_j, Y) < K_n\right) \to 0, \tag{2.18}$$

$$\mathbb{P}\left(\max_{j \in \mathcal{A}^c} \widehat{\text{dcorr}}(X_j, Y) > L\sqrt{\frac{\log(p \vee n)}{n}}\right) \to 0,. \tag{2.19}$$

For all $j \notin \widehat{\mathcal{A}}$, there are two possibilities: either $\widehat{\text{dcorr}}(X_j, Y) \leq K_n$, or there exists i such that $j \in \mathcal{D}(i)$ and $\widehat{\text{dcorr}}(X_i, Y) \leq K_n$. In the first case, suppose that $j \in \mathcal{A}$. Then, from (2.18), we have, with probability approaching 1, $\widehat{\text{dcorr}}(X_j, Y) > K_n$ which is a contradiction; consequently, it must be the case that $j \notin \mathcal{A}$. In the second case, we can similarly conclude that $i \notin \mathcal{A}$. By the extinction property, if $i \notin \mathcal{A}$, then $j \notin \mathcal{A}$ for all $j \in \mathcal{D}(i)$, implying that $j \notin \mathcal{A}$. Thus, in both cases, we

have $j \notin \mathcal{A}$, whence $\mathcal{A} \subseteq \widehat{\mathcal{A}}$ with probability approaching 1.

On the other hand, consider $j \in \widehat{\mathcal{A}}$. If $j \notin \mathcal{A}$, then from (2.19), with probability approaching one, we have

$$\widehat{\mathrm{dcorr}}(X_j, Y) < L\sqrt{\frac{\log(p \vee n)}{n}},$$

which contradicts with the assumption that $j \in \widehat{\mathcal{A}}$. Therefore, $j \in \mathcal{A}$. Combining the results,

$$\lim_{n\to\infty} \mathbb{P}\left(\widehat{\mathcal{A}} = \mathcal{A}\right) = 1.$$

For the second statement in Theorem 2.3.1, 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 $O(\sum_{j\in\widehat{\mathcal{A}}} |\mathcal{C}(j)|)$.

Chapter 3: Numerical Studies

We assess the performance of the DDC algorithm on both simulated (Section 3.1) and real (Section 3.2) 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.

3.1 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 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 disaggregated levels of the hierarchy. Then, κ_i was used to set the distribution

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

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 3.2 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

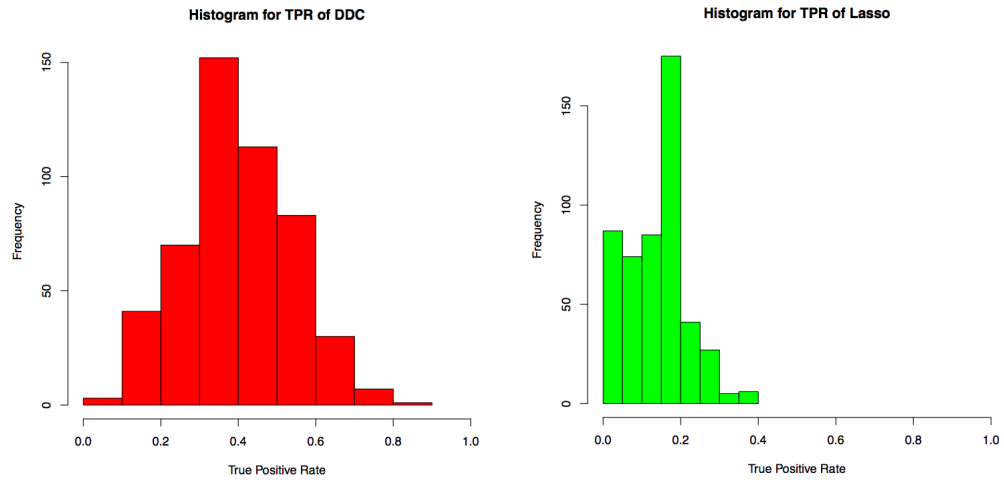
Method	Time (s)	TPR (the higher the better)							FPR (the lower the better)						
		Min	25%	50%	75%	Max	Mean	SD	Min	25%	50%	75%	Max	Mean	SD
DDC	0.16	0.08	0.32	0.4	0.48	0.88	0.41	0.14	0	0.00201	0.00420	0.00676	0.04314	0.00522	0.00444
Lasso	1.34	0.04	0.08	0.16	0.20	0.40	0.14	0.08	0	0.00110	0.00311	0.00585	0.02157	0.00388	0.00353
SR	15.73	0.04	0.16	0.20	0.25	0.44	0.21	0.07	0	0	0.00018	0.00018	0.00073	0.00013	0.00015
SIS	3.37	0.08	0.32	0.36	0.44	0.60	0.37	0.09	0.00110	0.00183	0.00219	0.00238	0.00347	0.00216	0.00040

Table 3.1: 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.

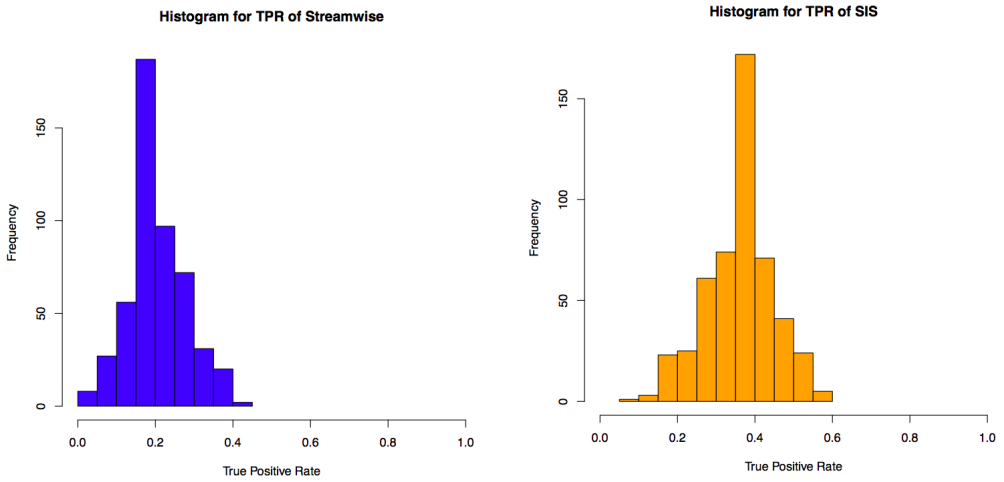
(SR) method performs a univariate (marginal) regression for each individual feature, analogously to the screening approach of Fan & Song (2010). We used univariate logistic regression as the criterion for SR because the response variable is binary. SIS uses the same screening criterion (Pearson correlation) 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 tuning parameter; Fan & Lv (2008) give several suggestions for how to choose it. We experimented with all of them and found that $d = [n/\log n]$ produced the best results.

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 3.1 presents some summary statistics across 500 simulated datasets in the first example, while Figure 3.1 shows the empirical distributions of TPR achieved by the four methods. DDC tends to achieve higher TPR than the other benchmarks,



(a) TPR (DDC). (b) TPR (Lasso).



(c) TPR (SR). (d) TPR (SIS).

Figure 3.1: Histograms showing TPR across 500 simulated datasets (first example).

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. Furthermore, DDC is the most computationally efficient method among the four, running over 8 times faster than Lasso, 21 times faster than SIS, and 98 times faster than SR. This illustrates the practical benefits that can be achieved by exploiting the hierarchical structure of the data.

Table 3.2 and Figure 3.2 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.

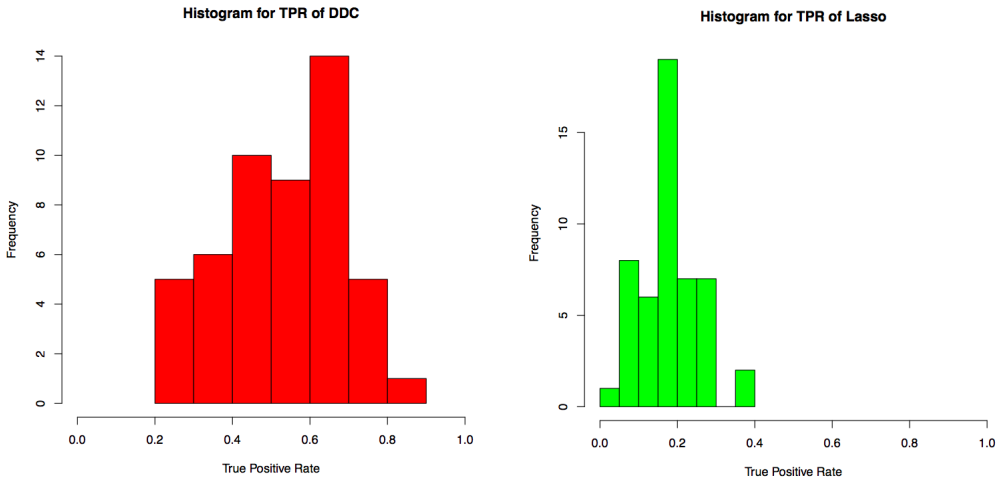
We conclude that, given its computational cost, DDC is highly competitive with the benchmark methods on high-dimensional problems in which the data are structured hierarchically. The computation times suggest that DDC may scale better to problems with massive data sizes; we explore such a setting in greater depth in the following case study.

3.2 Application to B2B Transaction Data

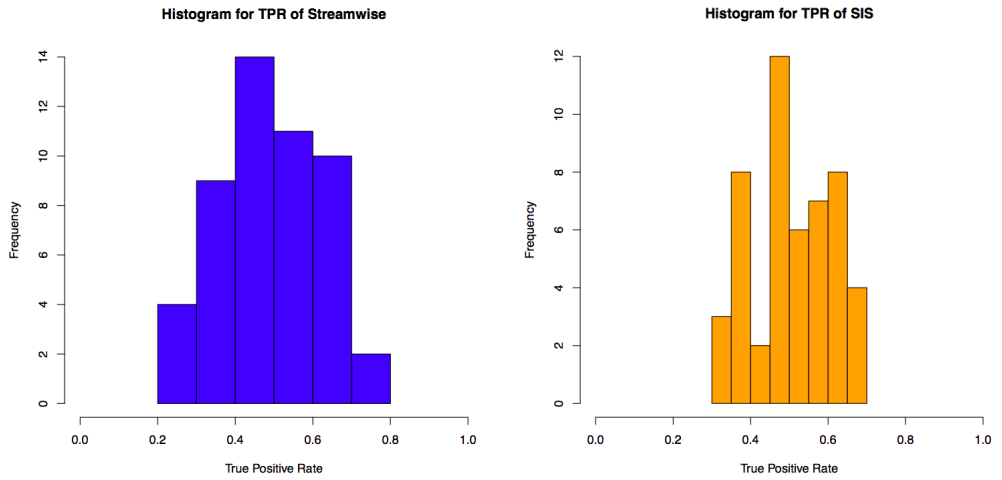
We also implemented our method on two historical datasets provided by Vendavo, Inc., a firm specializing in business-to-business (B2B) pricing science. These data cover a large volume of B2B transactions involving numerous products. The response variable is binary, since the customer in each transaction may either accept or reject the deal. In both datasets, products are aggregated using a hierarchy with

Method	Time (s)	TPR							FPR						
		Min	25%	50%	75%	Max	Mean	SD	Min	25%	50%	75%	Max	Mean	SD
DDC	6.96	0.23	0.47	0.57	0.66	0.83	0.55	0.15	0.00004	0.00075	0.00149	0.00254	0.00987	0.00213	0.00209
Lasso	43.45	0.03	0.13	0.18	0.23	0.37	0.19	0.08	0.00002	0.00030	0.00061	0.00100	0.00200	0.00067	0.00045
SR	4792.83	0.27	0.41	0.50	0.60	0.77	0.51	0.13	0.00006	0.00024	0.00040	0.00055	0.00108	0.00042	0.00023
SIS	876.68	0.30	0.44	0.52	0.60	0.70	0.52	0.10	0.00073	0.00074	0.00076	0.00077	0.00080	0.00076	0.00002

Table 3.2: Performance of model selection methods on second simulated example (50 datasets).



(a) TPR (DDC). (b) TPR (Lasso).



(c) TPR (SR). (d) TPR (SIS).

Figure 3.2: Histograms showing TPR across 50 simulated datasets (second example).

four levels; an individual product ID belongs to a ProductLevel1, ProductLevel2, and ProductLevel3, with an additional disaggregated level where one feature per product is added. The first dataset contains approximately $6,000$ 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 $\mathcal{A}$ 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 $\widehat{\mathcal{A}}$ 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

Method	AUC	Time (seconds)		% features selected				
		Selection	Estimation	Level 1	Level 2	Level 3	Level 4	All
DDC	0.6535	7.53	0.39	81.5%	49.4%	19.5%	4.24%	9.46%
Lasso	0.6408	14.27	0.08	36.2%	17.1%	9.47%	6.25%	7.44%
SR	0.6409	3775.58	0.64	63.8%	17.2%	0.86%	0.16%	1.09%
SIS	0.6330	12.91	0.18	100%	100%	34.1%	0%	11.4%

Table 3.3: Performance of model selection methods on first pricing dataset. All numbers are averaged over 5 folds.

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 3.3 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 3.4 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.

Method	AUC	Time (seconds)		% features selected				
		Selection	Estimation	Level 1	Level 2	Level 3	Level 4	All
DDC	0.6513	131	2	63.8%	12.8%	0.55%	0%	0.15%
Lasso	0.5	867	2181	54.6%	26.5%	10.9%	5.85%	6.43%
SR	0.5753	506021	2327	91.5%	63.1%	22.6%	0.92%	3.21%
SIS	0.5	952	1919	100%	100%	100%	17.3%	24.9%

Table 3.4: Performance of model selection methods on second pricing dataset. All numbers are averaged over 5 folds.

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 speed, and increase predictive power. We note that the benefits of DDC are greater, relative to the benchmark methods, when the dataset is larger.

Chapter 4: Payment Variation in Payer's Reimbursement for Physicians' Services

We did data analysis on the datasets provided by a cloud-based analytics platform company which acquires data from multiple resources . Specifically, the datasets consist of patients, providers, payers and claim (lab, medical and pharmaceutical) records. Our main analysis is on the allowed payment variation for physicians' common services.

4.1 Data

The full dataset contains records pertaining to more than $856,000$ physicians, $375,000$ clinical facilities, 158 million unique patients, and more than 14 billion medical events. We analyzed the sample datasets provided by the company. In the demographic data, we have 1000000 patients with average age around 47, 52% are females. In the enrollment data, there are 414607 records for 95380 health plan enrollees with their demographical attributes available in the demographic data. For our study, we focus on commercial health maintenance organization (HMO) plans. First, in the sample, we have the 88 unique health plans but the majority are commercial type HMO plans. Second, unlike private payers, the government

type of plans have a comprehensive payment model by CMS while the former also depends on market power of providers and payers, as well as the attributes regarding providers' profile, which might give us more insights for the health care cost variation. The only drawback of studying HMO plans is some of the providers are receiving a bundled payment per period of time based on the profile of individual patient. Therefore, some variations of allowed payment among providers for the same procedure might purely because of the capitation. However, since most of our claim data is concentrated in year 2012 and 2013 when capitation was not adopted by most of the payers in United States, particularly in New York State, this drawback won't truly affect our results. In the study, we picked year 2012 and focused on New York State due to the subset has the most available claim data points.

From the enrollment data, we observed there are 1741 patients (1.8%) have switched their health plans, which showed switching is a rare event. There are two possible explanations: one is the inertia of sticking to the current plan is quite high due to the high switching cost, cost of information search, hassle cost due to discontinuing the current treatment or switching primary care physicians, etc; another reason is the asymmetry of information due to the lack of knowledge relates to the rating of quality of primary care physicians or plans. In our datasets, there is no specific information regarding the choice set of each individual in terms of health plans and all health plans are anonymized which doesn't allow us to know the details of the plans coverage, deductibles, copay rate and etc. Therefore, we cannot build model that quantifies the switching cost or other type of inertia that keeps consumer to stay in their original choice.

In our claim data, there are more than 20 millions of claim records for around 90000 patients. The claim data records the allowed payment from payers to the physicians. Based on our observations, the allowed payment varied among plans, geographical regions, physicians for similar services. In our study, we picked five common physician services whose Current Procedural Terminology (CPT) codes in our claim data are at the top five.

4.2 Preliminary Analysis

By aggregating the allowed payment amount at physicians zip-code level we calculated a few statistics which gives us an idea of how large the within zip-code geographical variation is. We picked year 2012 since there are relatively more data points for the procedures compared with the other year's records. We calculated each provider's average allowed payment for the five procedure respectively and the results are showed in the table 4.1. From table 4.1, all five physician services showed variation at the zip code level and the scale of variations is quite different for those services. Particularly, for CPT code 36415 (Venipuncture), the maximum payment amount is around $11861 while the minimum amount is $2. Other than that, the maximum within zip-code standard deviation ranges from $0 to $16770. The standard deviation in some zip-codes is zero which means the payer charges the same amount for every physician within those zip-codes, while the large standard deviation implies the payer charged amount in the zip code is dramatically different for the same service with similar population health and other economical attributes.

Within the same zip code, we controlled the attributes relate to population health, property price and other cost of living. We hypothesize the variation of payment is from multiple attributes relate to providers as well as those from payers. For example, payment could vary among provider's specialties. For this hypothesis, we used K-mean clustering algorithm we used K-mean clustering algorithm to visualize the allowed payment pattern among physicians with different specialties in New York State in 2012.

In Figure 4.1, the service we picked is physician office visit (CPT: 99213) and the two specialties are internal medicine and pediatrics, which are top two specialties in our sample data. The x axis in the figure is the average paid amount from payer in 2012 and the y axis is the standard deviation for the varied payment for each individual physician in the same year. For Figure 4.1(a), the mean allowed payment is more concentrated on the interval $[50, 150]$ compared with Figure 4.1(b). For internal medicine providers, the allowed payment data is better suited with three clusters while for the pediatrics, four clusters is optimal.

The other possible attributes which potentially contribute to the variation are market power of providers and payers. For payers, with larger market share and boarder covered provider's network, they possess better negotiation power with providers. However, with some patients seeing the providers outside the network, we expect to see some of the payment is relatively higher. Form our dataset, there is no specific detail/attribute which indicates if the provider is outside the network. On the other hand, for providers, especially when hospitals and physicians merge into a larger organization, the new organization as a whole gains more market leverage.

CPT Code	Medical Service	Summary Statistics				
		Max Payment	Min Payment	Max SD	Min SD	Number of Zip-codes
99213	Office visit	$176	$49	$160	$0	345
99214	Office visit for established patient	$236	$36	$227	$0	294
97110	Therapeutic procedure	$35578	$34	$100	$0	49
36415	Venipuncture	$11861	$2	$16770	$0	263
80061	Lipid panel	$956	$15	$226	$0	41

Table 4.1: Allowed payment variation summary statistics among zip-codes in 2012.

Baker, Bundorf & Kessler (2014) found that when hospitals employ physicians, their market share increase are associated with higher prices and spending. Robinson & Miller (2014) examined patients in California and found that hospital owned physician organization is associated with 10 to 20 percent higher health expenditures.

4.3 Further Work

The preliminary analysis results suggest allowed payment depends on geographical regions, provider's specialty. Beyond that, we hypothesize the allowed payment strongly depends on market power of both providers and payers. At this point, we are still undergoing the process to acquire more data from the company as well as merge other datasets from other resources. After that, for the future work, in order to build regression model to explain price variation, we first need to calculate adjusted price per year for each provider with the mix of treatments and mix of patients. Then we will regress the individual prices on provider's fixed effects (years of experience, number of languages the provider's facility supports, etc.), patient characteristics (age, gender, race and ethnicity, etc.) and procedure or service fixed effects. To test our hypothesis relates to market power of providers and payers, we need to find valid measures for payer negotiation power and provider's market structure. For provider's market structure, we can compare different measures for

the physician practice competition and corresponding results. For payer negotiation power, we can find measures which incorporates market share, tenure with provider, group size and we expect those measures vary from state to state, or among other geographical regions (county to county, etc.). Then, by combining those measures we have just proposed, we can add some more attributes like quality measures of provider, time and geographical region fixed effects, etc. from other data resources for the regression. On top of that, robustness analysis will be done by comparing baseline model with extended models as well as different measurements for payer negotiation power and provider market power.

Figure 4.1: Cluster visualization of payment variation for different specialties

Chapter 5: 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.

We also had a broad study of how statistical learning techniques and models are used in health care system. We did a preliminary analysis on the datasets provided by a private data analytic company, which showed physicians' allowed payments from private payers varies among geographical regions at zip-code level. The variation of allowed payment also depends on the specialty of physicians for the same type of services. We proposed future work on finding the determinants of payment variation with feature engineering and building proper predictive model for the allowed payment amount.

Bibliography

Alsayat, A. & El-Sayed, H. (2016), Efficient genetic k-means clustering for health care knowledge discovery, *in* '2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)', pp. 45–52.

Atherly, A., Florence, C. & Thorpe, K. (2005), 'Health paln switching among members of the federal employees health benefits program', *Inquiry* **42**(3), 255–265.

Baker, L., Bundorf, M. & Kessler, D. (2014), 'Vertical integration: Hospital ownership of physician practices is associated with higher prices and spending', *Health Affair* **33**(5), 756–763.

Baker, L., Bundorf, M., Royalty, A. & Levin, Z. (2014), 'Physician practice competition and prices paid by private insurers for office visits', *JAMA* **312**(16), 1653–1662.

Barnes, S., Bjarnadóttir, M. & Guo, X. (2016), 'Centers for Medicare and Medicaid services provider characteristics fail to explain billing variability', *Health System* **5**, 109–119.

Barut, E., Fan, J. & Verhasselt, A. (2016), 'Conditional sure independence screening', *Journal of the American Statistical Association* **111**(515), 1266–1277.

Bates, D., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. (2014), 'Big data in health care: Using analytics to identify and manage high-risk and high-cost patients', **33**(7), 1123–1131.

Beaulieu, N. (2002), 'Quality information and consumer health plan choices', *Journal of Health Economics* **21**, 43–63.

Belloni, A. & Chernozhukov, V. (2013), 'Least squares after model selection in high-dimensional sparse models', *Bernoulli* **19**(2), 521–547.

Bertsimas, D., OHair, A., Relyea, S. & Silberholz, J. (2016), 'An analytics approach to designing combination chemotherapy regimens for cancer', *Management Science* **62**(5), 1511–1531.

Bien, J., Taylor, J. & Tibshirani, R. (2013), 'A lasso for hierarchical interactions', *Annals of Statistics* **41**(3), 1111–1141.

Boonen, L., Laske-Aldershof, T. & Schut, F. (2016), 'Switching health insurers: the role of price, quality and consumer information search', *Eur J Health Econ* **17**, 339 – 353.

Bradić, J. (2016), 'Randomized maximum-contrast selection: Subagging for large-scale regression', *Electronic Journal of Statistics* **10**, 121–170.

Buchmueller, T., Gilmer, T. & Harris, K. (2005), 'Health plan disenrollment in a choice-based Medicaid managed care program', *Inquiry* **41**(4), 447–460.

Buntin, M. & Zaslavsky, A. (2004), 'Too much ado about two-part models and transformation? comparing methods of modeling Medicare expenditures', *Journal of Health Economics* **23**, 525–542.

Candes, E. & Tao, T. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', **35**(6), 2313–2351.

Cantoni, E. & Ronchetti, E. (2006), 'A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures', *Journal of Health Economics* **25**, 198–213.

Chen, P.-Y. S. & Hitt, L. M. (2002), 'Measuring switching costs and the determinants of customer retention in internet-enabled businesses: A study of the online brokerage industry', *Information Systems Research* **13**(3), 255–274.

Einav, L. & Finkelstein, A. (2011), 'Selection in insurance markets: theory and empirics in pictures', *Journal of Economic Perspectives* **25**(1), 115–138.

Fan, J., Feng, Y. & Song, R. (2011), 'Nonparametric independence screening in sparse ultra-high-dimensional additive models', *Journal of the American Statistical Association* **106**(494), 544–557.

Fan, J., Han, F. & Liu, H. (2014), 'Challenges of big data analysis', *National Science Review* **1**(2), 293–314.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', **96**(456), 1348–1360.

Fan, J. & Lv, J. (2008), 'Sure independence screening for ultrahigh dimensional feature space', *Journal of the Royal Statistical Society* **B70**(5), 849–911.

Fan, J. & Lv, J. (2010), 'A selective overview of variable selection in high dimensional feature space', *Statistica Sinica* **20**(1), 101–148.

Fan, J., Samworth, R. & Wu, Y. (2009), 'Ultrahigh dimensional feature selection: beyond the linear model', *The Journal of Machine Learning Research* **10**, 2013–2038.

Fan, J. & Song, R. (2010), 'Sure independence screening in generalized linear models with NP-dimensionality', *The Annals of Statistics* **38**(6), 3567–3604.

Handel, B. (2013), 'Adverse selection and inersia in health insurance markets: When nudging hurts', *American Economic Review* **103**(7), 2643–2682.

Hao, N. & Zhang, H. H. (2014), 'Interaction screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **109**(507), 1285–1301.

Huo, X. & Székely, G. J. (2016), 'Fast computing for distance covariance', *Technometrics* **58**(4), 435–447.

Jin, G. & Sorensen, A. (2006), 'Information and consumer choice: the value of publicized health plan ratings', *Journal of Health Economics* **25**, 248–275.

Kleiner, A., Talwalkar, A., Sarkar, P. & Jordan, M. I. (2014), 'A scalable bootstrap for massive data', *Journal of the Royal Statistical Society* **B76**(4), 795–816.

Kolstad, J. & Chernew, M. (2009), 'Quality and consumer decision making in the market for health insurance and health care services', *Med Care Res Rev* **66**, 28S–52S.

Li, G., Peng, H., Zhang, J. & Zhu, L. (2012), 'Robust rank correlation based screening', *The Annals of Statistics* **40**(3), 1846–1877.

Li, J., Netessine, S. & Koulayev, S. (2017), 'Price to compete... with many: How to identify price competition in high dimensional space', *Working paper, INSEAD* .

Li, R., Zhong, W. & Zhu, L. (2012), 'Feature screening via distance correlation learning', *Journal of the American Statistical Association* **107**(499), 1129–1139.

Liao, M., Li, Y., Kianifard, F., Obi, E. & Arcona, S. (2016), 'Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis', *BMC Nephrology* **17**(1), 25.

Manning, W. & Mullahy, J. (2001), 'Estimating log models: to transform or not to transform?', *Journal of Health Economics* **20**, 461–494.

Marlin, B., Kale, D., Khemani, R. & Wetzel, R. (2012), Unsupervised pattern discovery in electronic health care data using probabilistic clustering models, *in* 'Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium', IHI '12, ACM, ACM, pp. 389–398.

Meier, L., Van De Geer, S. & Bühlmann, P. (2008), 'The group lasso for logistic regression', *Journal of the Royal Statistical Society* **B70**(1), 53–71.

Nardi, Y. & Rinaldo, A. (2008), 'On the asymptotic properties of the group lasso estimator for linear models', *Electronic Journal of Statistics* **2**, 605–633.

Price, J. D., Romeiser, J. L., Gnerre, J. M., Shroyer, A. L. W. & Rosengart, T. K. (2013), 'Risk analysis for readmission after coronary artery bypass surgery: Developing a strategy to reduce readmissions', *Journal of the American College of Surgeons* **216**(3), 412 – 419.

Robinson, J. & Miller, K. (2014), 'Total expenditures per patient in hospital-owned and physician-owned physician organizations in california.', *JAMA* **312**(16), 1163–1169.

Roth, V. & Fischer, B. (2008), The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms, *in* 'Proceedings of the 25th International Conference on Machine Learning', pp. 848–855.

Rudin, C., Ertekin, Ş., Passonneau, R., Radeva, A., Tomar, A., Xie, B., Lewis, S., Riddle, M., Pangsrivinij, D. & McCormick, T. (2014), 'Analytics for power grid distribution reliability in New York City', *Interfaces* **44**(4), 364–383.

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. & Wu, L. (2012), 'Machine learning for the New York City power grid', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34**(2), 328–345.

Ryzhov, I. O., Han, B. & Bradić, J. (2016), 'Cultivating disaster donors using data analytics', *Management Science* **62**(3), 849–866.

She, Y. & Jiang, H. (2016), Group regularized estimation under structural hierarchy.

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

Strombom, B., Buchmueller, T. & Feldstein, P. (2002), 'Switching costs, price sensitivity and health plan choice', *Journal of Health Economics* **21**, 89–116.

Székely, G. J. & Rizzo, M. L. (2009), 'Brownian distance covariance', *The Annals of Applied Statistics* **3**(4), 1233–1303.

Székely, G. J. & Rizzo, M. L. (2012), 'On the uniqueness of distance covariance', *Statistics & Probability Letters* **82**(12), 2278–2282.

Székely, G. J. & Rizzo, M. L. (2014), 'Partial distance correlation with methods for dissimilarities', *The Annals of Statistics* **42**(6), 2382–2412.

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007), 'Measuring and testing dependence by correlation of distances', *The Annals of Statistics* **35**(6), 2769–2794.

Tatonetti, N., Ye, P., Daneshjou, R. & Altman, R. (2012), 'Data-driven prediction of drug effects and interactions', **4**, 125–131.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society* **B58**(1), 267–288.

Wedig, G. & Tai-Seale, M. (2002), 'The effect of report cards on consumer choice in the health insurance market', *Journal of Health Economics* **21**, 1031–1048.

Welch, W., Miller, M., Welch, H., Fisher, E. & Wennberg, J. (1993), 'Geographic variation in expenditures for physicians' services in the United States', *New England Journal of Medicine* **328**(9), 621–627.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with group variables', *Journal of the Royal Statistical Society* **B68**(1), 49–67.

Zhao, J. & Leng, C. (2016), 'An analysis of penalized interaction models', *Bernoulli* **22**(3), 1937–1961.

Zhao, S. D. & Li, Y. (2012), 'Principled sure independence screening for Cox models with ultra-high-dimensional covariates', *Journal of Multivariate Analysis* **105**(1), 397–411.

Zhou, J., Foster, D. P., Stine, R. A., Ungar, L. H. & Guyon, I. (2006), 'Streamwise feature selection', *Journal of Machine Learning Research* **7**, 1861–1885.

Zhou, N. & Zhu, J. (2010), 'Group variable selection via a hierarchical lasso and its oracle property', *Statistics and its Interface* **3**(4), 557–574.

Zhu, L.-P., Li, L., Li, R. & Zhu, L.-X. (2011), 'Model-free feature screening for ultrahigh-dimensional data', *Journal of the American Statistical Association* **106**(496), 1464–1475.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', **67**, 301–320.