# A Bayesian Approach for Identifying Multivariate Differences Between Groups

Yuriy Sverchkov[1](✉) and Gregory F. Cooper[2]

[1] University of Wisconsin—Madison, Madison, WI, USA
yuriy@biostat.wisc.edu
[2] University of Pittsburgh, Pittsburgh, PA, USA
gfc@pitt.edu

**Abstract.** We present a novel approach to the problem of detecting multivariate statistical differences across groups of data. The need to compare data in a multivariate manner arises naturally in observational studies, randomized trials, comparative effectiveness research, abnormality and anomaly detection scenarios, and other application areas. In such comparisons, it is of interest to identify statistical differences across the groups being compared. The approach we present in this paper addresses this issue by constructing statistical models that describe the groups being compared and using a decomposable Bayesian Dirichlet score of the models to identify variables that behave statistically differently between the groups. In our evaluation, the new method performed significantly better than logistic lasso regression in indentifying differences in a variety of datasets under a variety of conditions.

## 1 Introduction

There are many circumstances in which data collected from different sources are similar in some respects, but nonetheless differ in ways that are interesting to report. Such circumstances arise naturally in observational studies, where, for example, a clinical researcher may observe a difference in the prevalence of a condition between two groups of patients and would like to explore the reasons behind the difference; in randomized trials, where we might be interested not only in the effectiveness of a treatment but also whether its effects are particular to subgroups of the subjects and if so, what the relevant contextual relationships are; in comparative effectiveness research, where an observed difference between two clinical treatment approaches is to be explained; and other application areas. Identifying patterns of differences is also useful in abnormality and anomaly detection scenarios, where data on a potentially anomalous population of samples are compared to a "normal" baseline population.

We approach the problem of identifying interesting patterns of differences from a statistical standpoint, where given a pair of data groups over a vector of discrete random variables we would like to identify variables that exhibit statistical differences. A variable might have a different marginal distribution in the two groups and/or a different conditional distribution when conditioning on

E. Fromont et al. (Eds.): IDA 2015, LNCS 9385, pp. 275–285, 2015.
DOI: 10.1007/978-3-319-24465-5_24

the values of some of the other variables. We present and evaluate a method for identifying differences in both of those categories.

The method accomplishes this task by building models for each of the groups and for both groups, scoring local differences in distribution by comparing how well these alternative parameterizations fit the data locally, and using these local scores to obtain a score of how different the two groups are as a whole. In this paper, the performance of our method for identifying differences between groups at the variable-level is evaluated using data based on four UCI Machine Learning Repository data sets [1].

## 2 Background

We review some general background literature about distribution comparison followed by background relevant to Bayesian networks, which is the model that our method uses, with a particular focus on learning models from data and the Bayesian Dirichlet score which we use.

### 2.1 Comparing Distributions

There are various statistical methods that are applicable to the problem of identifying differences across a pair of groups. The statistical approach that most closely relates is that of *contrast set mining*. Bay and Pazzani [2] present contrast-set mining as the discovery of joint variable-value assignments that have different levels of support in different groups. The approach taken parallels association-rule mining in that the space of possible joint variable-value assignments is searched to maximize a score (in association-rule mining, this score is the lift of a rule, while in contrast-set mining a chi-square test is used). The main challenge in contrast-set mining is the search of the exponentially large space of possible sets (joint variable-value assignments), and much of the literature is dedicated to discussing heuristics and pruning rules to make the search feasible. The output of contrast-set mining is the list of joint variable-value assignments (the sets) which have differing support across groups, ranked by the extent of that difference and tested for significance. Novak et al. [11] summarize further literature on contrast-set mining and its relation to association-rule mining, emerging pattern mining, and subgroup mining. While these approaches address a similar task to that of our method, these are all value-based approaches. Their task is to identify specific value ranges in which the differences between the groups are most pronounced. In contrast, our approach is variable-based, meaning that we identify variables the distributions of which are different across the groups.

The variable-based nature of our approach bears some similarity to traditional statistical methods. There are multiple traditional statistical tests that are designed to compare distributions. For categorical variables, the Chi-Square test is applicable, it tests whether two groups are independent. This can be used to determine if a variable has different distributions across two groups by testing whether it is dependent on the group variable. For continuous variables, the

Kolmogorov-Smirnov test is often used to determine equality of distributions. Note that these tests are univariate, and cannot therefore be used to compare two multivariate groups of data directly. There are other measures of distribution differences that are multivariate in nature, such as Hotelling's T-squared test, mutual information, or Kullback-Leibler divergence. These measures are multivariate, but they do not allow for examining the contributions of differences in individual variables to the overall measure of difference across the groups. The approach we present bridges this gap by providing both a measure of overall difference, as well as a breakdown into contributions in the differences of distributions of individual variables.

### 2.2 Bayesian Networks

As mentioned above, the approach we present relies on building statistical models for the data groups to be compared. In particular, the model we construct is a Bayesian Network (BN). A BN over the variables $\mathbf{X} = (X_1, \ldots, X_n)$, where each variable $X_i$ is discrete and takes $K_i$ values, consists of a directed acyclic graph (DAG) where each node represents a variable $X_i$ and each node is associated with a conditional probability table (CPT) defined by a set of parameters

$$\theta_{ijk} = P(X_i = x_{ik} | \Pi_i = \pi_{ij}) \tag{1}$$

where $x_{ik}$ represents the $k$-th value $X_i$ takes and $\pi_{ij}$ represents the $j$-th configuration of $X_i$'s set of parents $\Pi_i$ [8].

In order to obtain a BN from data, the DAG structure is needed. In some cases the structure or elements of the structure for a given domain may be known, but often the structure must be learned from data. Daly et al. [5] provide an extensive review of BN structure learning and divide existing methods into constraint-based methods, where conditional independencies (CI) in the data are used to constrain the structure; and score search, where the space of BN structures is searched for a structure that has the best score according to some scoring criterion. Constraint-based methods use CIs obtained from statistical tests on the data to eliminate possible arcs in the network structure, such as the the PC algorithm by Spirtes and Glymour [13], for example. Score-based techniques seek to optimize some score function of the graph based on the data. The space over which many methods in this category search is that of possible DAGs, which is combinatorial in the number of variables, and the task of optimizing the score is NP-hard in general [3]. Algorithms that feasibly search the entire space of DAGs in the case of up to approximately 30 variables include dynamic programming approaches [10,12] and an application of A$^*$ search to the space of DAGs [14]. For data with more variables, many search methods apply various heuristics and do not perform an exhaustive search of the space; most commonly these methods employ some sort of greedy search strategy [5].

In our implementation we used greedy-thick-thinning, an algorithm described but not named in [8], which maximizes the K2 score [4] in a greedy fashion by starting with an empty graph, adding arcs that most increase the score until no more arc additions can increase the score, and then performs arc removals that increase the score most until no more removals increase the score. Any score search strategy can be used with our method.

### 2.3 Bayesian Dirichlet Scores for Bayesian Networks

In this work we use Bayesian Dirichlet (BD) scoring in order to leverage both the mathematical properties of the score and its statistical interpretation. The BD score is motivated by the search for a maximum *a posteriori* (MAP) model, a graph structure $\mathcal{M}$ that is most probable given the data $\mathcal{D}$ and prior belief. Directly computing a posterior $P(\mathcal{M}|\mathcal{D})$ for the structure is an intractable task; however, we can show that it is proportional to an easily computable quantity. From Bayes' rule we have that $P(\mathcal{M}|\mathcal{D}) \propto P(\mathcal{M})P(\mathcal{D}|\mathcal{M})$, where $P(\mathcal{M})$ is a prior for the graph structure. Often the graph structure prior is assumed to be uniform, an assumption that we make in this paper, but one that can be easily relaxed, and the goal becomes to maximize $P(\mathcal{D}|\mathcal{M})$, which is a marginal likelihood. Under the assumptions of global and local parameter independence, and parameter modularity [9], the marginal likelihood for the full model $P(\mathcal{D}|\mathcal{M})$ is the product of local marginal likelihoods:

$$P(\mathcal{D}|\mathcal{M}) = E_{\boldsymbol{\Theta}|\mathcal{M}} \prod_{i=1}^{n} \prod_{j=1}^{J_i} P(\mathcal{D}|\boldsymbol{\Theta}_{ij}, \mathcal{M}) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_{ij}, \mathcal{M}). \qquad (2)$$

Here we treat BN parameters as random variables with a prior distribution rather than as point values; that is, the particular value for a network parameter $\theta_{ijk}$ is just a point in the continuum of possible values that a random variable $\Theta_{ijk}$ takes. In the context of a BD score such as K2 [4] or the BDeu score [9], the prior distribution of $\boldsymbol{\Theta}_{ij} = (\Theta_{ij1}, \ldots, \Theta_{ijK_i})$ is Dirichlet with parameters $\boldsymbol{\alpha}_{ij} = (\alpha_{ij1}, \ldots, \alpha_{ijK_i})$. For a given structure, the distribution of a variable $X_i$ given a parent configuration $\pi_{ij}$ is Dirichlet-multinomial, with a closed-form marginal likelihood

$$E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}} P(\mathcal{D}|\boldsymbol{\Theta}_{ij}, \mathcal{M}) = \frac{\Gamma(\alpha_{ij\cdot})}{\Gamma(\alpha_{ij\cdot} + N_{ij\cdot})} \prod_{k=1}^{K_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (3)$$

where $J_i$ is the number of configurations of the parent set $\Pi_i$, $\alpha_{ij\cdot} := \sum_{k=1}^{K_i} \alpha_{ijk}$, $N_{ijk}$ is the number of samples in the data for which $X_i = x_{ik}$ and $\Pi_i = \pi_{ij}$, and $N_{ij\cdot} := \sum_{k=1}^{K_i} N_{ijk}$. Different choices of the Dirichlet parameter priors lead to different BD scores: for example, the K2 score is obtained from using uniform priors (all $\alpha_{ijk} = 1$), and the BDeu score is obtained from using priors with $\alpha_{ijk} = \frac{\alpha^*}{J_i K_i}$ where $\alpha^*$ is the Equivalent Sample Size (ESS) hyperparameter.

Having outlined the differences of the proposed method with common approaches to the statistical comparison of data and reviewed the relevant background regarding about BNs and the BD score, we next describe our method.

## 3 Method

Consider two groups of data $\mathcal{D}_1$ and $\mathcal{D}_2$ over the same set of variables $\mathbf{X} = (X_1, \ldots, X_n)$, and denote the concatenation of $\mathcal{D}_1$ and $\mathcal{D}_2$ by $\mathcal{D}_\cup$. If $\mathcal{D}_1$ and $\mathcal{D}_2$ are not different in a statistical sense, they follow the same distribution, which is therefore the distribution of $\mathcal{D}_\cup$. Let $\mathcal{M}_1$, $\mathcal{M}_2$, $\mathcal{M}_\cup$ denote maximum *a posteriori* (MAP) models within some space of models for the data in $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_\cup$ respectively. In the case where $\mathcal{D}_1$ and $\mathcal{D}_2$ are the same, we expect that $P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2) \leq P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ in the large sample limit, since modeling the two groups as governed by independent distributions does not yield a better fitting model than when the groups are modeled as coming from the same distribution. In the case where $\mathcal{D}_1$ and $\mathcal{D}_2$ are statistically different, we expect $P(\mathcal{D}_1|\mathcal{M}_1) \times P(\mathcal{D}_2|\mathcal{M}_2) > P(\mathcal{D}_\cup|\mathcal{M}_\cup)$ in the large sample limit.

Let us extend this idea from the model level to the parameters of the models, an extension that can be applied when the models have the following properties: the distribution of a variable $X_i$ is defined by a vector of parameters $\boldsymbol{\theta}_i$, parameters $\boldsymbol{\theta}_i$ are drawn from a distribution $\boldsymbol{\Theta}_i$, and parameter independence holds, such that, $\boldsymbol{\Theta}_i \perp \boldsymbol{\Theta}_{i'}$ for $i \neq i'$. BNs with Dirichlet parameter priors have these two properties. In order to compare parameters across models, the parameters compared must match in meaning. First we will consider the case where $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ have the same structure, and therefore, have parameters that can be perfectly matched across models; next we will extend the approach to the more general case of structures that have consistent ordering, where matching happens between sets of parameters.

In the case where $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ have the same structure, we can consider each Dirichlet-multinomial component of the full model in isolation. Consider comparing the marginal likelihood of modeling $\boldsymbol{\theta}_{ij} = P(X_i|\pi_{ij})$ independently across the two groups of data

$$T_{ij} = P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}_{ij}^{(1)} \perp \boldsymbol{\Theta}_{ij}^{(2)}) = \\ \left(E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_1} P(\mathcal{D}_1|\boldsymbol{\Theta}_{ij}, \mathcal{M}_1)\right) \left(E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_2} P(\mathcal{D}_2|\boldsymbol{\Theta}_{ij}, \mathcal{M}_2)\right) \quad (4)$$

to the marginal likelihood of modeling $\boldsymbol{\theta}_{ij}$ as being the same for both groups

$$S_{ij} = P(\mathcal{D}_1, \mathcal{D}_2|\boldsymbol{\Theta}_{ij}^{(1)} = \boldsymbol{\Theta}_{ij}^{(2)}) = E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_\cup} P(\mathcal{D}_\cup|\boldsymbol{\Theta}_{ij}, \mathcal{M}_\cup). \quad (5)$$

The ratio $T_{ij}/S_{ij}$ of these quantities is a Bayes factor that we can use to quantify the difference in the distribution $X_i|\pi_{ij}$ across the two groups of data.

Next, let us consider the more general case where the structures of $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ differ, but have consistent ordering, meaning that if $X_i$ is an ancestor of $X_j$ in any one of the networks, $X_j$ cannot be an ancestor of $X_i$ in any other network. Note that constraining the ordering of the variables in a BN does not constrain the space of joint probability distributions that can be represented. In our evaluation we enforce that constraint by learning $\mathcal{M}_\cup$ without order constraints and use the topological order of the learned network to constrain $\mathcal{M}_1$ and $\mathcal{M}_2$. There are many other possible approaches to enforcing these constraints, ranging from obtaining an order *a priori* to minimizing the number of explicit constraints using an iterative process. Exploring these alternative approaches is outside of the scope of this paper.

In this more general setting, the parent sets of a variable $X_i$ can turn out to be different in the three models, and may have some partial overlap. To handle such overlap, we introduce a new index $\eta$ as follows: Denote the parent sets of $X_i$ in $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_\cup$ by $\Pi_i^{(1)}$, $\Pi_i^{(2)}$, $\Pi_i^{(\cup)}$ respectively. Let $J_i^{(\cdot)}$ be the number of possible configurations of $\Pi_i^{(\cdot)}$, and enumerate those configurations by $j = 1, \ldots, J_i^{(\cdot)}$. Let $\Pi_i^{\cap}$ denote $\Pi_i^{(1)} \cap \Pi_i^{(2)} \cap \Pi_i^{(\cup)}$. Let $H_i$ be the number of possible configurations of $\Pi_i^{\cap}$ and enumerate those configurations by $\eta = 1, \ldots, H_i$. For example, suppose that in data where all variables are binary, for a variable $X_1$ we have $\Pi_1^{(\cup)} = \{X_2, X_3, X_4\}$, $\Pi_1^{(1)} = \{X_2, X_3, X_5\}$, and $\Pi_1^{(2)} = \{X_2, X_4, X_5\}$. Then we have that $\Pi_1^{\cap} = \{X_2\}$, and there are two possible configurations $\eta = 1$ and $\eta = 2$ for this set, corresponding to $x_{21}$ and $x_{22}$. Let $J_i^{\cdot}(\eta)$ indicate the subset of parent configurations $j \in \{1, \ldots, J_i^{\cdot}\}$ that are consistent with configuration $\eta$. That is, for example, if $\eta = 1$ represents $x_{21}$ in our example, then $J_1^{\cup}(1)$ is the set of $j$-values that correspond to the set of parent assignments $\{(x_{21}, x_{31}, x_{41})$, $(x_{21}, x_{31}, x_{42})$, $(x_{21}, x_{32}, x_{41})$, $(x_{21}, x_{32}, x_{42})\}$.

We can then then compare the marginal likelihood of modeling the entire parameter set indexed by $\eta$ as independent

$$S_{i\eta} = P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} \perp \boldsymbol{\Theta}_{i\eta}^{(2)}) =$$
$$= \left( \prod_{j \in J_i^1(\eta)} E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_1} P(\mathcal{D}_1 | \boldsymbol{\Theta}_{ij}, \mathcal{M}_1) \right) \left( \prod_{j \in J_i^2(\eta)} E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_2} P(\mathcal{D}_2 | \boldsymbol{\Theta}_{ij}, \mathcal{M}_2) \right) \quad (6)$$

to the marginal likelihood of modeling the parameter set indexed by $\eta$ as identical

$$T_{i\eta} = P(\mathcal{D}_1, \mathcal{D}_2 | \boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)}) = \prod_{j \in J_i^{\cup}(\eta)} E_{\boldsymbol{\Theta}_{ij}|\mathcal{M}_\cup} P(\mathcal{D}_\cup | \boldsymbol{\Theta}_{ij}, \mathcal{M}_\cup). \quad (7)$$

In the case of identical structures, $S_{i\eta}$ and $T_{i\eta}$ are equivalent to $S_{ij}$ and $T_{ij}$.

One interesting and useful task is the detection of differences in the distributions when only a few parameters (out of many) differ between the two groups. We can use the marginal likelihoods derived above to obtain a measure that is sensitive to the presence of changes in only some conditional distributions of a variable $X_i$, while other conditional distributions may indeed be identical across groups. Particularly, we can compute the posterior odds of seeing a difference in $X_i$ as follows:

$$O_i = \frac{1 - P(\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)} | \mathcal{D}_1, \mathcal{D}_2)}{P(\boldsymbol{\Theta}_i^{(1)} = \boldsymbol{\Theta}_i^{(2)} | \mathcal{D}_1, \mathcal{D}_2)} = \left( \prod_{\eta=1}^{H_i} \frac{1}{P(\boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)} | \mathcal{D}_1, \mathcal{D}_2)} \right) - 1. \quad (8)$$

Since the $\eta$-level is defined to be the finest level at which parameters can be compared across the two groups, we consider only the two cases of $\boldsymbol{\Theta}_{ij}$ either being independent for the two groups or being identical for the two groups. By introducing priors for these two cases we are able to compute Eq. (8). Let $p_{i\eta} = P(\boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)})$ denote the prior probability that the distribution of $X_i|\pi_{i\eta}$ is the same across the two groups. Then we have that

$$\frac{1}{P(\boldsymbol{\Theta}_{i\eta}^{(1)} = \boldsymbol{\Theta}_{i\eta}^{(2)}|\mathcal{D}_1, \mathcal{D}_2)} = \frac{S_{i\eta}(1-p_{i\eta}) + T_{i\eta}p_{i\eta}}{T_{i\eta}p_{i\eta}}. \tag{9}$$

Plugging Eq. (9) into Eq. (8) gives

$$O_i = \left(\prod_{\eta=1}^{H_i}\left(\frac{S_{i\eta}(1-p_{i\eta})}{T_{i\eta}p_{i\eta}} + 1\right)\right) - 1. \tag{10}$$

In the absence of information that would lead one to expect differences in some parameters more than in others, the priors $p_{i\eta}$ can be related to the prior probability $p_i$ of seeing no difference in the conditional distribution of variable $X_i$ by the relation $p_{i\eta} = p_i^{1/H_i}$.

The same approach can be extended to obtain posterior odds of observing a difference in any parameter of the model, expressed as

$$O = \frac{1 - P(\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)}|\mathcal{D}_1, \mathcal{D}_2)}{P(\boldsymbol{\Theta}^{(1)} = \boldsymbol{\Theta}^{(2)}|\mathcal{D}_1, \mathcal{D}_2)} = \left(\prod_{i=1}^{n}\prod_{\eta=1}^{H_i}\left(\frac{S_{i\eta}(1-p_{i\eta})}{T_{i\eta}p_{i\eta}} + 1\right)\right) - 1. \tag{11}$$

Using (11) entails that the prior for seeing no difference between the two groups is $p = \prod_{i=1}^{n}\prod_{\eta=1}^{H_i} p_{i\eta}$. Given such an overall prior $p$, a natural choice for non-informative priors is $p_{i\eta} = p^{1/(nH_i)}$: this choice of priors assumes that we are equally and independently likely to see a difference in each variable, and equally and independently likely to see a difference in each conditional probability distribution of each variable.

## 4 Evaluation

We evaluated the performance of the odds ratio $O_i$ in Eq. 10 as a score for detecting variable-level differences. Next we describe the baseline method against which we compared our method and the experimental setup, followed by the experimental results.

### 4.1 Baseline Method

As mentioned in the introduction, to our knowledge there is no prior work that addresses the difference detection problem in the same manner as our approach: a variable-based analysis, accounting for multivariate relationships, identifying variable-level differences, and requiring no informative prior knowledge. As a result, we chose to simulate a process often followed by analysts and researchers, where logistic regression models with interactions are constructed to predict a variable $X_i$ using candidate predictors, and the researcher would judge a predictor's relevance based on the strength of its corresponding weight.

For this purpose, we use lasso-regularized logistic regression [7], which maximizes an $\mathcal{L}_1$-regularized log-likelihood of a logistic model, where the strength of the regularization is modulated by a parameter $\lambda$. The effect of regularization is that as $\lambda$ decreases from $+\infty$, predictors enter the model (their coefficients in the logistic model become nonzero). To detect variable-wise differences across two pre-defined groups using lasso-regularized logistic regression, we take the data from the two groups and add a group-indicator variable $Z$ to the data. The group indicator $Z$ is a binary variable that takes the value 1 for cases coming from one of the groups, and the value 0 for cases coming from the other group. A regression model is built for predicting each variable $X_i$ given all the other data variables $X_j : 1 \leq j < i$ that precede it in the variable ordering (we provide an ordering from a true generating model for the purposes of this evaluation), the group indicator $Z$, and interactions of $Z$ with each of the data variables $X_j$. Non-binary variables were handled by using multinomial logistic regression for the target and binary coding for input variables. The largest value of $\lambda$ at which a given predictor becomes nonzero can then be used as a score of how useful that predictor is for predicting $X_i$. Hence, for each $X_i$ we can use the largest $\lambda$ that corresponds to a nonzero coefficient in the logistic model for $Z$ or an interaction with $Z$ as the score for seeing a difference in the distribution of $X_i$ across groups.

### 4.2 Data and Experimental Setup

Since in real-world data the differences between groups of data are not known in advance, for the evaluation we generated pairs of data groups from known distributions that are based on real-world data. We chose to learn networks from which to generate data because publicly available BN models are overwhelmingly diagnostic, meaning that they often contain many hidden variables, whereas we would like to have a ground-truth model that directly relates observed variables to each other. We picked data where all variables are categorical, since the BD score is designed for BNs that represent multinomial distributions. In this evaluation we used the *balance-scale*, *car*, *hayes-roth*, and *nursery* datasets available from the UCI Machine Learning Repository [1]. We learned a BN from the data for each of these sets, which is referred to as the "original BN" in the following description of the data-generation process.

We ran 72 blocks of tests, where each block is characterized by a data source (one of the UCI Datasets), a type of perturbation, the number of perturbations, and the number of samples per group. Each block consists of 20 group pairs, where each pair consists of a group of samples generated from the original BN of the data source and a group of samples generated from a perturbed BN of a data source (a different perturbed BN is obtained for each group pair). The perturbed BN was obtained by performing perturbations to the original BN. There were two categories of perturbations: parametric perturbations and structural perturbations. A parametric perturbation was performed by uniformly randomly

**Table 1.** Table of AUCs obtained from 72 blocks of tests. The first column indicates the data source for each block, the second column indicates whether the perturbation introduced was structural (Struct.) or parametric (Param.), and the third column indicates the number of perturbations. AUCs that were statistically significantly higher at the $\alpha = 0.05$ level are shown in bold.

| | | | $O_i$ AUC | $\lambda$ AUC | $p$-value | $O_i$ AUC | $\lambda$ AUC | $p$-value | $O_i$ AUC | $\lambda$ AUC | $p$-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *balance-scale* | Param. | 1 | **0.8931** | 0.7031 | 0.0074 | **0.9444** | 0.7981 | 0.0303 | 0.9563 | 0.8056 | 0.0542 |
| | | 3 | **0.8303** | 0.6457 | 0.0007 | **0.7895** | 0.6675 | 0.0336 | **0.8936** | 0.7319 | 0.0024 |
| | | 5 | **0.9038** | 0.7217 | 0.0003 | **0.8780** | 0.7733 | 0.0248 | **0.8973** | 0.7827 | 0.0131 |
| | Struct. | 1 | 0.9956 | 0.9788 | 0.1481 | 0.9981 | 0.9888 | 0.2029 | 1.0000 | 0.9981 | 0.3850 |
| | | 3 | 0.9868 | 0.9836 | 0.7843 | 0.9964 | 0.9804 | 0.1088 | 0.9992 | 0.9892 | 0.1684 |
| | | 5 | 0.9825 | 0.9721 | 0.4884 | 0.9958 | 0.9812 | 0.1509 | 0.9921 | 0.9975 | 0.4761 |
| *car* | Param. | 1 | **0.6587** | 0.5033 | 0.0282 | **0.7171** | 0.5744 | 0.0008 | 0.7221 | 0.6477 | 0.0604 |
| | | 3 | **0.7544** | 0.6296 | 0.0026 | **0.7981** | 0.6595 | $< 10^{-4}$ | **0.8260** | 0.7136 | 0.0001 |
| | | 5 | **0.7540** | 0.6149 | $< 10^{-4}$ | **0.7567** | 0.6867 | 0.0265 | **0.8079** | 0.7209 | 0.0018 |
| | Struct. | 1 | 0.9229 | 0.8525 | 0.1168 | 0.9367 | 0.9096 | 0.6049 | 0.9788 | 0.9442 | 0.0696 |
| | | 3 | 0.8272 | **0.8952** | 0.0439 | 0.8739 | 0.8920 | 0.6019 | 0.9606 | 0.9572 | 0.8760 |
| | | 5 | 0.7774 | **0.8891** | 0.0034 | 0.8695 | 0.9340 | 0.0593 | 0.9043 | 0.9538 | 0.1347 |
| *hayes-roth* | Param. | 1 | 0.7256 | 0.6944 | 0.6718 | 0.7750 | 0.7375 | 0.5904 | 0.6713 | 0.6275 | 0.5563 |
| | | 3 | **0.7648** | 0.5889 | 0.0012 | **0.7210** | 0.5990 | 0.0277 | 0.6973 | 0.5978 | 0.0634 |
| | | 5 | 0.7560 | 0.6766 | 0.1211 | 0.7490 | 0.7088 | 0.4394 | 0.7331 | 0.6830 | 0.2590 |
| | Struct. | 1 | 0.7763 | **0.8844** | 0.0401 | 0.8081 | **0.9137** | 0.0047 | 0.8344 | **0.9238** | 0.0058 |
| | | 3 | 0.8814 | 0.8998 | 0.6295 | 0.9387 | 0.9026 | 0.2959 | 0.8666 | 0.9131 | 0.2554 |
| | | 5 | 0.8546 | **0.9479** | 0.0137 | 0.9214 | 0.9505 | 0.1106 | 0.9136 | **0.9674** | 0.0297 |
| *nursery* | Param. | 1 | 0.6234 | 0.6153 | 0.9283 | 0.6644 | 0.6247 | 0.6983 | **0.8263** | 0.6931 | 0.0499 |
| | | 3 | 0.6460 | 0.5845 | 0.2978 | **0.7179** | 0.5596 | 0.0035 | 0.6833 | 0.5852 | 0.0987 |
| | | 5 | 0.6156 | 0.6105 | 0.9226 | 0.6149 | 0.5776 | 0.4796 | **0.7198** | 0.5755 | 0.0032 |
| | Struct. | 1 | 0.5897 | **0.7984** | 0.0035 | 0.7781 | 0.8469 | 0.2932 | 0.8534 | 0.8372 | 0.7410 |
| | | 3 | 0.7686 | 0.8148 | 0.2443 | 0.8487 | 0.8527 | 0.8924 | 0.9251 | 0.8895 | 0.2596 |
| | | 5 | 0.8142 | 0.8326 | 0.5419 | 0.8125 | 0.8345 | 0.4145 | 0.9270 | 0.8919 | 0.1333 |
| | | | 500 samples per group | | | 1000 samples per group | | | 5000 samples per group | | |

selecting a variable $X_i$ to perturb, selecting for it a conditional distribution $X_i|\pi_{ij}$ to perturb, and then replacing its probability mass vector with a permutation of itself. A structural perturbation was performed by randomly (with probability 1/2) deciding whether to remove or add an arc, and then selecting a random arc to add (or remove) from the existing (or absent) arcs in the network. A node (variable) is considered perturbed by a structural perturbation only if an arc into the node is added or removed.

We provide the ordering of the variables in the generating model to the logistic regression method so that it may take advantage of that information. We do not provide this information to our method in the tests reported in Table 1.

### 4.3 Results

Table 1 shows areas under receiver operating characteristic curves (AUC) of perturbation detection obtained using the posterior odds $O_i$ as compared to AUC's obtained using the $\lambda$-based score from lasso-regularized logistic regression for data group pairs generated from the respective data sources. The table also shows the $p$-value for a two-tailed test of the difference between the AUCs of the two methods, based on [6]. Of a total of 72 blocks of tests, in 53 the $O_i$ AUC is higher than the $\lambda$ AUC. At the $\alpha = 0.05$ significance level, the $O_i$ AUC's are statistically significantly better than the $\lambda$ AUC's in 21 test blocks, whereas the $O_i$ AUC is statistically significantly worse than the $\lambda$ AUC in only eight test blocks. The $p$-value of a two-sided paired Wilcoxon signed rank test on the AUCs is less than $10^{-4}$, supporting that the overall better performance of $O_i$ is not due to chance.

Every case where the the posterior odds performs statistically significantly worse than the regression-based method is a case of a structural perturbation, and we suspect that this is because perturbed structure is more difficult to recover with no order information. In a different series of tests where we provided order information to the posterior odds based method, of a total of 72 blocks of tests, in 62 the $O_i$ AUC was higher than the $\lambda$ AUC. At the $\alpha = 0.05$ significance level, the $O_i$ AUC was statistically significantly better than the $\lambda$ AUC in 43 test blocks, and worse in only one block.

As is typical for statistical methods, we see better performance for data with more samples as well as for lower-dimensional data. The results also suggest that structural differences are easier to detect than parametric ones. We believe that this is because a structural difference reflects a more substantial distributional difference than a simple parametric one, since it can be expressed as a collection of parametric differences in a network containing the removed or added arcs. Overall, our experiments show consistently good AUC for the $O_i$ score over the various generated group pairs.

## 5 Discussion

We introduced a novel variable-based approach for identifying statistical differences across a pair of groups. Evaluation of the approach on simulated data showed good performance compared to a logistic lasso baseline. The data used in the evaluation is low-dimensional because the logistic lasso baseline scales poorly to many dimensions. The most computationally demanding step in our approach is learning the three network structures. Consequently, our method scales to more dimensions to the extent that the BN structure learning algorithm used with it does. Any structure search strategy that maximizes a Bayesian Dirichlet score is a good fit for our method.

For Bayesian networks with Bayesian Dirichlet priors, we showed how to compute the posterior odds that a given variable has different distribution across the two groups, as well as the posterior odds that the two groups are different overall. The property that enables this is parameter independence in the BD framework. This approach can be applied to other models as well. The distribution of a variable in the BN formulation is simply a grouping of finer-level model parameters. Hence, any model that has similar groupings of parameters in a framework where parameter independence holds can be used with this approach.

Identification of variable-level differences across groups of multivariate data is useful in many application areas. The method presented here considers differences over the sets of relationships that are present in the MAP models constructed for modeling the groups as independent vs. identical. Particularly, for settings in which the typical approaches in practice tend to be univariate analyses and ad-hoc exploration of relationships that are suspected to be important *a priori*, we present a more systematic approach.

**Acknowledgments.** This research was supported by grant IIS-0911032 from the National Science Foundation and grant T15 LM007359 from the National Library of Medicine.

## References

1. Bache, K., Lichman, M.: UCI machine learning repository (2013). http://archive.ics.uci.edu/ml
2. Bay, S.D., Pazzani, M.J.: Detecting group differences: mining contrast sets. Data Min. Knowl. Disc. **5**(3), 213–246 (2001)
3. Chickering, D.M.: Learning Bayesian networks is NP-complete. In: Fisher, D., Lenz, H.-J. (eds.) Learning from Data. Lecture Notes in Statistics, vol. 112, pp. 121–130. Springer, Heidelberg (1996)
4. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
5. Daly, R., Shen, Q., Aitken, S.: Learning Bayesian networks: approaches and issues. Knowl. Eng. Rev. **26**(2), 99–157 (2011)
6. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics **44**, 837–845 (1988)
7. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**(1), 1–22 (2010)
8. Heckerman, D.: A tutorial on learning with Bayesian networks. In: Jordan, M.I. (ed.) Learning in Graphical Models, pp. 301–354. MIT Press, Cambridge (1999)
9. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: the combination of knowledge and statistical data. Mach. Learn. **20**, 197–243 (1995)
10. Koivisto, M., Sood, K.: Exact Bayesian structure discovery in Bayesian networks. J. Mach. Learn. Res. **5**, 549–573 (2004)
11. Novak, P.K., Lavrač, N., Webb, G.I.: Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J. Mach. Learn. Res. **10**, 377–403 (2009)
12. Silander, T., Myllymaki, P.: A simple approach for finding the globally optimal Bayesian network structure. In: Dechter, R., Richardson, T. (eds.) Proceedings of the Twenty-second Annual Conference on Uncertainty in Artificial Intelligence (UAI 2006), pp. 445–452. AUAI Press (2006)
13. Spirtes, P., Glymour, C.: An algorithm for fast recovery of sparse causal graphs. Soc. Sci. Comput. Rev. **9**(1), 62–72 (1991)
14. Yuan, C., Malone, B., Wu, X.: Learning optimal Bayesian networks using A* search. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 2186–2191. Helsinki, Finland (2011)